MASK-SPACE DISTRIBUTIONALLY ROBUST OPTIMIZATION FOR LEARNING UNDER STOCHASTIC MISSINGNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training with stochastic feature masking enables classifiers to handle missing data during inference. However, standard masked empirical risk minimization (ERM), even when aligned with the evaluation missing-rate distribution, does not explicitly control sensitivity to additional feature deletion. We propose mask-space distributionally robust optimization (MaDRO), a transport-cost DRO framework centered at the joint data-mask distribution induced by a nominal masking rule, with transport restricted to mask space and toward additional missingness. MaDRO has a closed-form robust objective function: the base masked loss plus a hinge penalty on the worst deletion-cost-adjusted loss increase caused by one additional feature deletion. For each masked training example, MaDRO either keeps the sampled mask unchanged or deletes one additional observed feature, choosing the deletion only when its loss increase outweighs its deletion cost. Experiments characterize predictive, robustness, and attribution-fidelity trade-offs across tabular, vision, and semi-synthetic settings, with improved nominal-mask prediction over aligned ERM in controlled extended-training comparisons.

1 INTRODUCTION

Missing values are pervasive in real-world data, arising from sensor failures, incomplete records, privacy constraints, or other sources. Traditional approaches impute missing entries before downstream learning. Mask-aware architectures can instead take both observed values and a feature mask as input, allowing training to match the stochastic missingness encountered at test time.

A natural training principle is masked empirical risk minimization (ERM): choose a base masking rule q, the stochastic procedure that generates masks before any local robust perturbation, and minimize the average loss under masks sampled from q (Srivastava et al., 2014; van der Maaten et al., 2013). When the same rule is used for evaluation, this aligns the mask distributions, but it only controls average performance under sampled masks. It does not control the loss increase caused by deleting one additional observed feature from a sampled mask. This local delete-one sensitivity matters for prediction under stochastic missingness and for post-hoc interpretability: Shapley-based attribution methods (Shapley, 1953; Lundberg & Lee, 2017; Chen et al., 2023; Covert et al., 2023) compare model predictions across many masked versions of the same input. Each Shapley marginal contribution compares two masks that differ by one feature. Prediction errors across these masks can affect attribution fidelity; reducing deletion sensitivity alone does not guarantee accurate attributions.

We propose mask-space distributionally robust optimization (MaDRO), a transport-cost DRO framework that adds a local robust objective to a chosen base masking rule rather than replacing it. Given q, MaDRO centers the ambiguity set at the joint distribution induced by sampling (x, y) from the data distribution and then sampling a mask from q. Transport acts only on the mask, keeps the underlying labeled example fixed, and moves only from a sampled mask to a more-missing mask (Mohajerin Esfahani & Kuhn, 2018; Globerson & Roweis, 2006). This monotonicity encodes our threat model: an input may lose additional observed features, but the adversary does not reveal missing features or change the labeled example. In our primary local instantiation, the adversary either keeps the sampled mask unchanged or deletes one additional visible feature, paying a feature-dependent deletion penalty γw_j. The resulting objective is the base masked loss plus a hinge penalty on the worst deletion-cost-adjusted loss increase. The base masking rule and the local robust objective are separate choices: MaDRO-1 uses uniform deletion costs, MaDRO-CV uses conditional-variance

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.

deletion costs, and AdvDelete denotes the zero-penalty special case of the same delete-one local objective.

Our main contributions are the following:

- **Mask-space DRO formulation.** We formulate missingness-robust learning as a transport-cost DRO problem over masked inputs. The ambiguity set is centered at the joint data-mask distribution induced by any user-chosen base masking rule, with transport restricted to monotone feature deletions toward additional missingness.
- **Tractable robust surrogate with explicit regularization effect.** Under the practical local delete-one adversary, MaDRO yields a closed-form objective: the base masked loss plus an expected hinge penalty on the model's worst-case deletion-cost-adjusted sensitivity to one additional feature deletion.
- **Empirical validation.** We evaluate mask-space robustness on public tabular and vision benchmarks and in controlled, validation-selected comparisons with aligned masking, equal-step and equal-time budgets, extended training, repeated data partitions, natural missingness, and exact multi-deletion adversaries in low dimension. The results characterize when positive deletion costs help and when competing methods remain stronger; structured shifts and Monte Carlo Shapley attribution provide complementary diagnostics.

2 Related work

Learning with missing values has been studied from several complementary perspectives, including imputation-based preprocessing, missingness-aware predictors, corruption-based regularization, and robust optimization against feature deletion. MaDRO connects these lines by applying transport-cost DRO to the mask distribution induced by a chosen base masking rule and allowing only cost-penalized monotone feature deletion.

Classical approaches handle missing values through imputation, producing a completed dataset for downstream learning. Widely used pipelines include multiple imputation via chained equations and random-forest imputation (van Buuren & Groothuis-Oudshoorn, 2011; Stekhoven & Bühlmann, 2012), with broader statistical foundations discussed in standard missing-data references (Rubin, 1976; Little & Rubin, 2002). Dick et al. (2008) jointly optimize a predictor and a distribution over imputations. More recent deep imputation methods learn a generative model of the data and impute missing entries (e.g., adversarial or variational approaches) (Yoon et al., 2018; Mattei & Frellsen, 2019). OT-based ideas also appear in missing-data modeling, for instance through differentiable Sinkhorn-based objectives for imputation and representation learning (Muzellec et al., 2020). In contrast to imputation, another line avoids reconstructing missing values and instead incorporates the mask (and, when relevant, time gaps) directly into the predictor, as in recurrent models for multivariate time series with missing values (Che et al., 2018). MaDRO follows this mask-aware approach: it does not model a generative imputer, but instead regularizes a discriminative model under a chosen base masking rule through sensitivity to additional feature removal.

Stochastic masking and input corruption have also been used as regularization. Denoising autoencoders learn representations stable to corruption (Vincent et al., 2008), and dropout injects random masking noise to reduce co-adaptation and improve generalization (Srivastava et al., 2014). Several works analyze such corruption mechanisms as implicit regularizers or propose objectives that marginalize over corrupted data copies (Wager et al., 2013; van der Maaten et al., 2013). In vision, structured deletion augmentations such as Cutout, Random Erasing, and DropBlock simulate occlusions by masking contiguous regions or blocks during training (DeVries & Taylor, 2017; Zhong et al., 2020; Ghiasi et al., 2018). These methods typically optimize an expectation under a user-chosen corruption process. MaDRO differs by adding an explicit, cost-penalized worst-case deletion on top of each sampled mask, which targets loss increases under nearby, more-missing masks.

Robustness to missing and deleted features has been studied explicitly in worst-case settings. Nightmare-style formulations train predictors to perform well after an adversary deletes features at test time (Globerson & Roweis, 2006). Related work considered learning with missing or corrupted features under adversarial assumptions and provided algorithms and guarantees in simplified model classes (Dekel & Shamir, 2008). Our formulation combines a chosen stochastic base masking rule with a local, cost-penalized monotone deletion adversary.

MaDRO is also related to adversarial training, which frames robustness as a min–max optimization where an inner adversary perturbs inputs to maximize loss while the model minimizes it (Madry et al., 2018). Subsequent work studies training stability and scheduling of perturbation strength (Shaeiri et al., 2020; Bai et al., 2021) as well as variants that use hard or auxiliary adversarial examples (Rade & Moosavi-Dezfooli, 2022). Adversarial formulations have also been proposed over discrete regularizers such as dropout masks (Park et al., 2018). MaDRO has an analogous min–max structure. Its local adversary acts on observed input-feature masks, allows only further deletion, and pays a feature-dependent cost. Our controlled experiments include an SAdD adaptation to hidden units of the same predictor, keeping the perturbation locations and constraints explicit.

Blanchet et al. (2023) show that dropout is minimax optimal for generalized linear models under independent multiplicative covariate noise. MaDRO instead conditions its local adversary on a sampled input mask and permits only cost-penalized additional deletion; its finite-max objective applies to the chosen neural predictor without a GLM assumption.

MaDRO uses standard transport-cost DRO reformulations (Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019; Chen & Paschalidis, 2020). Its modeling choice restricts transport in the joint data-mask space to monotone feature deletions that keep (x, y) fixed. The local delete-one set yields an exact finite maximum and hinge representation; together they provide a mask-specific specialization of the transport framework.

A related but distinct setting studies prediction when feature values are costly and a policy decides which features to acquire before predicting (Saar-Tsechansky et al., 2009; Shim et al., 2018). These methods jointly optimize a predictor and an acquisition strategy under budgets, whereas MaDRO assumes a chosen base masking rule and trains the predictor to tolerate additional feature loss around masks sampled from that rule.

3 Problem setup

We study supervised classification with tokenized inputs under stochastic missingness. Let $(x, y) \sim \mathcal{D}$ denote a sample from the data distribution, where $y \in \{1, \dots, C\}$ and $x = (x_1, \dots, x_D)$ contains D tokens/features. For tabular data, tokens correspond to scalar features; for vision data, tokens correspond to image patches. A token can therefore be scalar- or vector-valued; the mask operates at the token level.

Missingness is represented by a binary mask $m \in \{0, 1\}^D$, where $m_j = 1$ indicates token j is missing and $m_j = 0$ indicates it is observed. Given (x, m), the model with parameters θ outputs a predictive distribution $\hat{p}_\theta(\cdot \mid x, m)$, and we use the cross-entropy loss $\ell_\theta(x, y, m) = -\log \hat{p}_\theta(y \mid x, m)$.

Let q denote a base masking rule: given a sample, q samples the mask used for the forward pass. This notation covers NoMask, Bernoulli masking, exact-k masking, and any other user-chosen stochastic masking mechanism. A base masking rule q induces a nominal distribution P_0^q over $Z = (X, Y, M)$ by drawing $(X, Y) \sim \mathcal{D}$ and then $M \sim q(\cdot \mid X, Y)$. For the rate-mixture evaluations, the masking distribution first samples a missingness ratio $p \sim \pi_{\text{eval}}$ and then samples $m \sim \mathcal{M}_p$; $k(p) = D - \text{round}(D(1-p))$ and $\mathcal{M}_p$ is uniform over masks with exactly $k(p)$ missing tokens, i.e., $\sum_{j=1}^{D} m_j = k(p)$. The MaDRO formulation in Section 4 only requires P_0^q and therefore can be applied on top of any such base masking rule. For naturally incomplete records, the base rule preserves the intrinsic missingness mask; fixed placeholder values in missing entries are excluded by the mask. Additional perturbations can only delete observed tokens.

The nominal risk induced by a base masking rule q is

$$R_q(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{E}_{m\sim q(\cdot|x,y)} \big[\ell_\theta(x, y, m)\big], \tag{1}$$

and the corresponding masked-ERM objective is $\min_\theta R_q(\theta)$.

Masked ERM minimizes a Monte Carlo estimate of (1) by sampling fresh masks from q during training. Training and evaluation are aligned when they use the same base masking rule. The next section introduces MaDRO as a local robust objective for this same base masking rule q: it keeps the masking distribution fixed while training against additional monotone feature deletion around each sampled mask.

4 METHODOLOGY

MaDRO preserves the base masking rule q and changes the loss evaluated around each sampled mask. Transport keeps the labeled example fixed and permits additional feature deletion at a feature-dependent cost. We first define a transport-cost DRO reference objective, then specialize it to the no-op/delete-one set used in training.

4.1 TRANSPORT-COST DRO IN MASK SPACE

Fix a base masking rule q. Let $Z = (X, Y, M)$ denote the augmented random variable, where $(X, Y) \sim \mathcal{D}$ and $M \sim q(\cdot \mid X, Y)$. Let P_0^q be the induced nominal distribution over Z; when q is clear, we write P_0. Use the shorthand $\ell_\theta(z) = \ell_\theta(x, y, m)$ for $z = (x, y, m)$. For any distribution Q over Z, define the expected loss $R_Q(\theta) = \mathbb{E}_{Z\sim Q}\big[\ell_\theta(Z)\big]$.

MaDRO specifies a positive deletion-cost vector $w = (w_1, \ldots, w_D) \in (0, \infty)^D$, where w_j is the cost of newly deleting feature j. For $z = (x, y, m)$ and $z' = (x', y', m')$, define the extended-real-valued transport cost

$$c_w(z', z) = \begin{cases} \sum_{j=1}^{D} w_j(m'_j - m_j), & (x', y') = (x, y) \text{ and } m' \succeq m, \\ +\infty, & \text{otherwise}, \end{cases} \tag{2}$$

where $m' \succeq m$ denotes element-wise inequality ($m'_j \geq m_j$ for all j). Thus, admissible transports keep the underlying labeled example fixed and can only increase missingness. This one-way constraint is part of the modeling choice: it represents additional feature loss around a sampled mask, not arbitrary mask editing or recovery of already missing features. In the ordered pair $c_w(z', z)$, the second argument is the nominal sample and the first argument is the transported, more-missing sample. Since $c_w(z', z) = +\infty$ whenever $(x', y') \neq (x, y)$, any distribution Q with finite transport cost to P_0 necessarily shares the same (X, Y) marginal as P_0; the ambiguity set only alters the conditional mask distribution.

Given the cost in Eq. (2), define the directed transport-cost discrepancy between distributions Q and P_0 over Z by

$$W_{c_w}(Q, P_0) = \inf_{\Pi \in \Pi(Q, P_0)} \mathbb{E}_{(Z', Z)\sim\Pi}\big[c_w(Z', Z)\big], \tag{3}$$

where $\Pi(Q, P_0)$ denotes the set of couplings with marginals Q and P_0. Because c_w only allows monotone deletions, W_{c_w} acts as a directed discrepancy rather than a symmetric distance. For a radius $\rho \geq 0$, the weighted transport-cost ambiguity set is $\mathcal{U}_{\rho,w}(P_0) = \{Q : W_{c_w}(Q, P_0) \leq \rho\}$. The radius-form transport-cost DRO objective is

$$\min_\theta \sup_{Q \in \mathcal{U}_{\rho,w}(P_0)} R_Q(\theta). \tag{4}$$

4.2 PENALIZED OBJECTIVE, LOCAL SURROGATE, AND COST CHOICES

We first define a penalized (Lagrangian) reference objective with a fixed $\gamma > 0$. For either transport cost c used below, let $\mathcal{Q}_c(P_0) = \{Q : W_c(Q, P_0) < \infty\}$ denote its finite-cost domain:

$$J^{\text{full}}_{\gamma,w}(\theta) = \sup_{Q \in \mathcal{Q}_{c_w}(P_0)} \Big(R_Q(\theta) - \gamma W_{c_w}(Q, P_0)\Big). \tag{5}$$

The base rule q, costs w, and penalty scale γ are fixed independently of θ. For fixed θ, assume the loss is finite, nonnegative, and measurable and the expected pointwise supremum in Eq. (6) is finite. Restricting Q to the finite-cost domain avoids undefined differences of infinite risk and infinite transport cost. Using the pointwise reformulation of penalized transport-cost DRO (Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019), with a finite-mask proof given in Appendix A, Eq. (5) can be written as an expectation over P_0 of a pointwise robust surrogate:

$$\begin{aligned} J^{\text{full}}_{\gamma,w}(\theta) &= \mathbb{E}_{Z\sim P_0}\big[\phi^{\text{full}}_{\gamma,w}(\theta; Z)\big], \\ \phi^{\text{full}}_{\gamma,w}(\theta; z) &= \sup_{m' \succeq m} \left(\ell_\theta(x, y, m') - \gamma \sum_{j=1}^{D} w_j(m'_j - m_j)\right). \end{aligned} \tag{6}$$

The finite-cost domain excludes changes to (x, y) and recovery of missing features. At $\gamma = 0$, we define the objective through the finite-set pointwise maximum, without multiplying infinite costs by zero.

The inner supremum in Eq. (6) ranges over all masks $m' \succeq m$ and is expensive. We therefore use a local adversary that either keeps the mask unchanged or deletes exactly one additional visible feature. Let $V(m) \doteq \{j \in [D] : m_j = 0\}$ denote the visible features under mask m. For $j \in V(m)$, let m^{+j} be the mask obtained by setting the j-th entry to 1 and leaving all other entries unchanged. For $z = (x, y, m)$, the admissible local perturbations are $z' \in \{(x, y, m)\} \cup \{(x, y, m^{+j}) : j \in V(m)\}$. The corresponding restricted extended cost is

$$c_{\mathrm{loc},w}(z', z) = \begin{cases} 0, & z' = z, \\ w_j, & z' = (x, y, m^{+j}) \text{ for } j \in V(m), \\ +\infty, & \text{otherwise.} \end{cases} \tag{7}$$

Let $W_{c_{\mathrm{loc},w}}$ be the transport-cost discrepancy induced by this restricted local cost. For $\gamma > 0$, the corresponding local penalized objective is $J_{\gamma,w}(\theta) = \sup_{Q \in \mathcal{Q}_{c_{\mathrm{loc},w}}(P_0)}(R_Q(\theta) - \gamma W_{c_{\mathrm{loc},w}}(Q, P_0)) = \mathbb{E}_{Z \sim P_0}[\phi_{\gamma,w}(\theta; Z)]$. Repeating the pointwise transform with Eq. (7) yields the local MaDRO surrogate

$$\phi_{\gamma,w}(\theta; x, y, m) = \max\left\{\ell_\theta(x, y, m), \max_{j \in V(m)} \left[\ell_\theta(x, y, m^{+j}) - \gamma w_j\right]\right\}. \tag{8}$$

We use the convention $\max_{j \in \emptyset}(\cdot) = -\infty$; hence when $V(m) = \emptyset$, the delete-one branch is unavailable and the surrogate reduces to the nominal loss $\ell_\theta(x, y, m)$. For $\gamma = 0$, we define $J_{0,w} = \mathbb{E}_{P_0}[\phi_{0,w}]$ directly. MaDRO training uses the local $J_{\gamma,w}$. It is exact for the local no-op/delete-one set. Appendix G characterizes its relation to the full multi-deletion objective.

The base rule q and the local objective are separate choices. MaDRO-1 uses uniform costs $w_j = 1$. For MaDRO-CV, we estimate the covariance $\widehat{\Sigma}$ from standardized numerical training covariates and set $\widehat{\Omega} = (\widehat{\Sigma} + 10^{-4} I)^{-1}$. The implemented costs are

$$u_j = 1/\widehat{\Omega}_{jj}, \qquad a_j = \mathrm{clip}(u_j/\bar{u}, 0.5, 2.0), \qquad w_j^{\mathrm{CV}} = a_j/\bar{a}, \tag{9}$$

where bars denote feature-wise means. Under the ridge-regularized Gaussian covariance proxy, u_j is a conditional variance is the conditional variance of feature j given all other features; costs remain fixed across masks. Larger values make deletion more expensive; these heuristic costs encode feature predictability. The final normalization gives unit mean cost; the interval $[0.5, 2.0]$ applies to a_j, not necessarily to the final w_j.

Setting $\gamma = 0$ in the same local surrogate gives a zero-penalty delete-one special case over the finite no-op/delete-one perturbation set. Because the deletion-cost vector no longer affects branch selection, we report this special case separately as AdvDelete. Complete method names combine the base masking rule and the local robust objective, e.g., Exact-k+AdvDelete, Exact-k+MaDRO-1, or Exact-k+MaDRO-CV. Optimization uses stochastic gradient methods by selecting the active branch of Eq. (8) for each example and updating on the chosen loss; the deletion penalty is constant in θ and affects branch selection but not the branch gradient. Appendix B gives the training step in Appendix Algorithm 1.

4.3 Basic properties

We establish the hinge form, the validity of active-branch gradients, and a local radius-risk upper bound. Appendix C–E gives proofs and the conditions for passing from pointwise gradients to expected objectives.

For a fixed example (x, y, m), define the nominal loss $\ell_0(\theta) \doteq \ell_\theta(x, y, m)$ and, for each visible feature $j \in V(m)$, the raw delete-one sensitivity $\Delta_j(\theta) \doteq \ell_\theta(x, y, m^{+j}) - \ell_0(\theta)$. The worst deletion-cost-adjusted loss increase is $\Delta^\star_{\gamma,w}(\theta) \doteq \max_{j \in V(m)}(\Delta_j(\theta) - \gamma w_j)$, with the convention $\Delta^\star_{\gamma,w}(\theta) = -\infty$ when $V(m) = \emptyset$.

Proposition 4.1 (Hinge form). *For the local MaDRO surrogate in Eq.* (8)*,*

$$\phi_{\gamma,w}(\theta; x, y, m) = \ell_0(\theta) + \left(\Delta^\star_{\gamma,w}(\theta)\right)_+, \tag{10}$$

where $(u)_+ \doteq \max(u, 0)$. *Under the empty-set convention above,* $(-\infty)_+ = 0$.

Eq. (10) shows that local MaDRO equals the masked ERM loss under the chosen base rule q plus a hinge penalty that activates only when some additional deletion increases the loss by more than its deletion penalty. For MaDRO-1, this reduces to the uniform tolerance form $\ell_0(\theta) + (\max_{j \in V(m)} \Delta_j(\theta) - \gamma)_+$. For nonuniform w, the tolerance is feature dependent: deleting feature j has to increase loss by more than γw_j to activate the robust branch.

For uniform costs, write $\Delta = \max_{j \in V(m)} \Delta_j$, with $\Delta = -\infty$ if $V(m) = \emptyset$. The thresholded penalty $(\Delta - \gamma)_+$ differs from a fixed mixture of ERM and AdvDelete, whose penalty is $\lambda(\Delta)_+$ for $\lambda \in [0, 1]$: γ changes which examples activate the deletion branch, whereas λ scales every active penalty. Appendix F.1 separates deletion sensitivity from predictive quality and gives a finite-class example whose thresholded minimizer is selected by no fixed mixture; Section 5.1 compares thresholding with fixed mixing.

Proposition 4.2 (Branch gradients are valid Clarke subgradients)**.** *For fixed (x, y, m), assume the branch losses in Eq.* (8) *are continuously differentiable in θ. Any active-branch gradient $g(\theta) = \nabla_\theta \ell_\theta(x, y, m^\star)$ belongs to $\partial_C \phi_{\gamma,w}(\theta; x, y, m)$; when the active branch is unique it is the ordinary gradient. Algorithm 1 selects the no-op branch when the maximum ties the nominal loss or no feature is visible.*

The finite maximum is locally Lipschitz. Appendix D gives the subgradient proof, and Appendix E states the integrable-gradient condition for passing to the expected objective. Omitting a selected branch's constant deletion penalty leaves its gradient unchanged. These statements supply the mathematical basis for the selected-branch update.

For any fixed $\gamma > 0$ and $\rho \geq 0$, Lagrangian relaxation gives

$$\sup_{Q:\, W_{c_{\mathrm{loc},w}}(Q, P_0) \leq \rho} R_Q(\theta) \leq \gamma\rho + J_{\gamma,w}(\theta). \tag{11}$$

Indeed, every feasible Q satisfies $R_Q \leq J_{\gamma,w} + \gamma W_{c_{\mathrm{loc},w}} \leq J_{\gamma,w} + \gamma\rho$. Thus minimizing $J_{\gamma,w}$ minimizes an upper bound for the local ambiguity set at fixed γ and ρ. The full analogue uses c_w and J^{full}. For unit-cost local deletion, taking the infimum over $\gamma \geq 0$ at fixed θ gives the exact radius risk, with ρ a deletion-probability budget (Appendix F). Candidate subsampling optimizes a different, restricted surrogate, as detailed in Appendix H.

5 Experiments

Questions and primary comparisons. We ask when positive deletion costs improve on aligned masked ERM and on zero-cost AdvDelete, and how the answer changes with training budget. Our primary evidence uses validation-selected comparisons on WDBC (Wolberg et al., 1993), a 50-dimensional METABRIC-derived Gaussian benchmark, and naturally incomplete UCI Credit Approval and Horse Colic (Quinlan, 1987; McLeish & Cecile, 1989). Five repeated splits each have group-disjoint training, validation, and test sets, with two optimization seeds per split. Preprocessing and CV costs use training rows only. Numerical training and evaluation share the eleven-rate mixture $p \in \{0, 0.1, \ldots, 1\}$, including all missing; natural-data primary metrics retain intrinsic missingness. Numerical tests average 100 common mask replicates at each fixed, validation-selected predictor. Summaries average masks and seeds within each split, then report means and descriptive SD across the five overlapping partitions; conditional mask Monte Carlo errors are reported separately.

Controls and selection. Neural methods share the predictor, Adam settings, training examples, base masks, and checkpoint grid. Each trajectory provides short, equal-training-time, and extended horizons; each horizon selects its checkpoint and configuration using validation data only. Time budgets are per configuration on T4 GPUs, with full search costs reported separately. Numerical comparisons include ERM, AdvDelete, MaDRO-1/CV, and the fixed-mixture control $\ell_0 + \lambda(\Delta)_+$ with $\lambda \in \{0, .25, .5, .75, 1\}$ (Section 4.3). The mixture and each MaDRO family search five settings using the same validation rule. Natural-data controls add RandomDelete, hidden dropout, supervised adversarial dropout adapted to the same predictor (SAdD) (Park et al., 2018), and native-missing-value boosted trees (HGB). Appendix K specifies budgets, stopping rules, and separate common-optimizer and shuffled-CV controls.

Table 1: Nominal-mask accuracy (%). 100 shared mask replicates and two optimization seeds are averaged within each of five partitions; entries are means ± descriptive partition SD. Short uses 3,000 updates; time allows 300 training seconds per configuration; extended uses validation-based stopping. Each horizon selects its own checkpoint and configuration; fixed mixture tunes five ERM/AdvDelete weights. Conditional mask errors appear in Appendix L.10.

	WDBC			Gaussian		
Method	Short	Time	Extended	Short	Time	Extended
ERM	91.95 ± 1.03	91.78 ± 0.90	91.78 ± 0.90	77.09 ± 0.29	79.29 ± 0.56	79.48 ± 0.69
AdvDelete	92.24 ± 1.13	92.24 ± 1.13	92.24 ± 1.13	75.20 ± 1.29	75.38 ± 1.19	78.25 ± 0.41
MaDRO-1	92.36 ± 1.19	92.30 ± 1.20	92.24 ± 1.22	79.01 ± 0.53	79.20 ± 0.39	80.16 ± 0.55
MaDRO-CV	92.29 ± 0.90	92.31 ± 0.96	92.28 ± 0.95	79.07 ± 0.38	79.12 ± 0.40	80.29 ± 0.25
Fixed mixture	92.16 ± 1.04	92.07 ± 0.95	92.12 ± 0.88	78.36 ± 0.39	79.29 ± 0.56	79.84 ± 0.34

Table 2: Accuracy (%) under intrinsic missingness, with five-partition means ± descriptive SD. Neural results use the extended horizon and two seeds per partition. HGB is fitted once per partition and is outside the neural timing comparison.

Method	Credit Approval	Horse Colic
ERM	85.51 ± 2.94	72.89 ± 4.19
RandomDelete	85.58 ± 1.21	74.64 ± 4.41
AdvDelete	84.71 ± 1.76	74.91 ± 5.48
MaDRO-1	85.43 ± 1.76	75.39 ± 5.82
Dropout	85.87 ± 2.75	71.54 ± 4.20
SAdD	84.93 ± 1.91	71.36 ± 4.14
HGB	85.94 ± 2.49	77.32 ± 4.54

Exploratory sweeps and diagnostics. Separate four-run sweeps cover WDBC, METABRIC (Curtis et al., 2012), MiniBooNE (Roe et al., 2010), MNIST (LeCun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky, 2009). Their Exact-k sampler draws a missing count uniformly from $\{0, \ldots, D - 1\}$, whereas evaluation uses the eleven-rate mixture; Bernoulli training deletes each feature with probability 0.5. Positive costs use $\gamma \in \{0.05, 0.1, 0.2, 0.5, 1\}$, and CV is restricted to numerical features. Displayed positive-cost configurations maximize the reported test mixture mean, so these sweeps describe parameter sensitivity rather than validation-selected performance. Appendices I–J specify their separate training budgets and split seeds. Gaussian oracle diagnostics examine structured missingness, Shapley fidelity, and candidate sampling. An eight-feature WDBC study exactly enumerates deletion subsets for budgets $k \in \{1, 2, 4, 8\}$.

5.1 Training budget and natural missingness

MaDRO improves mean accuracy over aligned ERM on both WDBC and Gaussian data after extended training (Table 1). Averaging 100 shared mask replicates, Gaussian gains are 0.68/0.81 percentage points for MaDRO-1/CV over ERM and 1.91/2.04 over AdvDelete, showing the value of tempering zero-cost deletion. Paired gains over ERM span $[-0.17, 1.21]$ and $[0.18, 1.47]$ across partitions. On WDBC, gains are 0.46/0.50 points over ERM, with accuracy close to AdvDelete (0.00/0.04 points higher). Training budget shapes these comparisons: Gaussian gains over ERM are 1.92/1.99 points at 3,000 updates, while ERM leads by 0.09/0.18 points at equal training time and itself improves by 2.39 points with extended training. Selected validation loss and training–validation gaps are analyzed jointly in Appendix L.12.1.

Under intrinsic missingness, extended MaDRO-1 improves over ERM by 2.50 points on Horse and is close on Credit (-0.07 points; Table 2). Paired partition SDs are 5.01 and 1.41 points, respectively; HGB provides the strongest mean accuracy on both datasets. Complete results appear in Appendix L; separate single-partition controls appear in Appendices M and N.

For Gaussian extended training, full-family search takes 10,459.8/10,115.7 seconds for MaDRO-1/CV versus 514.2 for ERM (Table 40). This includes all five positive-cost candidates and complements per-configuration time matching when choosing a training and tuning budget.

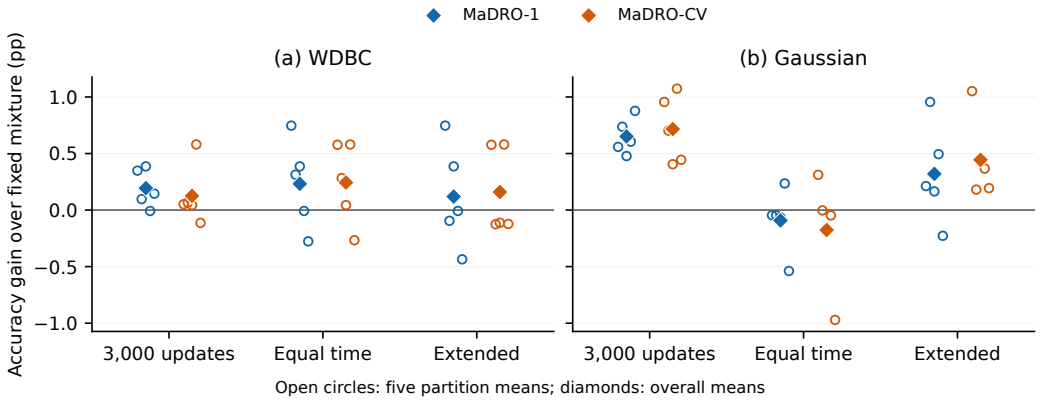

Figure 1: Accuracy differences from the validation-selected fixed mixture across training horizons. Each open circle is one of five paired partition means, averaging two optimization seeds and 100 shared test-mask replicates; diamonds average the five partitions. Positive values favor MaDRO. Horizontal offsets distinguish partitions. Conditional mask Monte Carlo errors are reported separately in Appendix L.10.

Threshold versus fixed mixture. Figure 1 shows every partition-level contrast. On Gaussian data, MaDRO-1/CV exceed the validation-selected fixed mixture by 0.65/0.72 points at 3,000 updates and 0.32/0.44 after extended training. The extended CV advantage is positive in all five partition means, ranging from 0.18 to 1.05 points; the unit-cost difference varies in sign. Conditional mask Monte Carlo SE is about 0.015 points for either difference (Appendix L.10). At equal training time, the mixture selects ERM in all ten runs and leads by 0.09/0.18 points. Extended WDBC differences favor MaDRO-1/CV by 0.12/0.16 points but vary in sign across partitions. A separate joint-optimizer control on one partition and two seeds retains extended MaDRO-1/CV gains over the mixture: 0.88/0.68 points on WDBC and 0.44/0.91 on Gaussian data. Equal-time Gaussian accuracy favors ERM and the mixture in both protocols. Appendices L.8 and L report all metrics, selections, and full search costs.

5.2 Exploratory public benchmarks

The test-selected MaDRO-1 configurations in Appendix Figure 3 have higher mixture means than Exact-k ERM; positive costs exceed AdvDelete on three of six datasets. The six-benchmark sweeps characterize behavior across tabular and vision inputs; the complementary validation-selected comparisons in Table 1 evaluate the gains under controlled selection and training budgets. Tables 7–9 retain all scanned configurations and the displayed choices.

5.3 Deletion costs and structured missingness

Positive costs raise masked Gaussian accuracy from 0.726 to 0.792 in the oracle sweeps; unit and CV costs are close (Figure 2a). In the five-partition extended comparison, MaDRO-1 reaches 83.10%/83.33% under MAR/MNAR, versus ERM's 81.73%/81.99%; the relative ranking of CV and unit costs varies with the mask condition (Table 29). These stress masks use the distinct definitions in Appendix K. These results measure empirical transfer to MCAR and MAR laws beyond the nominal mixture's one-way transport domain; Appendix G specifies the domain covered by the radius-risk bound.

At the extended horizon on one partition with two seeds and the default optimizer, CV exceeds shuffled CV by 0.12 points on WDBC and 0.62 on Gaussian data. Gaussian CV weights are nearly uniform (feature-wise SD 0.011–0.012), placing this comparison close to the unit-cost regime. The shuffled control characterizes the contribution of feature-specific cost placement on the tested partition.

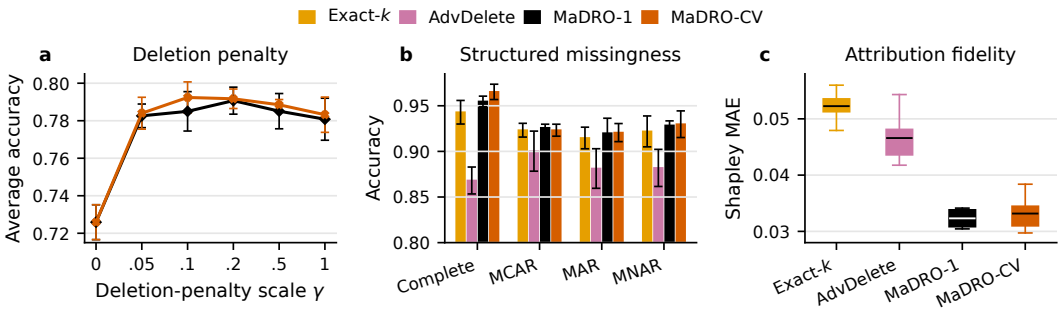

Figure 2: Separate four-run Gaussian oracle diagnostics. (a) Exact-k+MaDRO-1/CV cost sweeps; $\gamma = 0$ gives AdvDelete. (b) Complete and structured test masks with five always-observed anchors, using one checkpoint per run across conditions. Markers/bars and error bars show means and one SD. (c) Monte Carlo Shapley MAE on positive-class examples for PC1–PC5; boxes summarize runs and horizontal lines mark means.

Common-cost profiles distinguish average deletion sensitivity from its upper tail (Appendix L.11). On Gaussian data, AdvDelete has the lowest mean sensitivity, while MaDRO-1/CV combine lower mean nominal BCE with lower mean hinge excess at the three largest evaluated costs than AdvDelete.

5.4 Attribution, local scope, and computation

Using 512 permutations of the 50-feature game on 190–191 positive-class test examples per run, Exact-k Shapley MAE on PC1–PC5 falls from 0.052 to 0.032 with MaDRO-1 and 0.033 with CV (Figure 2c). Top-1 absolute-attribution agreement among those five features is 0.857, 0.871 and 0.850, respectively, complementing the MAE comparison for the same class and feature subset.

In eight-feature WDBC, full subset enumeration reduces the held-out multi-deletion hinge from 0.1662 to 0.0882 and raises training time from 9.9 to 52.1 seconds, while accuracy changes from 90.45% to 90.57%. Intermediate deletion budgets reveal a nonmonotone accuracy tradeoff (Tables 80–81). Sampling 32 Gaussian MaDRO-CV candidates uses 53% of full-delete-one-inner-max training time with similar accuracy (Table 16).

Scope and extensions. The formal guarantee concerns the local monotone ambiguity set; exact subset enumeration examines larger deletion sets empirically. Sensitivity and trajectory analyses are descriptive. The fixed-mixture study covers two numerical datasets; SAdD uses an architecture-specific adaptation. Extensions include categorical/vision cost models and broader natural-missingness studies.

6 Conclusion

MaDRO provides a transport-cost formulation of missing-feature robustness that preserves the chosen base masking law and each labeled example. Its exact delete-one objective adds an interpretable hinge on deletion-cost-adjusted local sensitivity, with valid selected-branch updates and a risk bound for the local ambiguity set. Validation-selected comparisons show higher extended-horizon mean accuracy than aligned ERM on WDBC and Gaussian data, and a Gaussian benefit over both zero-cost deletion and fixed adversarial-loss mixing. Budget comparisons quantify the accompanying computational tradeoff. Structured-missingness tests demonstrate empirical transfer, while Gaussian-oracle analyses show lower Shapley MAE on the evaluated positive-class examples and PC1–PC5. These results establish cost-thresholded feature deletion as a practical way to control local mask sensitivity, with cost design and candidate sampling offering further directions for efficient missing-data learning.

AI USE STATEMENT

Generative AI tools were used to assist with writing and polishing the manuscript, including language editing and LaTeX formatting.

REFERENCES

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936.

José Blanchet, Yang Kang, José Luis Montiel Olea, Viet Anh Nguyen, and Xuhui Zhang. Dropout training is distributionally robust optimal. *Journal of Machine Learning Research*, 24(180):1–60, 2023. URL `https://www.jmlr.org/papers/v24/21-0377.html`.

James V. Burke, Xiaojun Chen, and Hailin Sun. The subdifferential of measurable composite max integrands and smoothing approximation. *Mathematical Programming*, 181(2):229–264, 2020. doi: 10.1007/s10107-019-01441-9.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023. doi: 10.1038/s42256-023-00657-x.

Ruidi Chen and Ioannis Ch. Paschalidis. Distributionally robust learning. *Foundations and Trends in Optimization*, 4(1–2):1–243, 2020. doi: 10.1561/2400000026.

Ian Connick Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate Shapley values with vision transformers. In *International Conference on Learning Representations*, 2023.

Christina Curtis et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983.

Ofer Dekel and Ohad Shamir. Learning to classify with missing and corrupted features. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 216–223. ACM, 2008.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Uwe Dick, Peter Haider, and Tobias Scheffer. Learning from incomplete data with infinite imputations. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 232–239. ACM, 2008. doi: 10.1145/1390156.1390186.

Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. DropBlock: A regularization method for convolutional networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Amir Globerson and Sam Roweis. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 353–360. ACM, 2006. doi: 10.1145/1143844.1143889.

Jiantao Jiao, Thomas A. Courtade, Kartik Venkat, and Tsachy Weissman. Justification of logarithmic loss via the benefit of side information. *IEEE Transactions on Information Theory*, 61(10): 5357–5365, 2015. doi: 10.1109/TIT.2015.2462848.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 2 edition, 2002. ISBN 9780471183860. doi: 10.1002/9781119013563.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423. PMLR, 2019.

Mary McLeish and Matt Cecile. Horse colic. UCI Machine Learning Repository, 1989. URL `https://doi.org/10.24432/C58W23`.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018. doi: 10.1007/s10107-017-1172-1.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pp. 7130–7140. PMLR, 2020.

Sungrae Park, Jun-Keon Park, Su-Jin Shin, and Il-Chul Moon. Adversarial dropout for supervised and semi-supervised learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3917–3924. AAAI Press, 2018.

J. R. Quinlan. Credit approval. UCI Machine Learning Repository, 1987. URL `https://doi.org/10.24432/C5FS30`.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6.

Byron P. Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. MiniBooNE particle identification. UCI Machine Learning Repository, 2010.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009. doi: 10.1287/mnsc.1080.0952.

Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, Princeton, NJ, 1953.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: 10.1093/bioinformatics/btr597.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Laurens van der Maaten, Minmin Chen, Stephen Tyree, and Kilian Weinberger. Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 410–418, Atlanta, Georgia, USA, 2013. PMLR. URL https://proceedings.mlr.press/v28/vandermaaten13.html.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008. doi: 10.1145/1390156.1390294.

Stefan Wager, Sida I. Wang, and Percy Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26, pp. 351–359, 2013.

William H. Wolberg, Olvi L. Mangasarian, and Nick Street. Breast cancer wisconsin (diagnostic). UCI Machine Learning Repository, 1993.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5689–5698. PMLR, 2018.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13001–13008, 2020.

A From penalized transport-cost DRO to the pointwise surrogate

This section justifies the transition from the penalized transport-cost DRO objective (Eq. (5)) to the pointwise robust surrogate (Eq. (6)). The equality below is a finite-mask instance of standard transport-cost DRO reformulations (Mohajerin Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019). We give a direct coupling proof because, for both costs used in this paper, the finite-cost perturbation set around each sampled mask is finite and nonempty.

Lemma A.1 (Penalized transport-cost DRO admits a pointwise form)**.** *Let c be either c_w or $c_{\mathrm{loc},w}$, with finite D and fixed $w \in (0,\infty)^D$, and define $\mathcal{Q}_c(P_0) = \{Q : W_c(Q,P_0) < \infty\}$. For fixed θ, assume that ℓ_θ is a finite, nonnegative measurable function and that the expectation on the right below is finite. Then, for every $\gamma > 0$,*

$$\sup_{Q\in\mathcal{Q}_c(P_0)} \Big(R_Q(\theta) - \gamma W_c(Q,P_0)\Big) = \mathbb{E}_{Z\sim P_0}\Big[\max_{z'\in F_c(Z)} \big(\ell_\theta(z') - \gamma c(z',Z)\big)\Big], \tag{12}$$

where $F_c(z) = \{z' : c(z',z) < \infty\}$.

Proof. Write $C_w = \sum_{j=1}^D w_j$ and $\psi(z) = \max_{z'\in F_c(z)}[\ell_\theta(z') - \gamma c(z',z)]$. For either cost, $F_c(z)$ is a nonempty finite set and every finite cost belongs to $[0, C_w]$. The no-op branch gives $\psi(z) \geq \ell_\theta(z) \geq 0$, and every feasible pair satisfies

$$0 \leq \ell_\theta(z') \leq \psi(z) + \gamma C_w. \tag{13}$$

Consequently, if $Q \in \mathcal{Q}_c(P_0)$, there exists a finite-cost coupling with second marginal P_0, and integrating this bound shows $R_Q(\theta) < \infty$. The penalized objective therefore contains no undefined difference of infinities.

Let $\mathcal{C}_c(P_0)$ be the set of couplings with second marginal P_0 and finite expected cost. For feasible Q, couplings with infinite cost can be excluded when taking the infimum that defines W_c. Thus

$$R_Q(\theta) - \gamma W_c(Q,P_0) = \sup_{\substack{\Pi\in\Pi(Q,P_0)\\ \mathbb{E}_\Pi c<\infty}} \mathbb{E}_\Pi[\ell_\theta(Z') - \gamma c(Z',Z)], \tag{14}$$

$$\sup_{Q\in\mathcal{Q}_c(P_0)} \big(R_Q(\theta) - \gamma W_c(Q,P_0)\big) = \sup_{\Pi\in\mathcal{C}_c(P_0)} \mathbb{E}_\Pi[\ell_\theta(Z') - \gamma c(Z',Z)] \leq \mathbb{E}_{P_0}\psi(Z). \tag{15}$$

The inequality follows pointwise from the definition of ψ.

For the reverse inequality, choose $T(z) \in \arg\max_{z'\in F_c(z)}[\ell_\theta(z') - \gamma c(z',z)]$ using a fixed lexicographic tie-breaking rule over masks. This selector is measurable because there are finitely many masks and each branch value and feasibility condition is measurable. The law $\Pi^\star$ of $(T(Z),Z)$ has expected cost at most C_w, so $\Pi^\star \in \mathcal{C}_c(P_0)$. Its first marginal $Q^\star$ belongs to $\mathcal{Q}_c(P_0)$, and

$$R_{Q^\star}(\theta) - \gamma W_c(Q^\star,P_0) \geq \mathbb{E}_{\Pi^\star}[\ell_\theta(Z') - \gamma c(Z',Z)] = \mathbb{E}_{P_0}\psi(Z).$$

This proves equality. Distributions outside $\mathcal{Q}_c(P_0)$ are excluded; no convention for $+\infty - \infty$ is used. □

Effect of extended costs. In our construction (Eq. (2)), $c_w(z',z) = +\infty$ whenever $(x',y') \neq (x,y)$ or $m' \not\geq m$. Since the supremum in Eq. (6) is taken over z' with finite cost, infeasible points are excluded and the adversary is restricted to admissible monotone feature deletions of the same example. The finite part of c_w assigns total deletion cost $\sum_j w_j(m'_j - m_j)$. The lemma is stated for $\gamma > 0$; for $\gamma = 0$, the paper uses the finite-cost pointwise adversary directly and does not identify it with an unrestricted zero-penalty supremum over all distributions.

MaDRO as a restricted pointwise adversary. Local MaDRO further restricts the admissible set of z' to $\{(x,y,m)\} \cup \{(x,y,m^{+j}) : j \in V(m)\}$ and assigns the one-feature deletion cost w_j (Eq. (7)). Applying Lemma A.1 with $c = c_{\mathrm{loc},w}$ yields $J_{\gamma,w}(\theta) = \mathbb{E}_{P_0}[\phi_{\gamma,w}(\theta;Z)]$ for $\gamma > 0$, with the closed form in Eq. (8).

B Local MaDRO Training Step

The displayed minibatch loss omits branch-constant deletion penalties because they do not affect the optimizer step. Full surrogate values include these constants when reporting robust objective values.

Algorithm 1 Local MaDRO training step

Input: mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$, model f_θ, loss $\ell_\theta(\cdot)$, fixed base sampler q, penalty $\gamma \geq 0$, fixed costs w
for $i = 1$ to B **do**
  Sample $m_i \sim q(\cdot \mid x_i, y_i)$
  $\ell_{0,i} \leftarrow \ell_\theta(x_i, y_i, m_i)$
  $\widetilde{\ell}^{\star}_{1,i} \leftarrow -\infty$
  if $V(m_i) \neq \emptyset$ **then**
    $j_i^{\star} \in \arg\max_{j \in V(m_i)} \left[\ell_\theta(x_i, y_i, m_i^{+j}) - \gamma w_j\right]$
    $\widetilde{\ell}^{\star}_{1,i} \leftarrow \ell_\theta(x_i, y_i, m_i^{+j_i^{\star}}) - \gamma w_{j_i^{\star}}$
  end if
  if $\widetilde{\ell}^{\star}_{1,i} > \ell_{0,i}$ **then**
    $m_i^{\star} \leftarrow m_i^{+j_i^{\star}}$
  else
    $m_i^{\star} \leftarrow m_i$
  end if
end for
Take one optimizer step using $L \leftarrow \frac{1}{B}\sum_{i=1}^{B} \ell_\theta(x_i, y_i, m_i^{\star})$

C Proof of Proposition 4.1

Proof. Fix (x, y, m) and write $\ell_0(\theta) \doteq \ell_\theta(x, y, m)$. Recall that the local MaDRO surrogate (Eq. (8)) is

$$\phi_{\gamma,w}(\theta; x, y, m) = \max\left(\ell_0(\theta),\ \max_{j \in V(m)} \left[\ell_\theta(x, y, m^{+j}) - \gamma w_j\right]\right),$$

with the delete-one maximum unavailable when $V(m) = \emptyset$. For $j \in V(m)$, define

$$\Delta_j(\theta) \doteq \ell_\theta(x, y, m^{+j}) - \ell_0(\theta),$$

and define

$$\Delta^{\star}_{\gamma,w}(\theta) \doteq \max_{j \in V(m)} \big(\Delta_j(\theta) - \gamma w_j\big),$$

with the convention $\Delta^{\star}_{\gamma,w}(\theta) = -\infty$ when $V(m) = \emptyset$. Using $\max(a, b) = a + \max(0, b - a)$ gives

$$\begin{aligned}
\phi_{\gamma,w}(\theta; x, y, m) &= \ell_0(\theta) + \max\left(0,\ \max_{j \in V(m)} \left[\ell_\theta(x, y, m^{+j}) - \gamma w_j - \ell_0(\theta)\right]\right) \\
&= \ell_0(\theta) + \max\left(0,\ \max_{j \in V(m)} \left[\Delta_j(\theta) - \gamma w_j\right]\right) \\
&= \ell_0(\theta) + \big(\Delta^{\star}_{\gamma,w}(\theta)\big)_+.
\end{aligned} \tag{16}$$

When $V(m) = \emptyset$, the convention makes the positive part equal to zero, so the same expression reduces to $\phi_{\gamma,w}(\theta; x, y, m) = \ell_0(\theta)$. This proves Eq. (10). □

D Proof of Proposition 4.2

Proof. Fix (x, y, m). If $V(m) = \emptyset$, then $\phi_{\gamma,w}(\theta; x, y, m) = \ell_\theta(x, y, m)$ and $\nabla_\theta \ell_\theta(x, y, m)$ is both the gradient and a valid Clarke subgradient, so the claim holds. Assume below that $V(m) \neq \emptyset$.

Define the index set $\mathcal{I} \doteq \{0\} \cup V(m)$ and consider the collection of continuously differentiable functions

$$f_0(\theta) \doteq \ell_\theta(x, y, m), \qquad f_j(\theta) \doteq \ell_\theta(x, y, m^{+j}) - \gamma w_j,\ \ j \in V(m). \tag{17}$$

Then the local MaDRO surrogate can be written as the pointwise maximum

$$\phi_{\gamma,w}(\theta; x, y, m) = \max_{i \in \mathcal{I}} f_i(\theta). \tag{18}$$

Let $\mathcal{I}^\star(\theta) \doteq \arg\max_{i\in\mathcal{I}} f_i(\theta)$ denote the active set. A standard property of pointwise maxima of continuously differentiable functions states that

$$\partial_C \phi_{\gamma,w}(\theta; x, y, m) = \mathrm{conv}\Big\{\nabla_\theta f_i(\theta) : i \in \mathcal{I}^\star(\theta)\Big\}, \tag{19}$$

where ∂_C is the Clarke subdifferential and $\mathrm{conv}(\cdot)$ denotes convex hull. In particular, for any active index $i \in \mathcal{I}^\star(\theta)$, $\nabla_\theta f_i(\theta) \in \partial_C \phi_{\gamma,w}(\theta; x, y, m)$.

Choose any

$$j^\star \in \arg\max_{j\in V(m)} \left[\ell_\theta(x, y, m^{+j}) - \gamma w_j\right], \tag{20}$$

which exists since $V(m)$ is finite. If

$$\ell_\theta(x, y, m^{+j^\star}) - \gamma w_{j^\star} > \ell_0(\theta),$$

then $f_{j^\star}$ is an active branch and

$$\nabla_\theta f_{j^\star}(\theta) = \nabla_\theta \ell_\theta(x, y, m^{+j^\star}) = \nabla_\theta \ell_\theta(x, y, m^\star), \tag{21}$$

since $m^\star = m^{+j^\star}$. Otherwise, f_0 is an active branch (and in the tie case both f_0 and at least one delete-one branch are active); choosing $m^\star = m$ yields $\nabla_\theta f_0(\theta) = \nabla_\theta \ell_\theta(x, y, m^\star)$. In all cases, $\nabla_\theta \ell_\theta(x, y, m^\star)$ is the gradient of an active branch, hence it belongs to the subdifferential by Eq. (19). When the active branch is unique, the active set is a singleton and this gradient equals $\nabla_\theta \phi_{\gamma,w}(\theta; x, y, m)$. □

E From branch gradients to expected objectives

Proposition 4.2 is a pointwise statement. The following result gives sufficient conditions for an expectation statement; these conditions make explicit the local integrability required for subdifferential interchange (Burke et al., 2020).

Corollary E.1 (Expected active-branch gradients). *Fix $\gamma \geq 0$ and $w \in (0,\infty)^D$, and suppose P_0 is independent of θ. Let θ_0 lie in an open parameter set, assume every branch loss is continuously differentiable there, and assume $J_{\gamma,w}(\theta_0) < \infty$. Suppose that on some neighborhood U of θ_0 there is a measurable integrable envelope $L(Z)$ such that*

$$\sup_{\vartheta\in U} \max_{z'\in F_{c_{\mathrm{loc},w}}(Z)} \|\nabla_\vartheta \ell_\vartheta(z')\| \leq L(Z), \qquad \mathbb{E}_{P_0} L(Z) < \infty. \tag{22}$$

For any measurable exact active-branch selection $g(\theta_0; Z)$ as in Proposition 4.2,

$$\mathbb{E}_{P_0}[g(\theta_0; Z)] \in \partial_C J_{\gamma,w}(\theta_0).$$

Proof. Write $f_i(\theta; Z)$ for the nominal branch and the penalized delete-one branches, and let $\mathcal{I}^\star(\theta; Z)$ be their active set. After shrinking U to a ball, the envelope bounds the Lipschitz modulus of every branch and their maximum. Hence $J_{\gamma,w}$ is finite and locally Lipschitz near θ_0, and g is integrable. For any direction d, dominated convergence yields

$$J'_{\gamma,w}(\theta_0; d) = \mathbb{E}_{P_0}\left[\max_{i\in\mathcal{I}^\star(\theta_0;Z)} \langle \nabla_\theta f_i(\theta_0; Z), d\rangle\right].$$

A finite maximum of continuously differentiable functions is Clarke regular. Applying the same integrable bound to difference quotients based at nearby parameters gives $J^\circ_{\gamma,w}(\theta_0; d) \leq \mathbb{E}_{P_0}[\phi^\circ_{\gamma,w}(\theta_0; Z; d)] = J'_{\gamma,w}(\theta_0; d)$. The reverse inequality holds by the definition of the Clarke directional derivative, so equality holds. Since g is the gradient of an active branch,

$$\langle \mathbb{E}_{P_0} g, d\rangle \leq \mathbb{E}_{P_0} \max_{i\in\mathcal{I}^\star(\theta_0;Z)} \langle \nabla_\theta f_i, d\rangle = J^\circ_{\gamma,w}(\theta_0; d)$$

for every d. The defining inequalities for the Clarke subdifferential prove the claim. □

Finite training sets and optimization. For $\mathcal{D}_n = n^{-1}\sum_{i=1}^n \delta_{(x_i,y_i)}$, the empirical objective is

$$\widehat{J}_{\gamma,w}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{m\in\{0,1\}^D} q(m \mid x_i, y_i)\,\phi_{\gamma,w}(\theta; x_i, y_i, m).$$

With fixed q and finite branch losses, this is a finite weighted sum, and continuous differentiability supplies the local envelope automatically. Independent sampling of a training example and its base mask gives a branch gradient whose expectation belongs to $\partial_C \widehat{J}_{\gamma,w}$. This unbiasedness statement presumes independent sampling conditional on the current parameters; it does not assert conditional unbiasedness for every within-epoch update under sampling without replacement. Algorithm 1 specifies the selected-branch gradient, which can be used by the chosen optimizer. The result characterizes the update direction; optimizer convergence requires separate assumptions.

F Unit-cost threshold and deletion-probability budget

Proposition F.1 (A probability budget for unit-cost local deletion)**.** *Fix θ, let $w_j = 1$, and put $T = (\max_{j\in V(M)}[\ell_\theta(X,Y,M^{+j}) - \ell_\theta(X,Y,M)])_+$, with $T = 0$ when no feature is visible. Assume a standard Borel example space, finite nonnegative measurable branch losses, and $R_{P_0}(\theta) + \mathbb{E}_{P_0}T < \infty$. For $0 \le \rho \le 1$,*

$$\sup_{Q: W_{c_{\mathrm{loc},1}}(Q,P_0)\le\rho} R_Q(\theta) = R_{P_0}(\theta) + \inf_{\gamma\ge 0}\{\gamma\rho + \mathbb{E}_{P_0}(T-\gamma)_+\}.$$

Writing $p = P_0(T > 0)$, for $0 < \rho < p$ an optimal positive threshold t satisfies $P_0(T > t) \le \rho \le P_0(T \ge t)$. The adversary deletes a maximizing feature on $T > t$ and randomizes on $T = t$ to achieve deletion probability ρ. For $\rho \ge p$, the gain is $\mathbb{E}T$, attained by deleting only on $T > 0$; for $\rho = 0$, the gain is zero.

Proof. Let $s(m) = \sum_j m_j$. Every finite-cost pair satisfies $c_{\mathrm{loc},1}(z', z) = s(m') - s(m)$. Whenever $W(Q, P_0) < \infty$, some finite-cost coupling exists, and every such coupling has cost $\mathbb{E}_Q s(M) - \mathbb{E}_{P_0} s(M)$; thus each attains W. Its loss gain is bounded by $\mathbb{E}[a(Z)T(Z)]$, where $a(Z)$ is its conditional deletion probability and $\mathbb{E}a \le \rho$. Conversely, any measurable $a \in [0, 1]$ supported on $T > 0$ is implemented by deleting a measurably selected best feature with probability $a(Z)$. Allowing a on $T = 0$ cannot improve the scalar objective, so the optimal gain is $\sup_{0\le a\le 1,\, \mathbb{E}a\le\rho} \mathbb{E}[aT]$. The bound $aT \le \gamma a + (T-\gamma)_+$ gives one inequality. For $0 < \rho < p$, a finite positive $(1-\rho)$-quantile t exists. Set $a = \mathbf{1}_{\{T>t\}} + b\mathbf{1}_{\{T=t\}}$, where $b = (\rho - P_0(T > t))/P_0(T = t)$, or $b = 0$ when the denominator vanishes. Then $\mathbb{E}a = \rho$ and $\mathbb{E}[aT] = t\rho + \mathbb{E}(T-t)_+$, proving equality. For $\rho \ge p$, take $a = \mathbf{1}_{\{T>0\}}$ and $\gamma = 0$. At $\rho = 0$, take $a = 0$ and use dominated convergence as $\gamma \to \infty$. A finite minimizing threshold at zero radius exists exactly when T is essentially bounded. □

For $0 < \rho < 1$, the gain is $\rho\ \mathrm{CVaR}_{1-\rho}(T)$, by the established upper-tail minimization formula (Rockafellar & Uryasev, 2002). Thus ρ budgets the probability mass of example–mask pairs receiving one additional deletion. The no-op tie rule used in training induces $\rho_\gamma = P_0(T > \gamma)$, which varies with θ. The experiments use fixed penalties while this induced radius evolves with the predictor.

F.1 Prediction quality and deletion sensitivity

Fix a visible set S and a feature $j \in S$, and put $B = S \setminus \{j\}$. Evaluate both masked predictors $f_A(\cdot \mid X_A)$, $A \in \{S, B\}$, under the same data law P on a standard Borel feature space. For a finite label alphabet and finite log risks $R_f(A) = \mathbb{E}[-\log f_A(Y \mid X_A)]$, standard cross-entropy decomposition gives

$$R_f(B) - R_f(S) = I(Y; X_j \mid X_B) + \varepsilon_f(B) - \varepsilon_f(S),$$
$$\varepsilon_f(A) = \mathbb{E}\,\mathrm{KL}(P(Y \mid X_A) \| f_A(\cdot \mid X_A))\,.$$

Indeed, $R_f(A) = H(Y \mid X_A) + \varepsilon_f(A)$; the Bayes-risk difference is the information benefit of the additional observation (Jiao et al., 2015). For fixed P, S, j, the information term is unchanged across predictors, so a smaller signed loss increase can also reflect larger prediction error before deletion.

This identity concerns a fixed deletion under one data law; T instead takes an adaptive maximum and positive part. Risks conditional on different realized masks require specifying their respective conditional data laws.

For a finite-loss illustration, let $Y = X \sim \mathrm{Bernoulli}(1/2)$ and let q reveal the single feature independently with probability $a \in (0, 1]$. Predictors use only the mask and visible values. The constant predictor f_c always predicts $(1/2, 1/2)$. The predictor f_r assigns probability $r \in (1/2, 1)$ to the label indicated by visible X and predicts $(1/2, 1/2)$ when missing. With $d = \log(2r) > 0$,

$$R_q(f_r) = \log 2 - ad < R_q(f_c) = \log 2,$$

although f_r has mean delete-one sensitivity ad and f_c has zero. Write $J_\gamma = J_{\gamma,1}$ for unit costs. Every predictor's zero-cost robust loss includes the empty-coalition branch, so $J_0(f) \geq \mathbb{E}[-\log f_\emptyset(Y)] \geq H(Y) = \log 2$. Both f_c and f_r attain this bound and are global minimizers over any admissible predictor class containing them. Thus zero-cost robustness can be indifferent between predictors of different nominal quality. For every $\gamma > 0$,

$$J_\gamma(f_r) = \log 2 - a \min\{\gamma, d\} < J_\gamma(f_c) = \log 2.$$

Positive cost removes the constant predictor from the minimizers in this example. The expected fixed-mixture loss is likewise $\log 2 - (1 - \lambda)ad$ for f_r, so any $\lambda \in [0, 1)$ also favors it. The example isolates indifference under zero-cost robustness and separates deletion invariance from predictive quality. The additional predictive effect of thresholding relative to fixed mixing is evaluated separately in Appendix L.8.

Thresholding can select a different predictor from every fixed mixture. The distinction extends beyond unequal scalar penalties. Consider two equiprobable binary examples, $(X, Y) = ((0, 0), 0)$ and $((1, 1), 1)$, with both features nominally visible and unit deletion costs. In a restricted class of three mask-aware predictors, prescribe the following finite log losses; both single-feature deletion branches have the indicated increase at each example:

Predictor	Nominal loss at either example	Increase at $Y = 0$	Increase at $Y = 1$
f_A	$1/5$	1	0
f_B	$3/10$	$2/5$	$2/5$
f_C	$2/5$	0	0

These losses are realizable by assigning probability $e^{-\ell}$ to the indicated label at each reachable masked input: either remaining feature identifies the example. The all-missing output is arbitrary and nondegenerate, since it is outside this local neighborhood. The nominal risks are $(1/5, 3/10, 2/5)$ and mean positive sensitivities are $(1/2, 2/5, 0)$. Writing $F_\lambda = R + \lambda \mathbb{E}T$, for $0 \leq \lambda \leq 2/5$ we have $F_\lambda(f_B) - F_\lambda(f_A) = (1 - \lambda)/10 \geq 3/50$; for $2/5 \leq \lambda \leq 1$, we have $F_\lambda(f_B) - F_\lambda(f_C) = -1/10 + 2\lambda/5 \geq 3/50$. Thus no fixed mixture selects f_B in this class. At $\gamma = 2/5$, the thresholded risks are $(1/2, 3/10, 2/5)$, making f_B uniquely optimal; this remains true for $3/10 < \gamma < 4/5$. Thresholding depends on the sensitivity distribution beyond its mean. The example separates minimizing predictors under a restricted class; f_A still has the lowest nominal risk, and all three have zero nominal classification error. The empirical comparisons evaluate when this different selection mechanism improves prediction.

G Scope of the local ambiguity set

The local cost is at least the full cost pointwise: $c_{\mathrm{loc},w} \geq c_w$. Therefore $W_{c_{\mathrm{loc},w}} \geq W_{c_w}$, and the local radius ambiguity set is contained in the full one. Restricting the pointwise adversary also gives

$$\phi_{\gamma,w} \leq \phi^{\mathrm{full}}_{\gamma,w}, \qquad J_{\gamma,w} \leq J^{\mathrm{full}}_{\gamma,w}.$$

The distinction between local and simultaneous deletion can be seen in a two-feature example: take two visible features, $w_1 = w_2 = 1$, and losses $\ell_0 = \ell_1 = \ell_2 = a > 0$ and $\ell_{12} = a + M$ with $M > 0$. Then

$$\phi_{\gamma,w} = a, \qquad \phi^{\mathrm{full}}_{\gamma,w} = a + (M - 2\gamma)_+.$$

The gap grows without bound as M increases. The local objective thus gives an exact characterization of its chosen uncertainty set, while simultaneous deletion introduces additional feature interactions. Equation (11) is a population-risk bound for that local ambiguity set and nominal distribution. The unit-cost identity in Appendix F characterizes radius risk at fixed parameters. Training at a prescribed radius would additionally optimize the threshold; finite-sample guarantees require separate estimation analysis.

Which mask shifts admit finite transport cost? An upward-closed set $A \subseteq \{0,1\}^D$ satisfies $m \in A,\ m' \succeq m \Rightarrow m' \in A$. For either deletion cost in this paper, every Q with $W_c(Q, P_0) < \infty$ necessarily satisfies

$$Q(M \in A) \geq P_0(M \in A) \quad \text{for every upward-closed } A. \tag{23}$$

Indeed, a finite-cost coupling has $M' \succeq M$ almost surely, so $\mathbf{1}_A(M') \geq \mathbf{1}_A(M)$; taking expectations proves the claim. This is a necessary condition, not a sufficient characterization of the local ambiguity set.

In particular, the all-missing mask cannot lose probability under a feasible transport. The eleven-rate nominal mixture in Appendix K assigns it probability $1/11$. On the 50-feature Gaussian benchmark, independent Bernoulli(0.5) test deletion assigns it probability 2^{-50}, and the MAR test rule with one always-visible feature assigns it probability zero. Both test laws therefore lie outside the finite-cost domain of even the full deletion cost centered at that nominal mixture, irrespective of the radius. Their empirical results therefore evaluate transfer to external mask shifts, complementing the guarantee within the local domain of Eq. (11). Monotone coverage, the delete-one constraint, and finite-sample estimation are separate aspects of this scope.

H Sampled-candidate surrogate

For the candidate-sampling experiments, let $C \subseteq V(m)$ contain uniformly sampled visible features without replacement, with $|C| = \min\{K_{\text{inner}}, |V(m)|\}$. The no-op branch remains available, including when C is empty. Conditional on $Z = (x, y, m)$, the candidate-sampling law is independent of θ. The sampled surrogate is

$$\phi_{\gamma,w,C}(\theta; Z) = \max\left\{\ell_0(\theta), \max_{j \in C}[\ell_\theta(x, y, m^{+j}) - \gamma w_j]\right\}, \tag{24}$$

with the same empty-set convention as the full local maximum. Pointwise,

$$\ell_0 \leq \phi_{\gamma,w,C} \leq \phi_{\gamma,w}.$$

Under the conditions of Corollary E.1, now applied to (Z, C), exact active-branch selection within C gives stochastic branch subgradients of $\widetilde{J}_{\gamma,w} = \mathbb{E}_{Z\sim P_0}\mathbb{E}_{C|Z}\phi_{\gamma,w,C}$, with $\widetilde{J}_{\gamma,w} \leq J_{\gamma,w}$. These are subgradients of the sampled objective $\widetilde{J}_{\gamma,w}$; recovering a bound for $J_{\gamma,w}$ requires controlling the sampling approximation error. In particular, Eq. (11) retains $J_{\gamma,w}$ unless such an error bound is supplied. Including all visible features recovers the full local surrogate exactly.

I Data preprocessing

This section describes preprocessing for the public-benchmark sweeps and the Gaussian benchmark. The splits and preprocessing for the controlled comparisons are specified in Appendices K and M. We represented each dataset as feature tokens paired with a label. For tabular datasets (WDBC, METABRIC, MiniBooNE), each scalar feature was treated as one token. We maintained a modality specification for each feature (numerical vs. categorical), and we used constant imputation to produce dense tensors for batching: missing numerical values were imputed with 0, and missing categorical values were mapped to a designated default category (encoded as 0). Masking operated on a boolean missingness mask (which prevented masked tokens from contributing to the model output), while the imputed feature values remained fixed for efficient computation.

For image datasets (MNIST, CIFAR-10, CIFAR-100), we converted each image into non-overlapping 4×4 patch tokens. Each patch was flattened into a vector token: 16-dimensional for MNIST (grayscale) and 48- dimensional for CIFAR (RGB). Patch extraction and feature normalization were performed offline.

Table 3: Summary of the six public benchmarks. For each dataset, the table lists the number of tokens/features (D), number of classes (C), number of samples, and, when applicable, the image and patch sizes used for patchification.

Dataset	D	C	# Samples	Patchification (image/patch)
WDBC	30	2	569	–
METABRIC	489	7	1,904	–
MiniBooNE	50	2	127,599	–
MNIST	49	10	70,000	28×28 / 4×4
CIFAR-10	64	10	60,000	32×32 / 4×4
CIFAR-100	64	100	60,000	32×32 / 4×4

Splits and reproducibility. For the exploratory public sweeps and Gaussian oracle diagnostics, runs 1–4 use the corresponding outer split seeds 1–4 and training random seeds 1–4. A stratified 20% outer test split is followed by a stratified 25% holdout from the remaining training pool with inner seed 0, giving approximately 60/20/20 train/validation/test proportions. Thus four-run SD includes both partition and optimization variation. These splits are separate from the repeated-partition protocol in Appendix K.

Model and optimization details. For tabular datasets, we used a Transformer encoder over feature tokens with embedding dimension 128, 2 attention heads, and 2 layers. For vision datasets, we used a ViT-style Transformer with 4×4 non-overlapping patches, embedding dimension 128, 2 heads, and 6 layers. Missing tokens/features were excluded from attention. All models used Adam without weight decay and batch size 64; tabular learning rates were 10^{-3} and vision learning rates were 2×10^{-4}. The public-benchmark tables summarize parameter sweeps. The controlled comparisons use the validation-based model selection specified in Appendices K and M.

Exploratory-sweep training and selection. The configured update budgets and validation intervals are listed below. Each run selects a checkpoint by minimum masked-validation cross-entropy, with the later checkpoint retained on ties. Validation averages the eleven missing rates, using 2,048 bootstrap draws per rate and fresh deterministic masks at each check. Checkpoint selection and selection of the displayed positive γ are distinct: the latter maximizes the reported four-run test mixture mean (Table 9). METABRIC uses the family-specific budgets listed below; matched-budget comparisons follow the controlled protocol.

Table 4: Training configurations for the exploratory sweeps. Intervals are optimizer updates between validation checks. ERM baselines include NoMask, Bernoulli, and Exact-k.

Dataset	Method family	Update budget	Validation interval
WDBC	All	10,000	5
METABRIC	ERM baselines	5,000	17
METABRIC	AdvDelete, MaDRO-1/CV	2,000	17
MiniBooNE	All	200,000	1,196
MNIST	All	100,000	656
CIFAR-10/100	All	100,000	562
Gaussian oracle studies	All	5,000	17

METABRIC-derived Gaussian synthetic benchmark. For the Gaussian synthetic benchmark, we started from the original METABRIC RNA cohort, standardized the 489 numerical covariates, and projected them to a 50-dimensional PCA representation. Each score coordinate is oriented so that its largest-magnitude source-cohort score is positive. The same oriented class means and covariances define both the synthetic sampling distribution and the Gaussian oracle. We collapsed the subtype labels into two superclass labels: class 0 combined LumA, LumB, Normal, and NC; class 1 combined Basal, Her2, and claudin-low. In the PCA space, we fit one Gaussian model per superclass and resampled a balanced dataset with 952 examples per class, for 1,904 examples total. The generator uses seed 0 for PCA, Gaussian sampling, and the final row shuffle. Each class covariance is the sample covariance with denominator $n_c - 1$, regularized by $10^{-6}I$ before sampling. The first 50 PCs explain 56.45% of the source-cohort variance. PCA and the class distributions are fitted to the source

cohort to define the synthetic data-generating distribution; classifier train/validation/test splits are subsequently formed from the generated rows. The resulting benchmark assesses learning under the specified Gaussian distribution; the source cohort defines its generator.

For the four-run oracle diagnostic study in Figure 2, the structured-missingness variants reused the same base examples and labels. The first five principal components were retained as always-observed anchors. The remaining 45 components were masked to target 50% average missingness among these 45 components (approximately 45% of all 50 features) under three mechanisms: MCAR used value-independent missingness with anchor-linked random effects to induce co-missingness; MAR made missingness depend on anchor-derived linear predictions; MNAR made missingness depend on each feature's own standardized value, again with the same co-missingness structure.

J Additional Experimental Details

J.1 Public benchmark supplementary results

Figure 3 shows the rate-wise curves. The tables below give mixture summaries and complete parameter sweeps.

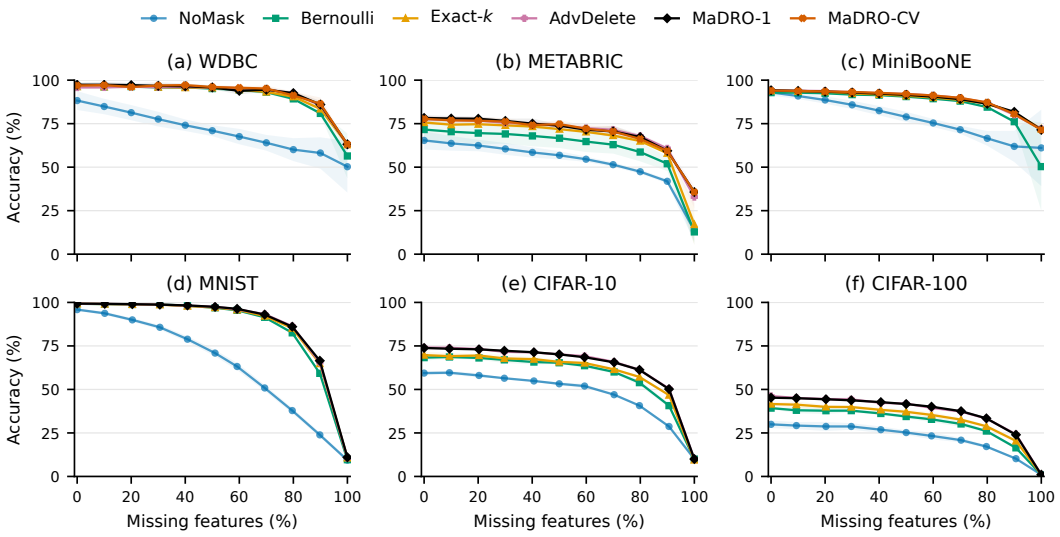

Figure 3: Exploratory accuracy across eleven missing rates on six public benchmarks. Lines and bands show four-run means and one sample SD on linear axes. MaDRO-1/CV and AdvDelete use the Exact-k base rule; CV is restricted to numerical features. Positive-cost configurations maximize test mixture accuracy within each reported sweep; the curves summarize test-selected parameter sweeps. Full sweeps and training budgets appear in Appendices I–J.

Public benchmark mixture evaluation. Table 5 gives the public benchmark mixture evaluation under $p \sim \pi_{\text{eval}}$. Exact-k+AdvDelete is the $\gamma = 0$ special case of the same finite no-op/delete-one local objective. It is reported separately because deletion costs do not affect branch selection when $\gamma = 0$. Exact-k+MaDRO-1 is evaluated on all six datasets. Exact-k+MaDRO-CV is evaluated on the three tabular public datasets and marked N/A on tokenized vision datasets because no patch-level conditional-variance cost model is defined.

Table 5: Mixture evaluation under $p \sim \pi_{\text{eval}}$ using exact-k masks. Exact-k+MaDRO-1 and Exact-k+MaDRO-CV use deletion-penalty scales $\gamma > 0$ from $\{0.05, 0.1, 0.2, 0.5, 1.0\}$; Table 7 and Table 8 give the corresponding full sweeps.

Tabular datasets

Method	WDBC	METABRIC	MiniBooNE
NoMask	0.7065 ± 0.0254	0.5237 ± 0.0284	0.7782 ± 0.0129
Bernoulli	0.8997 ± 0.0146	0.6058 ± 0.0559	0.8545 ± 0.0209
Exact-k	0.9100 ± 0.0084	0.6569 ± 0.0094	0.8860 ± 0.0005
Exact-k+AdvDelete	0.9152 ± 0.0024	0.6920 ± 0.0093	$\mathbf{0.8904 \pm 0.0032}$
Exact-k+MaDRO-1	0.9191 ± 0.0078	$\mathbf{0.6945 \pm 0.0089}$	0.8886 ± 0.0008
Exact-k+MaDRO-CV	$\mathbf{0.9199 \pm 0.0046}$	0.6911 ± 0.0083	$\mathbf{0.8904 \pm 0.0013}$

Vision datasets

Method	MNIST	CIFAR-10	CIFAR-100
NoMask	0.6365 ± 0.0118	0.4727 ± 0.0029	0.2193 ± 0.0145
Bernoulli	0.8448 ± 0.0025	0.5739 ± 0.0108	0.3000 ± 0.0104
Exact-k	0.8542 ± 0.0013	0.5904 ± 0.0086	0.3240 ± 0.0097
Exact-k+AdvDelete	0.8578 ± 0.0010	$\mathbf{0.6281 \pm 0.0052}$	$\mathbf{0.3631 \pm 0.0059}$
Exact-k+MaDRO-1	$\mathbf{0.8589 \pm 0.0013}$	0.6268 ± 0.0032	0.3623 ± 0.0026
Exact-k+MaDRO-CV	N/A	N/A	N/A

Public Exact-k+MaDRO-1 γ sweeps. Table 6 and Table 7 give the four-seed Exact-k+MaDRO-1 γ sweep. The $\gamma = 0$ row gives Exact-k+AdvDelete; $\gamma > 0$ Exact-k+MaDRO-1 runs use the grid $\{0.05, 0.1, 0.2, 0.5, 1.0\}$.

Table 6: Public Exact-k+MaDRO-1 γ sweep: complete-data accuracy.

Tabular datasets

γ	WDBC	METABRIC	MiniBooNE
0	0.9579 ± 0.0037	0.7769 ± 0.0161	0.9409 ± 0.0012
0.05	0.9583 ± 0.0044	0.7723 ± 0.0172	0.9388 ± 0.0006
0.1	0.9715 ± 0.0044	0.7749 ± 0.0195	0.9380 ± 0.0013
0.2	$\mathbf{0.9737 \pm 0.0124}$	$\mathbf{0.7828 \pm 0.0162}$	0.9376 ± 0.0014
0.5	0.9649 ± 0.0072	0.7756 ± 0.0152	0.9318 ± 0.0114
1.0	$\mathbf{0.9737 \pm 0.0072}$	0.7664 ± 0.0207	$\mathbf{0.9415 \pm 0.0008}$

Vision datasets

γ	MNIST	CIFAR-10	CIFAR-100
0	$\mathbf{0.9933 \pm 0.0010}$	$\mathbf{0.7408 \pm 0.0180}$	$\mathbf{0.4615 \pm 0.0045}$
0.05	0.9931 ± 0.0006	0.7384 ± 0.0054	0.4526 ± 0.0022
0.1	0.9927 ± 0.0004	0.7374 ± 0.0038	0.4513 ± 0.0046
0.2	0.9927 ± 0.0004	0.7346 ± 0.0059	0.4478 ± 0.0055
0.5	0.9926 ± 0.0004	0.7201 ± 0.0103	0.4364 ± 0.0113
1.0	0.9918 ± 0.0002	0.7152 ± 0.0074	0.4220 ± 0.0116

Table 7: Public Exact-k+MaDRO-1 γ sweep: average accuracy over missing rates 0:0.1:1.0.

Tabular datasets

γ	WDBC	METABRIC	MiniBooNE
0	0.9152 ± 0.0024	0.6920 ± 0.0093	$\mathbf{0.8904 \pm 0.0032}$
0.05	0.9155 ± 0.0076	0.6921 ± 0.0114	0.8883 ± 0.0022
0.1	0.9173 ± 0.0047	0.6864 ± 0.0120	0.8885 ± 0.0014
0.2	$\mathbf{0.9191 \pm 0.0078}$	$\mathbf{0.6945 \pm 0.0089}$	0.8880 ± 0.0015
0.5	0.9147 ± 0.0081	0.6878 ± 0.0085	0.8812 ± 0.0105
1.0	0.9137 ± 0.0056	0.6830 ± 0.0128	0.8886 ± 0.0008

Vision datasets

γ	MNIST	CIFAR-10	CIFAR-100
0	0.8578 ± 0.0010	$\mathbf{0.6281 \pm 0.0052}$	$\mathbf{0.3631 \pm 0.0059}$
0.05	$\mathbf{0.8589 \pm 0.0013}$	0.6268 ± 0.0032	0.3623 ± 0.0026
0.1	0.8583 ± 0.0013	0.6260 ± 0.0048	0.3608 ± 0.0041
0.2	0.8580 ± 0.0008	0.6230 ± 0.0062	0.3561 ± 0.0049
0.5	0.8561 ± 0.0018	0.6091 ± 0.0077	0.3438 ± 0.0081
1.0	0.8542 ± 0.0014	0.6031 ± 0.0059	0.3320 ± 0.0094

Tabular public Exact-k+MaDRO-CV γ sweeps. Table 8 gives the full tabular public Exact-k+MaDRO-CV sweep. These datasets have numerical covariates, so conditional-variance deletion costs are defined directly.

Table 8: Tabular public Exact-k+MaDRO-CV γ sweep. The $\gamma = 0$ row gives Exact-k+AdvDelete, for which deletion costs do not affect branch selection.

γ	WDBC	METABRIC	MiniBooNE
0	0.9152 ± 0.0024	$\mathbf{0.6920 \pm 0.0093}$	$\mathbf{0.8904 \pm 0.0032}$
0.05	0.9171 ± 0.0076	0.6911 ± 0.0083	$\mathbf{0.8904 \pm 0.0013}$
0.1	$\mathbf{0.9199 \pm 0.0046}$	0.6883 ± 0.0058	0.8899 ± 0.0005
0.2	0.9195 ± 0.0085	0.6828 ± 0.0126	0.8900 ± 0.0010
0.5	0.9117 ± 0.0053	0.6830 ± 0.0197	0.8890 ± 0.0012
1.0	0.9107 ± 0.0092	0.6776 ± 0.0115	0.8898 ± 0.0007

Deletion-penalty settings in the public benchmark sweeps. Table 9 lists the positive deletion-penalty settings used for the displayed public-benchmark entries. Tables 7 and 8 report the full parameter sweeps. The displayed positive-cost configurations maximize the reported four-run test mixture mean over the positive-cost grid. These test-selected summaries describe parameter sensitivity; the primary validation-selected comparisons follow Appendix K.

Table 9: Positive deletion-penalty settings used for the displayed public-benchmark results; the full parameter sweeps are reported above.

Method	Dataset	γ	Test mixture accuracy
Exact-k+MaDRO-1	WDBC	0.2	0.9191 ± 0.0078
	METABRIC	0.2	0.6945 ± 0.0089
	MiniBooNE	1.0	0.8886 ± 0.0008
	MNIST	0.05	0.8589 ± 0.0013
	CIFAR-10	0.05	0.6268 ± 0.0032
	CIFAR-100	0.05	0.3623 ± 0.0026
Exact-k+MaDRO-CV	WDBC	0.1	0.9199 ± 0.0046
	METABRIC	0.05	0.6911 ± 0.0083
	MiniBooNE	0.05	0.8904 ± 0.0013
	MNIST	N/A	N/A
	CIFAR-10	N/A	N/A
	CIFAR-100	N/A	N/A

J.2 Training and validation trajectories

Figure 10 reports training-probe and validation cross-entropy for the four-seed single-partition controls. Complete datasets use matched training and evaluation mask laws, while naturally incomplete datasets retain intrinsic missingness. Fixed evaluation masks and a fixed training probe make trajectories comparable across checkpoints. Checkpoints are selected by minimum validation loss, and configuration selection is completed before test evaluation, as specified in Appendix M. Figures 8 and 9 provide the corresponding repeated-partition, extended-horizon diagnostics. The curves characterize optimization and validation behavior under the specified masking laws.

J.3 Gaussian synthetic supplementary results

Gaussian synthetic γ sweeps. Table 10 gives the Exact-k+MaDRO-CV sweep used in Section 5.3. Accuracy averaged over 0:0.1:1.0 is the mean across exact-k evaluation masks with the corresponding observed-feature fractions. Values are means over four runs, with standard deviations across runs.

Table 10: Exact-k+MaDRO-CV γ sweep on Gaussian synthetic data.

γ	Complete accuracy	Average accuracy over 0:0.1:1.0
0	0.8681 ± 0.0148	0.7259 ± 0.0093
0.05	0.9587 ± 0.0184	0.7840 ± 0.0085
0.1	0.9580 ± 0.0173	$\mathbf{0.7924 \pm 0.0083}$
0.2	$\mathbf{0.9652 \pm 0.0084}$	0.7917 ± 0.0052
0.5	0.9514 ± 0.0099	0.7885 ± 0.0028
1.0	0.9455 ± 0.0141	0.7832 ± 0.0094

Table 11: Exact-k+MaDRO-1 γ sweep on Gaussian synthetic data. The $\gamma = 0$ row gives Exact-k+AdvDelete.

γ	Complete accuracy	Average accuracy over 0:0.1:1.0
0	0.8681 ± 0.0148	0.7259 ± 0.0093
0.05	0.9501 ± 0.0113	0.7826 ± 0.0063
0.1	0.9508 ± 0.0175	0.7850 ± 0.0105
0.2	$\mathbf{0.9547 \pm 0.0058}$	$\mathbf{0.7907 \pm 0.0072}$
0.5	0.9488 ± 0.0026	0.7851 ± 0.0094
1.0	0.9501 ± 0.0184	0.7808 ± 0.0112

Gaussian synthetic exact-k accuracy. Table 12 gives selected values from the controlled-deletion evaluation on the Gaussian synthetic benchmark. The first three columns show accuracy when 5, 25, or 50 of the 50 PCA features are observed. The final column averages over all eleven observed-feature fractions.

Table 12: Selected Gaussian synthetic exact-k evaluation accuracies.

Method	5 observed	25 observed	50 observed	Avg.
NoMask	0.5440	0.7021	0.8688	0.6973
NoMask+MaDRO-CV	0.5886	0.7684	0.9199	0.7462
Bernoulli	0.6240	0.7940	0.9449	0.7757
Bernoulli+MaDRO-CV	0.6214	**0.8274**	0.9619	**0.7972**
Exact-k	0.6129	0.8104	0.9429	0.7768
Exact-k+MaDRO-1	**0.6260**	0.8176	0.9547	0.7907
Exact-k+MaDRO-CV	0.6010	0.8150	**0.9652**	0.7917

Structured missingness. Table 13 reports the complete structured-missingness stress test from Section 5.3. The underlying Gaussian synthetic covariates are complete during training, but each method still uses its specified base masking rule during training. Complete, MCAR, MAR, and MNAR masks are used only for held-out evaluation. Within each run, the same checkpoint and held-out rows are used for all four conditions, without additional random masking at evaluation.

Table 13: Structured test-time missingness on Gaussian synthetic data. Entries are mean accuracies over four seeds, using complete covariates before applying each method's training mask.

Method	Complete	MCAR	MAR	MNAR
NoMask	0.8688	0.8681	0.8694	0.8694
NoMask+MaDRO-CV	0.9199	0.8885	0.8970	0.8957
Bernoulli	0.9449	0.9199	0.9160	0.9134
Bernoulli+MaDRO-CV	0.9619	0.9252	**0.9272**	0.9173
Exact-k	0.9429	0.9232	0.9147	0.9219
Exact-k+AdvDelete	0.8681	0.9003	0.8812	0.8819
Exact-k+MaDRO-1	0.9547	**0.9259**	0.9199	0.9285
Exact-k+MaDRO-CV	**0.9652**	0.9232	0.9206	**0.9298**

Oracle probabilities and attribution protocol. For an observed coalition S, the oracle value is the Bayes posterior under the balanced Gaussian generator:

$$v_x(S) = \frac{\mathcal{N}(x_S; \mu_{1,S}, \Sigma_{1,SS})}{\mathcal{N}(x_S; \mu_{0,S}, \Sigma_{0,SS}) + \mathcal{N}(x_S; \mu_{1,S}, \Sigma_{1,SS})}, \qquad v_x(\emptyset) = \tfrac{1}{2}.$$

The means and regularized covariances are those used to generate the data, described in Appendix I. Model values are predicted positive-class probabilities using only the coalition's visible features. Both model and oracle use the same 512 uniformly sampled permutations of all 50 features, generated with NumPy random seed 20260317. For each target feature, we average its marginal contribution over these permutations. The complete-data test subsets contain 191, 190, 190, and 190 positive-class examples for runs 1–4, respectively; these subsets can overlap across runs. MAE averages absolute model–oracle attribution differences over examples and PC1–PC5, followed by an unweighted mean across runs. The reported uncertainty is run-to-run dispersion, not a Monte Carlo error estimate or an

independent-sample confidence interval. The comparison measures model attribution fidelity relative to the known Gaussian oracle.

Shapley-fidelity summary. Table 14 compares Monte Carlo Shapley estimates for PC1–PC5 on held-out positive-class examples. Both model and Gaussian-oracle coalition probabilities are evaluated along 512 sampled permutations of all 50 features; the reported values approximate the Shapley sums by permutation averaging. MAE averages the five reported components, and top-1 agreement compares the largest absolute attribution among those five. Lower MAE is better; higher agreement is better. The class, feature subset, and run-level uncertainty follow the protocol above.

Table 14: Monte Carlo Shapley fidelity on positive-class held-out Gaussian examples, restricted to reported PC1–PC5 components.

Method	Shapley MAE	Top-1 agreement
NoMask	0.1011 ± 0.0292	0.758 ± 0.025
NoMask+MaDRO-CV	0.0610 ± 0.0126	0.811 ± 0.015
Bernoulli	0.0572 ± 0.0114	0.853 ± 0.024
Bernoulli+MaDRO-CV	0.0378 ± 0.0028	0.850 ± 0.033
Exact-k	0.0523 ± 0.0033	0.857 ± 0.009
Exact-k+AdvDelete	0.0466 ± 0.0054	0.829 ± 0.033
Exact-k+MaDRO-1	$\mathbf{0.0323 \pm 0.0019}$	$\mathbf{0.871 \pm 0.020}$
Exact-k+MaDRO-CV	0.0332 ± 0.0037	0.850 ± 0.014

Per-PC attribution errors. Table 15 gives the per-PC MAEs.

Table 15: Per-PC Shapley MAE on Gaussian synthetic data.

Method	PC1	PC2	PC3	PC4	PC5	Avg.
NoMask	0.0438	0.3130	0.0911	0.0204	0.0371	0.1011
Bernoulli	0.0256	0.1592	0.0503	0.0217	0.0290	0.0572
Exact-k	0.0204	0.1644	0.0354	0.0142	0.0269	0.0523
Exact-k+AdvDelete	0.0238	0.1297	0.0333	0.0153	0.0308	0.0466
Exact-k+MaDRO-1	**0.0179**	**0.0870**	0.0244	0.0125	0.0197	**0.0323**
Exact-k+MaDRO-CV	0.0183	0.0935	**0.0232**	**0.0117**	**0.0192**	0.0332

Sampled inner maximization. Table 16 gives the sampled-inner-max ablation used in Section 5.4. The experiment uses Exact-k+MaDRO-CV with $\gamma = 0.2$. Relative training time is normalized by the full visible-feature inner maximization.

Table 16: Sampled inner maximization for Exact-k+MaDRO-CV on Gaussian synthetic data.

Inner candidates	Relative train time	Complete accuracy	Average accuracy over 0:0.1:1.0
$K_{\text{inner}} = 4$	**41%**	0.9455	0.7774
$K_{\text{inner}} = 8$	42%	0.9469	0.7856
$K_{\text{inner}} = 16$	45%	0.9554	0.7907
$K_{\text{inner}} = 32$	53%	**0.9652**	**0.7946**
$K_{\text{inner}} = 50$ full	100%	**0.9652**	0.7917

Table 17: Public benchmark sampled inner maximization on MiniBooNE with Exact-k+MaDRO-1 at $\gamma = 0.2$. Relative training time is omitted because the sampled runs were split across heterogeneous hardware.

Dataset	Method	Inner candidates	Relative train time	Mixture accuracy
MiniBooNE	Exact-k+MaDRO-1	$K_{\text{inner}} = 4$	N/A	0.8892 ± 0.0015
MiniBooNE	Exact-k+MaDRO-1	$K_{\text{inner}} = 8$	N/A	$\mathbf{0.8904 \pm 0.0017}$
MiniBooNE	Exact-k+MaDRO-1	$K_{\text{inner}} = 16$	N/A	0.8900 ± 0.0010
MiniBooNE	Exact-k+MaDRO-1	$K_{\text{inner}} = 32$	N/A	0.8899 ± 0.0007
MiniBooNE	Exact-k+MaDRO-1	full	N/A	0.8880 ± 0.0015

K CONTROLLED COMPARISONS ACROSS TRAINING HORIZONS

Protocol map. The comparisons below answer different questions and retain their own selection and uncertainty conventions. The numerical primary comparison integrates 100 test-mask banks; auxiliary numerical controls retain their original single bank. Natural-data primary evaluation uses intrinsic missingness.

Comparison	Replication	Selection	Reported variation
Primary numerical (App. L)	5 partitions, 2 seeds, 100 mask banks	Validation checkpoint and configuration	Partition SD; conditional mask MCSE
Natural missingness (this protocol)	5 partitions, 2 seeds	Validation checkpoint and configuration	Partition SD
Optimizer / shuffled cost (this protocol)	1 partition, 2 seeds, 1 mask bank	Validation checkpoint and configuration	Optimization-seed SD
L4 controls (Apps. M, N)	1 partition, 4 seeds	Validation checkpoint and configuration	Optimization-seed SD
Public sweeps (Apps. I, J)	4 runs per setting	Validation checkpoint; test-selected displayed cost	Run SD

Data and repeated partitions. We compare methods on WDBC, the 50-dimensional Gaussian semi-synthetic benchmark, and two datasets with naturally occurring missingness: UCI Credit Approval and Horse Colic (Quinlan, 1987; McLeish & Cecile, 1989). Credit Approval retains its 15 predictors and original missing values. For Horse Colic, we pool the two source files, discard records with unknown outcome, and predict survival versus death/euthanasia. Predictors are age and original fields 4–22; surgery status, hospital identifier, outcome, and retrospective lesion/pathology fields are excluded. Repeated hospital identifiers stay in the same partition. Other datasets group identical predictor rows. Five outer seeds, 2027–2031, and inner seed 0 define approximately 60/20/20 training/validation/test partitions. We stratify by label when all groups are singletons, and otherwise use disjoint group splits. Numerical normalization, categorical vocabularies, and covariance-based costs are fitted independently using only each partition's training rows. Original missing entries remain explicitly masked. Unseen categories receive an unknown code and are distinguished from missing features.

Shared architecture and controls. All neural methods use the same TaskNet feature encoder with dimension 128, two Transformer layers, two heads, no internal dropout, and a binary linear output head. The primary comparison uses Adam with learning rate 10^{-3}, zero weight decay, batch size 64, and optimization seeds 1 and 2 for every partition. ERM uses the base mask without a further deletion. AdvDelete maximizes loss over no-op and each single-feature deletion. MaDRO-1 uses unit deletion costs; MaDRO-CV uses training-only conditional-variance costs on numerical data. Both use $\gamma \in \{0.05, 0.1, 0.2, 0.5, 1\}$; AdvDelete is the separate zero-cost control. On Credit and Horse we additionally compare RandomDelete, ordinary hidden dropout, and the binary-linear-head SAdD adaptation described in Appendix M. RandomDelete removes one uniformly sampled visible feature, with no change if none remains. CV costs are not extended to these mixed categorical datasets.

Predictor implementation. Numerical tokens use feature-specific affine maps $1 \to 128$, and categorical tokens use independent lookup tables, each followed by LayerNorm. Fixed sinusoidal encodings in column order use frequencies $10000^{-2r/128}$ for sine/cosine pairs $r = 0, \ldots, 63$. A learned zero-initialized CLS token is then prepended. Transformer blocks use 512-dimensional feed-forward layers, GELU, post-residual LayerNorm, batch-first tensors, and no terminal LayerNorm. Missing tokens are excluded as attention keys; the visible CLS output feeds a biased $128 \to 1$ linear

classifier. Runs use PyTorch 2.7.0+cu128 defaults except for these settings and the dimensions, heads, layers and dropout specified above, on Tesla T4.

Base masks and evaluation. For WDBC and Gaussian data, training samples a rate index uniformly from $p \in \{0, 0.1, \ldots, 1\}$, then samples exactly $D - \mathrm{round}(D(1-p))$ missing features uniformly. Separate random streams for example order, base masks, and method-specific augmentation keep examples and base masks aligned across methods. Natural-data training, validation, and primary testing retain intrinsic missingness. Their additional-deletion curves are secondary stress tests; zero additional deletion is the primary evaluation. All configurations and checkpoints are selected using validation data and held fixed across test masking rates.

Nominal metric estimator. Validation assigns one uniformly sampled exact-size mask to each example at each rate. A device-specific PyTorch generator is reset to seed 701 at every call, with fixed row and rate order, so checkpoints and configurations share validation masks. The numerical primary test comparison integrates over 100 independent mask replicates at the already selected predictors. Each replicate assigns one exact-size mask per example and rate; all methods, optimization seeds and horizons share the same bank within a partition. At each rate we compute mean per-example binary cross-entropy, accuracy at probability threshold 0.5, balanced accuracy and AUROC, then average the eleven rate-specific metrics equally. AUROC is averaged across rates rather than after pooling. At full missingness only the learned CLS token remains visible, yielding an input-independent prediction.

Mask replication and precision. Test banks use NumPy PCG64 with float64 random ranks, in saved test-row order and ascending rate order. For dataset identifier `cancer` or `semisynth`, partition s and replicate $r = 0, \ldots, 99$, the seed is the first eight SHA-256 bytes of `ICLR2027/frozen-selected-mask/v1/<dataset>/<s>/<r>`, interpreted as an unsigned little-endian integer. At each rate, including endpoints, an $n_{\mathrm{test}} \times D$ rank matrix is drawn and the prescribed number of lowest-ranked features is masked. Deterministic complete/all-missing predictions are shared across replicates. For each replicate we average optimization seeds within partitions and then partitions; the SD of the resulting 100 means divided by $\sqrt{100}$ is the conditional mask Monte Carlo standard error (MCSE). Paired errors apply this operation to common-mask method differences. Partition SD instead averages all masks and seeds first, then summarizes the five partition means. Appendix L.10 reports all families, horizons and paired comparisons. The single-realization sensitivity comparison, rate curves, and separate numerical optimizer, shuffled-cost and four-seed diagnostics use the original PyTorch test stream, reset to seed 202709; they remain distinct from the replicated primary estimates. Training and validation, including all selection rules, are held fixed.

Three horizons from one trajectory. The short horizon uses 3,000 updates for WDBC/Gaussian and 2,000 for Credit/Horse. The equal-time horizon allows 300 seconds of synchronized training computation per numerical configuration and 120 seconds per natural-data configuration. Validation and checkpoint I/O are excluded uniformly; the final update can overshoot by one step. All runs use NVIDIA T4 GPUs with one training process per GPU. Extended training stops after at least 15,000 updates on numerical data when the running minimum validation cross-entropy has not improved by more than 0.001 for 10,000 updates, subject to a 60,000-update cap. Natural-data thresholds are 3,000 minimum updates, 2,000-update patience, and a 20,000-update cap. Improvements are measured cumulatively from the last patience anchor. Cap-reaching runs are reported separately and are not assumed to have converged. If the time budget has not been reached when extended training stops, the trajectory continues only to complete that budget; the extended endpoint remains fixed.

Checkpoint and configuration selection. The shared validation grid includes update 0, every 25 updates through 500, the union of multiples of 100 and 150 through 3,000, and every 500 thereafter. Short and extended selection use this common grid. The time horizon also includes its exact final update; that extra checkpoint does not affect short/extended selection or patience. Each horizon first selects the checkpoint with minimum validation cross-entropy for each configuration. Within each method family, it then selects the configuration with maximum validation accuracy at that checkpoint; ties use lower validation loss and then configuration name. Consequently, each horizon has its own validation-selected configuration. Longer horizons retain all earlier eligible checkpoints. Equal-time budgets apply per configuration: five positive-cost candidates per MaDRO family and four SAdD

candidates entail larger total searches than ERM or AdvDelete. We report full family search costs, and do not sum costs across the three horizons, which share a training trajectory.

Cost and optimizer controls. For numerical data we use the ridge-regularized precision $\widehat{\Omega} = (\widehat{\Sigma} + 10^{-4}I)^{-1}$ and costs $u_j = 1/\widehat{\Omega}_{jj}$ from Eq. (9). Costs are divided by their mean, clipped to $[0.5, 2]$, and divided by their mean again. The final normalized costs need not remain within the clipping interval. We report their dispersion and the regularized covariance's condition number. On outer split 2027 with optimization seeds 1 and 2, a shuffled-CV control permutes costs using fixed seed 317 while preserving their distribution. A separate common-optimizer comparison considers $(\mathrm{lr}, \mathrm{wd}) \in \{(10^{-3}, 0), (3\times10^{-4}, 0), (10^{-3}, 10^{-2})\}$ for ERM, AdvDelete, every positive-γ configuration of MaDRO-1/CV, and every fixed-mixture weight $\lambda \in \{0, .25, .5, .75, 1\}$. Thus ERM and AdvDelete each search three settings, and each MaDRO family and the mixture search fifteen. Mixture endpoints reuse the corresponding ERM/AdvDelete fits and count toward its full search cost. Joint optimizer/configuration selection uses the same validation rule: highest validation accuracy at the validation-loss-selected checkpoint, then lower validation loss, optimizer index, and configuration name to break ties. Every candidate enters search-cost accounting. These two-seed controls retain the original shared test-mask realization (seed 202709) and are reported separately from the five-partition, 100-mask primary comparison.

Natural-data tree baseline. HistGradientBoosting uses the same partitions and training-only preprocessing, retaining numerical NaNs and explicitly marking categorical features. We select among learning rates $\{0.03, 0.1\}$ and maximum leaf counts $\{7, 15, 31\}$, using 300 iterations with early stopping disabled. Validation accuracy, then cross-entropy, selects the configuration. This deterministic baseline is fitted once per configuration and partition, outside the neural per-configuration timing comparison.

Structured masks and statistical summaries. Gaussian checkpoints selected under the nominal rate mixture are also evaluated with no missing features and three fixed stress masks: independent Bernoulli(0.5) deletion (MCAR); deletion probabilities given by the sigmoid of the first standardized feature, which itself stays visible (MAR); and deletion probabilities given by the sigmoid of each feature's own standardized value (MNAR). Shared mask seed 202711 aligns conditions across methods without retuning. Numerical primary comparisons first average 100 mask replicates and two optimization seeds within each partition; natural-data comparisons average the two seeds under intrinsic missingness. We then report the mean and descriptive sample SD across five overlapping partitions, including paired method differences. This dispersion describes partition sensitivity and is not a population confidence interval. The structured-mask comparisons retain their fixed stress-mask realizations. Additional optimization seeds 3 and 4 on split 2027 form a separate four-seed comparison. We report accuracy, balanced accuracy, and AUROC. The exploratory public-benchmark sweeps use the separate protocol in Appendix J.

L Results for repeated-partition comparisons

All results in this section use the protocol in Appendix K. Numerical primary entries average 100 common test-mask replicates and two optimization seeds within each of five partitions before computing the mean and sample standard deviation; natural-data entries average the two seeds under intrinsic missingness. The partitions overlap, so this SD describes partition sensitivity, not a population confidence interval. Accuracy, balanced accuracy, and AUROC are percentages. Positive-cost and SAdD configurations are selected separately at each horizon using validation data only.

Table 18: WDBC: primary metrics across training horizons. Numerical primary tests average the eleven-rate mixture over 100 common mask replicates; natural-data tests retain intrinsic missingness.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	91.95 ± 1.03	90.32 ± 1.39	94.26 ± 0.23
Short	AdvDelete	92.24 ± 1.13	90.59 ± 1.46	94.51 ± 0.25
Short	MaDRO-1	92.36 ± 1.19	90.58 ± 1.66	94.51 ± 0.25
Short	MaDRO-CV	92.29 ± 0.90	90.69 ± 1.18	94.49 ± 0.18
Equal time	ERM	91.78 ± 0.90	90.22 ± 1.29	94.18 ± 0.15
Equal time	AdvDelete	92.24 ± 1.13	90.59 ± 1.46	94.51 ± 0.25
Equal time	MaDRO-1	92.30 ± 1.20	90.46 ± 1.64	94.46 ± 0.25
Equal time	MaDRO-CV	92.31 ± 0.96	90.73 ± 1.25	94.51 ± 0.20
Extended	ERM	91.78 ± 0.90	90.22 ± 1.29	94.18 ± 0.15
Extended	AdvDelete	92.24 ± 1.13	90.59 ± 1.46	94.51 ± 0.25
Extended	MaDRO-1	92.24 ± 1.22	90.52 ± 1.65	94.46 ± 0.25
Extended	MaDRO-CV	92.28 ± 0.95	90.69 ± 1.23	94.50 ± 0.19

Table 19: Gaussian: primary metrics across training horizons. Numerical primary tests average the eleven-rate mixture over 100 common mask replicates; natural-data tests retain intrinsic missingness.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	77.09 ± 0.29	77.09 ± 0.29	83.88 ± 0.36
Short	AdvDelete	75.20 ± 1.29	75.20 ± 1.30	83.01 ± 0.63
Short	MaDRO-1	79.01 ± 0.53	79.01 ± 0.52	85.20 ± 0.44
Short	MaDRO-CV	79.07 ± 0.38	79.08 ± 0.38	85.14 ± 0.21
Equal time	ERM	79.29 ± 0.56	79.29 ± 0.56	85.34 ± 0.48
Equal time	AdvDelete	75.38 ± 1.19	75.38 ± 1.19	83.08 ± 0.60
Equal time	MaDRO-1	79.20 ± 0.39	79.20 ± 0.39	85.33 ± 0.35
Equal time	MaDRO-CV	79.12 ± 0.40	79.12 ± 0.39	85.18 ± 0.26
Extended	ERM	79.48 ± 0.69	79.48 ± 0.69	85.55 ± 0.56
Extended	AdvDelete	78.25 ± 0.41	78.25 ± 0.42	84.86 ± 0.38
Extended	MaDRO-1	80.16 ± 0.55	80.16 ± 0.55	86.07 ± 0.48
Extended	MaDRO-CV	80.29 ± 0.25	80.29 ± 0.25	86.15 ± 0.20

Table 20: Credit: primary metrics across training horizons. Numerical primary tests average the eleven-rate mixture over 100 common mask replicates; natural-data tests retain intrinsic missingness.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	85.51 ± 2.94	85.58 ± 2.89	92.16 ± 1.64
Short	RandomDelete	85.58 ± 1.21	85.58 ± 1.07	91.68 ± 1.62
Short	AdvDelete	84.71 ± 1.76	83.90 ± 2.13	90.87 ± 1.68
Short	MaDRO-1	85.43 ± 1.76	85.35 ± 1.60	92.24 ± 2.00
Short	Dropout	85.87 ± 2.75	85.65 ± 2.94	92.01 ± 2.25
Short	SAdD	84.93 ± 1.91	84.89 ± 1.97	91.45 ± 1.46
Equal time	ERM	85.51 ± 2.94	85.58 ± 2.89	92.16 ± 1.64
Equal time	RandomDelete	85.58 ± 1.21	85.58 ± 1.07	91.68 ± 1.62
Equal time	AdvDelete	84.71 ± 1.76	83.90 ± 2.13	90.87 ± 1.68
Equal time	MaDRO-1	85.43 ± 1.76	85.35 ± 1.60	92.24 ± 2.00
Equal time	Dropout	85.51 ± 2.14	85.34 ± 2.40	92.14 ± 2.47
Equal time	SAdD	84.93 ± 1.91	84.89 ± 1.97	91.45 ± 1.46
Extended	ERM	85.51 ± 2.94	85.58 ± 2.89	92.16 ± 1.64
Extended	RandomDelete	85.58 ± 1.21	85.58 ± 1.07	91.68 ± 1.62
Extended	AdvDelete	84.71 ± 1.76	83.90 ± 2.13	90.87 ± 1.68
Extended	MaDRO-1	85.43 ± 1.76	85.35 ± 1.60	92.24 ± 2.00
Extended	Dropout	85.87 ± 2.75	85.65 ± 2.94	92.01 ± 2.25
Extended	SAdD	84.93 ± 1.91	84.89 ± 1.97	91.45 ± 1.46

Table 21: Horse: primary metrics across training horizons. Numerical primary tests average the eleven-rate mixture over 100 common mask replicates; natural-data tests retain intrinsic missingness.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	72.89 ± 4.19	72.28 ± 4.47	82.30 ± 5.46
Short	RandomDelete	74.64 ± 4.41	73.72 ± 4.28	82.16 ± 4.78
Short	AdvDelete	74.91 ± 5.48	73.87 ± 5.17	84.46 ± 3.26
Short	MaDRO-1	75.39 ± 5.82	74.53 ± 5.80	84.38 ± 3.09
Short	Dropout	71.54 ± 4.20	69.85 ± 5.03	81.51 ± 5.10
Short	SAdD	72.05 ± 2.95	70.63 ± 3.47	81.22 ± 4.17
Equal time	ERM	72.89 ± 4.19	72.28 ± 4.47	82.30 ± 5.46
Equal time	RandomDelete	74.64 ± 4.41	73.72 ± 4.28	82.16 ± 4.78
Equal time	AdvDelete	74.91 ± 5.48	73.87 ± 5.17	84.46 ± 3.26
Equal time	MaDRO-1	75.39 ± 5.82	74.53 ± 5.80	84.38 ± 3.09
Equal time	Dropout	71.54 ± 4.20	69.85 ± 5.03	81.51 ± 5.10
Equal time	SAdD	71.36 ± 4.14	69.71 ± 4.88	79.34 ± 7.90
Extended	ERM	72.89 ± 4.19	72.28 ± 4.47	82.30 ± 5.46
Extended	RandomDelete	74.64 ± 4.41	73.72 ± 4.28	82.16 ± 4.78
Extended	AdvDelete	74.91 ± 5.48	73.87 ± 5.17	84.46 ± 3.26
Extended	MaDRO-1	75.39 ± 5.82	74.53 ± 5.80	84.38 ± 3.09
Extended	Dropout	71.54 ± 4.20	69.85 ± 5.03	81.51 ± 5.10
Extended	SAdD	71.36 ± 4.14	69.71 ± 4.88	79.34 ± 7.90

Table 22: Extended-horizon natural-missingness metrics, including validation-selected native-missing-value boosted trees. HGB uses one deterministic fit per partition.

Dataset	Method	Accuracy	Balanced acc.	AUROC
Credit	ERM	85.51 ± 2.94	85.58 ± 2.89	92.16 ± 1.64
Credit	RandomDelete	85.58 ± 1.21	85.58 ± 1.07	91.68 ± 1.62
Credit	AdvDelete	84.71 ± 1.76	83.90 ± 2.13	90.87 ± 1.68
Credit	MaDRO-1	85.43 ± 1.76	85.35 ± 1.60	92.24 ± 2.00
Credit	Dropout	85.87 ± 2.75	85.65 ± 2.94	92.01 ± 2.25
Credit	SAdD	84.93 ± 1.91	84.89 ± 1.97	91.45 ± 1.46
Credit	HGB	85.94 ± 2.49	85.90 ± 2.43	92.90 ± 1.70
Horse	ERM	72.89 ± 4.19	72.28 ± 4.47	82.30 ± 5.46
Horse	RandomDelete	74.64 ± 4.41	73.72 ± 4.28	82.16 ± 4.78
Horse	AdvDelete	74.91 ± 5.48	73.87 ± 5.17	84.46 ± 3.26
Horse	MaDRO-1	75.39 ± 5.82	74.53 ± 5.80	84.38 ± 3.09
Horse	Dropout	71.54 ± 4.20	69.85 ± 5.03	81.51 ± 5.10
Horse	SAdD	71.36 ± 4.14	69.71 ± 4.88	79.34 ± 7.90
Horse	HGB	77.32 ± 4.54	75.56 ± 4.71	84.07 ± 5.05

L.1 Paired effects and partition sensitivity

Paired differences below subtract methods on the same partition and optimization seed before averaging. They retain both positive and negative outcomes.

Table 23: WDBC: paired extended-horizon differences (percentage points, mean $\pm$ partition SD).

Comparison	Accuracy	Balanced acc.	AUROC
AdvDelete – ERM	0.46 ± 0.31	0.37 ± 0.29	0.33 ± 0.13
MaDRO-1 – ERM	0.46 ± 0.37	0.30 ± 0.50	0.29 ± 0.14
MaDRO-1 – AdvDelete	0.00 ± 0.40	-0.07 ± 0.36	-0.05 ± 0.10
MaDRO-CV – ERM	0.50 ± 0.29	0.47 ± 0.41	0.33 ± 0.06
MaDRO-CV – AdvDelete	0.04 ± 0.31	0.10 ± 0.37	-0.00 ± 0.12
MaDRO-CV – MaDRO-1	0.04 ± 0.42	0.17 ± 0.46	0.04 ± 0.12

Table 24: Gaussian: paired extended-horizon differences (percentage points, mean ± partition SD).

Comparison	Accuracy	Balanced acc.	AUROC
AdvDelete – ERM	-1.23 ± 0.63	-1.22 ± 0.64	-0.69 ± 0.43
MaDRO-1 – ERM	0.68 ± 0.55	0.68 ± 0.54	0.52 ± 0.31
MaDRO-1 – AdvDelete	1.91 ± 0.76	1.91 ± 0.77	1.21 ± 0.57
MaDRO-CV – ERM	0.81 ± 0.59	0.81 ± 0.58	0.60 ± 0.43
MaDRO-CV – AdvDelete	2.04 ± 0.18	2.03 ± 0.19	1.29 ± 0.20
MaDRO-CV – MaDRO-1	0.12 ± 0.61	0.13 ± 0.61	0.08 ± 0.47

Table 25: Credit: paired extended-horizon differences (percentage points, mean ± partition SD).

Comparison	Accuracy	Balanced acc.	AUROC
RandomDelete – ERM	0.07 ± 1.75	-0.00 ± 1.88	-0.48 ± 0.45
AdvDelete – ERM	-0.80 ± 2.53	-1.69 ± 2.86	-1.29 ± 1.20
MaDRO-1 – ERM	-0.07 ± 1.41	-0.24 ± 1.91	0.08 ± 0.57
MaDRO-1 – AdvDelete	0.72 ± 1.49	1.45 ± 1.61	1.37 ± 1.04
Dropout – ERM	0.36 ± 1.72	0.07 ± 1.47	-0.14 ± 0.65
SAdD – ERM	-0.58 ± 1.42	-0.69 ± 1.55	-0.71 ± 0.89

Table 26: Horse: paired extended-horizon differences (percentage points, mean ± partition SD).

Comparison	Accuracy	Balanced acc.	AUROC
RandomDelete – ERM	1.75 ± 1.14	1.44 ± 2.63	-0.13 ± 1.05
AdvDelete – ERM	2.02 ± 5.93	1.59 ± 5.89	2.17 ± 3.05
MaDRO-1 – ERM	2.50 ± 5.01	2.25 ± 4.98	2.08 ± 3.00
MaDRO-1 – AdvDelete	0.48 ± 2.87	0.66 ± 3.57	-0.08 ± 0.61
Dropout – ERM	-1.35 ± 2.22	-2.43 ± 2.00	-0.78 ± 1.75
SAdD – ERM	-1.53 ± 4.02	-2.57 ± 2.39	-2.95 ± 4.85

L.2 Structured missingness and deletion costs

These structured tests retain their original shared stress-mask realizations (seed 202711) and are separate from the 100-replicate nominal-mask estimates.

The same nominally selected Gaussian checkpoint is used under Complete, MCAR, MAR and MNAR conditions. The structured masks here use the explicit sigmoid/Bernoulli definitions in Appendix K; they are a separate protocol from the four-seed oracle diagnostic study.

Table 27: Gaussian structured-mask metrics at the short horizon (%). No condition-specific retuning.

Condition	Method	Accuracy	Balanced acc.	AUROC
Complete	ERM	94.38 ± 0.30	94.39 ± 0.30	98.72 ± 0.29
Complete	AdvDelete	91.52 ± 1.69	91.53 ± 1.69	97.98 ± 0.40
Complete	MaDRO-1	95.41 ± 0.74	95.41 ± 0.73	99.32 ± 0.24
Complete	MaDRO-CV	96.04 ± 0.54	96.04 ± 0.54	99.28 ± 0.19
MCAR	ERM	78.22 ± 1.72	78.22 ± 1.72	87.66 ± 1.28
MCAR	AdvDelete	77.82 ± 2.98	77.81 ± 3.00	86.49 ± 2.46
MCAR	MaDRO-1	81.55 ± 0.85	81.55 ± 0.85	90.48 ± 0.91
MCAR	MaDRO-CV	82.02 ± 1.83	82.02 ± 1.83	90.10 ± 1.16
MAR	ERM	77.85 ± 0.43	77.85 ± 0.42	87.43 ± 0.88
MAR	AdvDelete	76.38 ± 1.21	76.37 ± 1.22	85.81 ± 1.60
MAR	MaDRO-1	80.34 ± 1.54	80.34 ± 1.53	89.68 ± 0.79
MAR	MaDRO-CV	80.81 ± 2.29	80.82 ± 2.29	89.81 ± 1.61
MNAR	ERM	77.59 ± 2.40	77.59 ± 2.39	88.96 ± 1.61
MNAR	AdvDelete	74.57 ± 4.12	74.55 ± 4.14	85.73 ± 1.84
MNAR	MaDRO-1	81.26 ± 2.41	81.26 ± 2.38	90.75 ± 1.09
MNAR	MaDRO-CV	81.26 ± 2.15	81.26 ± 2.14	90.42 ± 1.12

Table 28: Gaussian structured-mask metrics at the equal time horizon (%). No condition-specific retuning.

Condition	Method	Accuracy	Balanced acc.	AUROC
Complete	ERM	96.46 ± 0.95	96.46 ± 0.95	99.25 ± 0.43
Complete	AdvDelete	91.60 ± 1.65	91.61 ± 1.65	98.05 ± 0.44
Complete	MaDRO-1	95.59 ± 0.63	95.59 ± 0.63	99.30 ± 0.31
Complete	MaDRO-CV	96.17 ± 0.51	96.17 ± 0.51	99.31 ± 0.26
MCAR	ERM	82.49 ± 1.28	82.49 ± 1.28	90.92 ± 0.89
MCAR	AdvDelete	77.98 ± 2.90	77.97 ± 2.92	86.58 ± 2.37
MCAR	MaDRO-1	81.89 ± 1.41	81.89 ± 1.41	90.54 ± 1.22
MCAR	MaDRO-CV	81.89 ± 1.32	81.89 ± 1.32	90.50 ± 1.42
MAR	ERM	81.99 ± 0.74	81.99 ± 0.74	90.65 ± 0.90
MAR	AdvDelete	76.67 ± 1.37	76.66 ± 1.38	85.90 ± 1.57
MAR	MaDRO-1	80.94 ± 1.20	80.94 ± 1.20	90.05 ± 0.83
MAR	MaDRO-CV	80.55 ± 2.21	80.55 ± 2.21	89.88 ± 1.64
MNAR	ERM	81.99 ± 1.37	81.99 ± 1.35	91.90 ± 1.24
MNAR	AdvDelete	75.30 ± 3.55	75.29 ± 3.56	85.92 ± 1.93
MNAR	MaDRO-1	80.92 ± 2.29	80.92 ± 2.29	90.74 ± 1.14
MNAR	MaDRO-CV	81.05 ± 1.72	81.05 ± 1.70	90.51 ± 1.02

Table 29: Gaussian structured-mask metrics at the extended horizon (%). No condition-specific retuning.

Condition	Method	Accuracy	Balanced acc.	AUROC
Complete	ERM	96.27 ± 1.00	96.28 ± 1.00	99.31 ± 0.41
Complete	AdvDelete	94.99 ± 1.03	94.99 ± 1.03	99.11 ± 0.37
Complete	MaDRO-1	96.67 ± 0.94	96.67 ± 0.94	99.41 ± 0.24
Complete	MaDRO-CV	96.82 ± 0.61	96.82 ± 0.61	99.58 ± 0.24
MCAR	ERM	82.81 ± 1.50	82.80 ± 1.51	91.20 ± 1.09
MCAR	AdvDelete	81.68 ± 2.02	81.68 ± 2.02	90.67 ± 1.53
MCAR	MaDRO-1	83.36 ± 2.08	83.36 ± 2.08	91.72 ± 1.20
MCAR	MaDRO-CV	83.20 ± 1.50	83.20 ± 1.50	91.80 ± 0.47
MAR	ERM	81.73 ± 0.90	81.73 ± 0.90	90.94 ± 1.16
MAR	AdvDelete	80.08 ± 1.11	80.08 ± 1.12	89.99 ± 1.07
MAR	MaDRO-1	83.10 ± 1.23	83.10 ± 1.23	92.09 ± 0.58
MAR	MaDRO-CV	82.57 ± 0.68	82.57 ± 0.67	91.57 ± 0.59
MNAR	ERM	81.99 ± 1.21	81.99 ± 1.19	92.06 ± 1.46
MNAR	AdvDelete	82.05 ± 1.08	82.05 ± 1.08	90.61 ± 1.11
MNAR	MaDRO-1	83.33 ± 1.96	83.33 ± 1.94	92.48 ± 1.79
MNAR	MaDRO-CV	83.18 ± 1.17	83.17 ± 1.18	92.78 ± 1.19

Table 30: Training-only CV cost diagnostics. SD is across feature weights, not across runs; condition numbers refer to the ridge-regularized covariance. Clip counts give lower/upper clipping before the final mean normalization.

Dataset	Split	Min	Max	SD	Condition	Clipped
WDBC	2027	0.520	2.080	0.604	53800.0	15/5
WDBC	2028	0.524	2.097	0.602	56768.1	15/5
WDBC	2029	0.532	2.129	0.584	61481.8	14/4
WDBC	2030	0.510	2.041	0.574	63917.0	14/4
WDBC	2031	0.518	2.072	0.589	61812.9	13/4
Gaussian	2027	0.979	1.027	0.011	2.6	0/0
Gaussian	2028	0.968	1.022	0.012	2.8	0/0
Gaussian	2029	0.977	1.030	0.012	2.6	0/0
Gaussian	2030	0.980	1.030	0.011	2.6	0/0
Gaussian	2031	0.969	1.028	0.011	2.5	0/0

Gaussian costs are close to uniform (feature-wise SD approximately 0.011–0.012), whereas WDBC costs are substantially more dispersed. The final normalization follows clipping, so some WDBC costs exceed 2. These diagnostics place the Gaussian experiments near the unit-cost regime and

identify WDBC as the setting with greater cost heterogeneity. The shuffled-cost comparisons below assess feature-specific cost assignment on partition 2027.

L.3 Common optimizer and shuffled-cost controls

These separate two-seed controls use the original shared test-mask realization rather than the replicated primary estimator.

These controls examine sensitivity to optimizer choice and the assignment of feature costs on partition 2027 with optimization seeds 1 and 2. Means and descriptive seed SD summarize optimization variability on that partition, separately from the five-partition primary study. The optimizer comparison selects jointly from all three optimizer settings and the full configuration grid of each method family using validation only. Shuffled-CV uses the default optimizer and a fixed permutation of CV weights.

Table 31: WDBC: validation-selected common-optimizer comparison. Percent mean and sample SD over two optimization seeds on partition 2027, using the original shared test-mask realization.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	93.02 ± 0.51	90.98 ± 0.69	94.29 ± 0.17
Short	AdvDelete	93.70 ± 0.23	91.99 ± 0.11	94.51 ± 0.12
Short	MaDRO-1	94.18 ± 0.45	92.62 ± 0.58	94.50 ± 0.14
Short	MaDRO-CV	93.82 ± 0.39	92.20 ± 0.47	94.71 ± 0.08
Short	Mixture	93.46 ± 1.13	91.60 ± 1.50	94.58 ± 0.17
Equal time	ERM	92.78 ± 0.51	91.13 ± 0.53	94.00 ± 0.13
Equal time	AdvDelete	93.70 ± 0.23	91.99 ± 0.11	94.51 ± 0.12
Equal time	MaDRO-1	94.18 ± 0.45	92.62 ± 0.58	94.50 ± 0.14
Equal time	MaDRO-CV	93.82 ± 0.39	92.20 ± 0.47	94.71 ± 0.08
Equal time	Mixture	93.46 ± 1.13	91.60 ± 1.50	94.58 ± 0.17
Extended	ERM	92.78 ± 0.51	91.13 ± 0.53	94.00 ± 0.13
Extended	AdvDelete	93.70 ± 0.23	91.99 ± 0.11	94.51 ± 0.12
Extended	MaDRO-1	94.34 ± 0.23	92.81 ± 0.31	94.64 ± 0.06
Extended	MaDRO-CV	94.14 ± 0.06	92.52 ± 0.02	94.74 ± 0.03
Extended	Mixture	93.46 ± 1.13	91.60 ± 1.50	94.58 ± 0.17

Table 32: WDBC: shuffled-cost control, two optimization seeds on one partition.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	MaDRO-1	94.18 ± 0.45	92.62 ± 0.58	94.50 ± 0.14
Short	MaDRO-CV	93.86 ± 0.68	92.19 ± 0.95	94.29 ± 0.33
Short	Shuffled-CV	93.82 ± 0.39	92.29 ± 0.60	94.64 ± 0.12
Equal time	MaDRO-1	94.18 ± 0.45	92.62 ± 0.58	94.50 ± 0.14
Equal time	MaDRO-CV	94.14 ± 0.28	92.61 ± 0.35	94.50 ± 0.03
Equal time	Shuffled-CV	94.02 ± 0.11	92.56 ± 0.22	94.67 ± 0.08
Extended	MaDRO-1	94.18 ± 0.45	92.62 ± 0.58	94.50 ± 0.14
Extended	MaDRO-CV	94.14 ± 0.28	92.61 ± 0.35	94.50 ± 0.03
Extended	Shuffled-CV	94.02 ± 0.11	92.56 ± 0.22	94.67 ± 0.08

Table 33: Gaussian: validation-selected common-optimizer comparison. Percent mean and sample SD over two optimization seeds on partition 2027, using the original shared test-mask realization.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	78.55 ± 0.47	78.55 ± 0.45	84.51 ± 1.13
Short	AdvDelete	76.29 ± 0.59	76.33 ± 0.58	83.43 ± 0.15
Short	MaDRO-1	79.48 ± 0.61	79.50 ± 0.60	85.12 ± 0.39
Short	MaDRO-CV	79.62 ± 0.78	79.64 ± 0.78	85.23 ± 0.46
Short	Mixture	79.44 ± 0.46	79.47 ± 0.45	85.58 ± 0.44
Equal time	ERM	79.73 ± 0.12	79.75 ± 0.12	85.22 ± 0.02
Equal time	AdvDelete	76.29 ± 0.59	76.33 ± 0.58	83.43 ± 0.15
Equal time	MaDRO-1	79.19 ± 0.30	79.20 ± 0.32	85.41 ± 0.11
Equal time	MaDRO-CV	79.27 ± 0.27	79.27 ± 0.25	85.23 ± 0.46
Equal time	Mixture	79.73 ± 0.12	79.75 ± 0.12	85.22 ± 0.02
Extended	ERM	79.73 ± 0.12	79.75 ± 0.12	85.22 ± 0.02
Extended	AdvDelete	78.59 ± 0.83	78.60 ± 0.81	84.65 ± 0.37
Extended	MaDRO-1	79.86 ± 0.44	79.87 ± 0.42	85.86 ± 0.12
Extended	MaDRO-CV	80.33 ± 0.42	80.34 ± 0.40	86.14 ± 0.20
Extended	Mixture	79.42 ± 1.06	79.43 ± 1.08	85.38 ± 0.32

Table 34: Gaussian: shuffled-cost control, two optimization seeds on one partition.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	MaDRO-1	77.61 ± 0.66	77.62 ± 0.64	84.71 ± 0.18
Short	MaDRO-CV	78.55 ± 0.20	78.56 ± 0.22	84.72 ± 0.25
Short	Shuffled-CV	78.14 ± 0.71	78.16 ± 0.73	84.47 ± 0.64
Equal time	MaDRO-1	78.41 ± 0.07	78.42 ± 0.07	84.74 ± 0.17
Equal time	MaDRO-CV	78.66 ± 0.05	78.67 ± 0.07	84.86 ± 0.05
Equal time	Shuffled-CV	78.14 ± 0.71	78.15 ± 0.72	84.05 ± 0.05
Extended	MaDRO-1	79.34 ± 1.18	79.33 ± 1.17	85.10 ± 1.20
Extended	MaDRO-CV	80.33 ± 0.42	80.34 ± 0.40	86.14 ± 0.20
Extended	Shuffled-CV	79.71 ± 0.46	79.72 ± 0.44	85.43 ± 0.35

Thresholding with joint optimizer selection. With all three optimizer settings available to every family, extended MaDRO-1/CV exceed the fixed mixture by 0.88/0.68 percentage points on WDBC and 0.44/0.91 on Gaussian data. Gaussian CV gains are 0.45 and 1.36 points for the two seeds (paired SD 0.64); the unit-cost difference varies in sign. At the short horizon, Gaussian accuracy differences narrow to 0.04/0.18 points, while the mixture has higher AUROC by 0.46/0.35 points. At equal training time, the Gaussian mixture selects ERM for both seeds and leads MaDRO-1/CV in accuracy by 0.54/0.47 points. Tables 35–37 give the paired metrics and validation choices. These single-partition controls retain the extended-horizon accuracy benefit under joint tuning and describe optimization-seed variability separately from the repeated-partition primary comparison.

Table 35: WDBC: fixed-mixture minus comparator under common-optimizer selection. Percentage-point paired means and sample SDs across the same two optimization seeds on partition 2027. Positive values favor the mixture; these SDs describe optimization variability on one partition.

Horizon	Comparator	Accuracy	Balanced acc.	AUROC
Short	ERM	0.44 ± 1.64	0.62 ± 2.19	0.30 ± 0.34
Short	AdvDelete	-0.24 ± 1.35	-0.39 ± 1.61	0.07 ± 0.29
Short	MaDRO-1	-0.72 ± 0.68	-1.02 ± 0.92	0.09 ± 0.02
Short	MaDRO-CV	-0.36 ± 0.73	-0.60 ± 1.03	-0.13 ± 0.09
Equal time	ERM	0.68 ± 0.62	0.47 ± 0.97	0.58 ± 0.04
Equal time	AdvDelete	-0.24 ± 1.35	-0.39 ± 1.61	0.07 ± 0.29
Equal time	MaDRO-1	-0.72 ± 0.68	-1.02 ± 0.92	0.09 ± 0.02
Equal time	MaDRO-CV	-0.36 ± 0.73	-0.60 ± 1.03	-0.13 ± 0.09
Extended	ERM	0.68 ± 0.62	0.47 ± 0.97	0.58 ± 0.04
Extended	AdvDelete	-0.24 ± 1.35	-0.39 ± 1.61	0.07 ± 0.29
Extended	MaDRO-1	-0.88 ± 0.90	-1.21 ± 1.19	-0.06 ± 0.22
Extended	MaDRO-CV	-0.68 ± 1.18	-0.92 ± 1.48	-0.16 ± 0.13

Table 36: Gaussian: fixed-mixture minus comparator under common-optimizer selection. Percentage-point paired means and sample SDs across the same two optimization seeds on partition 2027. Positive values favor the mixture; these SDs describe optimization variability on one partition.

Horizon	Comparator	Accuracy	Balanced acc.	AUROC
Short	ERM	0.89 ± 0.02	0.92 ± 0.00	1.07 ± 0.69
Short	AdvDelete	3.15 ± 0.13	3.14 ± 0.13	2.16 ± 0.59
Short	MaDRO-1	-0.04 ± 0.15	-0.03 ± 0.16	0.46 ± 0.05
Short	MaDRO-CV	-0.18 ± 0.32	-0.17 ± 0.33	0.35 ± 0.02
Equal time	ERM	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Equal time	AdvDelete	3.44 ± 0.47	3.42 ± 0.46	1.80 ± 0.16
Equal time	MaDRO-1	0.54 ± 0.42	0.55 ± 0.44	-0.18 ± 0.13
Equal time	MaDRO-CV	0.47 ± 0.15	0.48 ± 0.13	-0.01 ± 0.44
Extended	ERM	-0.31 ± 1.18	-0.32 ± 1.20	0.16 ± 0.34
Extended	AdvDelete	0.84 ± 1.89	0.83 ± 1.89	0.74 ± 0.69
Extended	MaDRO-1	-0.44 ± 1.50	-0.43 ± 1.50	-0.48 ± 0.44
Extended	MaDRO-CV	-0.91 ± 0.64	-0.90 ± 0.67	-0.76 ± 0.11

Table 37: Validation-selected mixture configurations. Each entry is (o, λ), where optimizer indices $o = 0, 1, 2$ follow the ordered settings in Appendix K. Both endpoint controls are eligible.

Dataset	Horizon	Seed 1	Seed 2
WDBC	Short	$(1, 0.25)$	$(0, 0.75)$
WDBC	Equal time	$(1, 0.25)$	$(0, 0.75)$
WDBC	Extended	$(1, 0.25)$	$(0, 0.75)$
Gaussian	Short	$(1, 0.5)$	$(1, 0.25)$
Gaussian	Equal time	$(0, 0)$	$(1, 0)$
Gaussian	Extended	$(1, 0.5)$	$(0, 0.25)$

L.4 Common-optimizer search cost

Table 38: Common-optimizer control: mean full-family training search cost (seconds) over two seeds on partition 2027, including every configuration at all three optimizer settings. Mixture endpoints reuse the corresponding ERM/AdvDelete fits but count toward the mixture search. Horizons share trajectories and are not additive.

Dataset	Method	Grid	Short	Time	Extended
WDBC	ERM	3	126.9	900.0	773.2
WDBC	AdvDelete	3	358.8	900.0	1866.6
WDBC	MaDRO-1	15	1810.1	4500.3	10459.5
WDBC	MaDRO-CV	15	1796.8	4500.3	9757.5
WDBC	Mixture	15	1884.2	4500.3	10173.2
Gaussian	ERM	3	184.9	900.0	1556.0
Gaussian	AdvDelete	3	830.2	900.1	5890.3
Gaussian	MaDRO-1	15	4170.1	4500.6	31980.0
Gaussian	MaDRO-CV	15	4153.8	4500.7	31194.4
Gaussian	Mixture	15	3937.7	4500.5	24694.2

L.5 Selection, stopping, and computational cost

Costs sum synchronized training computation across every candidate in the indicated family and horizon, then average over the ten split/seed combinations. Validation and storage time are excluded. Budgets are matched per configuration; total searches are larger for MaDRO and SAdD. The three horizons share each trajectory, so their costs must not be added. Maximum-update stops are reported separately from validation plateaus.

Table 39: WDBC: mean full-family training search time (seconds), selected extended checkpoint range, and capped candidate counts.

Method	Grid	Short	Time	Extended	Best step	Capped
ERM	1	41.4	300.0	210.8	225–8500	0/10
AdvDelete	1	118.6	300.0	593.1	225–2900	0/10
MaDRO-1	5	595.4	1500.1	3109.7	150–13000	0/50
MaDRO-CV	5	592.5	1500.1	3037.4	200–11000	0/50

Table 40: Gaussian: mean full-family training search time (seconds), selected extended checkpoint range, and capped candidate counts.

Method	Grid	Short	Time	Extended	Best step	Capped
ERM	1	58.0	300.0	514.2	11000–23500	0/10
AdvDelete	1	273.1	300.0	2318.3	9500–30000	0/10
MaDRO-1	5	1369.2	1500.2	10459.8	5000–28000	0/50
MaDRO-CV	5	1368.2	1500.2	10115.7	10500–29000	0/50

Table 41: Credit: mean full-family training search time (seconds), selected extended checkpoint range, and capped candidate counts.

Method	Grid	Short	Time	Extended	Best step	Capped
ERM	1	17.1	120.0	25.7	25–75	0/10
RandomDelete	1	17.8	120.0	26.8	25–375	0/10
AdvDelete	1	35.6	120.0	53.4	50–475	0/10
MaDRO-1	5	179.7	600.0	268.6	50–150	0/50
Dropout	1	18.0	120.0	27.1	25–300	0/10
SAdD	4	76.0	480.0	114.9	25–1950	0/40

Table 42: Horse: mean full-family training search time (seconds), selected extended checkpoint range, and capped candidate counts.

Method	Grid	Short	Time	Extended	Best step	Capped
ERM	1	20.7	120.0	31.0	25–50	0/10
RandomDelete	1	21.5	120.0	32.2	25–50	0/10
AdvDelete	1	48.3	120.0	73.8	25–1500	0/10
MaDRO-1	5	244.8	600.1	366.1	25–75	0/50
Dropout	1	21.8	120.0	32.6	25–25	0/10
SAdD	4	90.6	480.0	137.7	25–2100	0/40

Table 43: WDBC: validation-selected configuration counts across ten split/seed combinations. MaDRO entries are gamma values; counts sum to ten in each row.

Horizon	Family	Configuration: count
Short	MaDRO-1	0.05: 1; 0.1: 2; 0.5: 3; 0.2: 3; 1: 1
Short	MaDRO-CV	0.05: 2; 0.1: 2; 0.2: 2; 0.5: 1; 1: 3
Equal time	MaDRO-1	0.1: 2; 0.5: 3; 0.2: 3; 1: 2
Equal time	MaDRO-CV	0.05: 3; 0.1: 2; 0.2: 2; 1: 3
Extended	MaDRO-1	0.1: 2; 0.5: 4; 0.2: 2; 1: 2
Extended	MaDRO-CV	0.05: 2; 0.1: 2; 0.2: 2; 1: 4

Table 44: Gaussian: validation-selected configuration counts across ten split/seed combinations. MaDRO entries are gamma values; counts sum to ten in each row.

Horizon	Family	Configuration: count
Short	MaDRO-1	0.1: 1; 0.5: 1; 0.2: 7; 1: 1
Short	MaDRO-CV	0.1: 3; 0.2: 5; 0.5: 1; 1: 1
Equal time	MaDRO-1	0.1: 3; 0.5: 1; 0.2: 6
Equal time	MaDRO-CV	0.1: 4; 0.2: 3; 0.5: 2; 1: 1
Extended	MaDRO-1	0.05: 1; 0.5: 3; 1: 6
Extended	MaDRO-CV	0.5: 3; 1: 7

Table 45: Credit: validation-selected configuration counts across ten split/seed combinations. MaDRO entries are gamma values; counts sum to ten in each row.

Horizon	Family	Configuration: count
Short	MaDRO-1	0.5: 3; 1: 7
Short	SAdD	k=16, lambda=0.1: 3; k=16, lambda=1: 4; k=4, lambda=0.1: 1; k=4, lambda=1: 2
Equal time	MaDRO-1	0.5: 3; 1: 7
Equal time	SAdD	k=16, lambda=0.1: 3; k=16, lambda=1: 4; k=4, lambda=0.1: 1; k=4, lambda=1: 2
Extended	MaDRO-1	0.5: 3; 1: 7
Extended	SAdD	k=16, lambda=0.1: 3; k=16, lambda=1: 4; k=4, lambda=0.1: 1; k=4, lambda=1: 2

Table 46: Horse: validation-selected configuration counts across ten split/seed combinations. MaDRO entries are gamma values; counts sum to ten in each row.

Horizon	Family	Configuration: count
Short	MaDRO-1	0.05: 1; 0.1: 6; 0.2: 3
Short	SAdD	k=16, lambda=0.1: 3; k=16, lambda=1: 4; k=4, lambda=1: 3
Equal time	MaDRO-1	0.05: 1; 0.1: 6; 0.2: 3
Equal time	SAdD	k=16, lambda=0.1: 2; k=16, lambda=1: 4; k=4, lambda=1: 4
Extended	MaDRO-1	0.05: 1; 0.1: 6; 0.2: 3
Extended	SAdD	k=16, lambda=0.1: 2; k=16, lambda=1: 4; k=4, lambda=1: 4

L.6 Four-seed comparison on the original partition

The numerical four-seed diagnostics use the original shared test-mask realization on partition 2027; natural-data diagnostics retain intrinsic missingness.

The following separate summaries use four optimization seeds on partition 2027. Their SD measures initialization variability conditional on that partition; these four seeds are not added to the five-partition estimates.

Table 47: WDBC: four-seed, single-partition metrics under the shared-checkpoint protocol.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	93.16 ± 0.16	91.33 ± 0.19	94.30 ± 0.15
Short	AdvDelete	93.74 ± 0.19	91.98 ± 0.21	94.50 ± 0.13
Short	MaDRO-1	93.74 ± 0.58	92.03 ± 0.77	94.39 ± 0.17
Short	MaDRO-CV	93.72 ± 0.48	92.02 ± 0.71	94.41 ± 0.23
Equal time	ERM	92.92 ± 0.34	91.15 ± 0.31	94.18 ± 0.22
Equal time	AdvDelete	93.74 ± 0.19	91.98 ± 0.21	94.50 ± 0.13
Equal time	MaDRO-1	93.80 ± 0.56	92.07 ± 0.75	94.41 ± 0.18
Equal time	MaDRO-CV	93.72 ± 0.51	92.00 ± 0.74	94.47 ± 0.05
Extended	ERM	92.92 ± 0.34	91.15 ± 0.31	94.18 ± 0.22
Extended	AdvDelete	93.74 ± 0.19	91.98 ± 0.21	94.50 ± 0.13
Extended	MaDRO-1	93.84 ± 0.50	92.27 ± 0.55	94.54 ± 0.10
Extended	MaDRO-CV	93.78 ± 0.46	92.09 ± 0.66	94.54 ± 0.09

Table 48: Gaussian: four-seed, single-partition metrics under the shared-checkpoint protocol.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	76.79 ± 0.45	76.80 ± 0.45	83.27 ± 0.28
Short	AdvDelete	74.39 ± 1.99	74.42 ± 1.97	82.48 ± 1.15
Short	MaDRO-1	78.16 ± 0.76	78.17 ± 0.76	84.68 ± 0.16
Short	MaDRO-CV	78.63 ± 0.38	78.63 ± 0.38	84.52 ± 0.28
Equal time	ERM	79.07 ± 0.44	79.08 ± 0.45	85.18 ± 0.25
Equal time	AdvDelete	74.57 ± 1.73	74.60 ± 1.71	82.37 ± 1.33
Equal time	MaDRO-1	78.46 ± 0.35	78.47 ± 0.36	84.69 ± 0.16
Equal time	MaDRO-CV	78.69 ± 0.35	78.69 ± 0.35	84.59 ± 0.32
Extended	ERM	79.37 ± 0.37	79.38 ± 0.38	85.18 ± 0.12
Extended	AdvDelete	77.59 ± 0.75	77.61 ± 0.76	84.41 ± 1.13
Extended	MaDRO-1	79.83 ± 0.91	79.83 ± 0.91	85.58 ± 0.89
Extended	MaDRO-CV	80.39 ± 0.55	80.40 ± 0.55	86.02 ± 0.44

Table 49: Credit: four-seed, single-partition metrics under the shared-checkpoint protocol.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	89.13 ± 0.59	89.15 ± 0.40	94.92 ± 1.11
Short	RandomDelete	86.78 ± 1.37	86.62 ± 1.39	93.55 ± 0.58
Short	AdvDelete	87.86 ± 1.61	87.08 ± 1.65	92.39 ± 1.58
Short	MaDRO-1	87.50 ± 0.69	86.97 ± 1.03	94.09 ± 0.89
Short	Dropout	89.86 ± 0.00	89.84 ± 0.29	95.18 ± 0.94
Short	SAdD	87.86 ± 2.24	87.89 ± 2.15	93.52 ± 2.47
Equal time	ERM	89.13 ± 0.59	89.15 ± 0.40	94.92 ± 1.11
Equal time	RandomDelete	86.78 ± 1.37	86.62 ± 1.39	93.55 ± 0.58
Equal time	AdvDelete	87.86 ± 1.61	87.08 ± 1.65	92.39 ± 1.58
Equal time	MaDRO-1	87.50 ± 0.69	86.97 ± 1.03	94.09 ± 0.89
Equal time	Dropout	88.95 ± 1.81	89.08 ± 1.65	95.50 ± 0.38
Equal time	SAdD	87.86 ± 2.24	87.89 ± 2.15	93.52 ± 2.47
Extended	ERM	89.13 ± 0.59	89.15 ± 0.40	94.92 ± 1.11
Extended	RandomDelete	86.78 ± 1.37	86.62 ± 1.39	93.55 ± 0.58
Extended	AdvDelete	87.86 ± 1.61	87.08 ± 1.65	92.39 ± 1.58
Extended	MaDRO-1	87.50 ± 0.69	86.97 ± 1.03	94.09 ± 0.89
Extended	Dropout	89.86 ± 0.00	89.84 ± 0.29	95.18 ± 0.94
Extended	SAdD	87.86 ± 2.24	87.89 ± 2.15	93.52 ± 2.47

Table 50: Horse: four-seed, single-partition metrics under the shared-checkpoint protocol.

Horizon	Method	Accuracy	Balanced acc.	AUROC
Short	ERM	67.41 ± 13.73	67.20 ± 13.70	74.89 ± 18.96
Short	RandomDelete	68.35 ± 14.36	68.14 ± 14.34	75.14 ± 19.14
Short	AdvDelete	70.25 ± 5.71	69.56 ± 6.08	84.74 ± 1.31
Short	MaDRO-1	74.05 ± 5.61	73.56 ± 6.05	83.95 ± 0.98
Short	Dropout	69.62 ± 10.49	69.53 ± 10.23	75.63 ± 19.54
Short	SAdD	77.53 ± 3.64	77.25 ± 3.55	86.01 ± 1.60
Equal time	ERM	67.41 ± 13.73	67.20 ± 13.70	74.89 ± 18.96
Equal time	RandomDelete	68.35 ± 14.36	68.14 ± 14.34	75.14 ± 19.14
Equal time	AdvDelete	70.25 ± 5.71	69.56 ± 6.08	84.74 ± 1.31
Equal time	MaDRO-1	74.05 ± 5.61	73.56 ± 6.05	83.95 ± 0.98
Equal time	Dropout	69.62 ± 10.49	69.53 ± 10.23	75.63 ± 19.54
Equal time	SAdD	77.53 ± 3.64	77.25 ± 3.55	86.01 ± 1.60
Extended	ERM	67.41 ± 13.73	67.20 ± 13.70	74.89 ± 18.96
Extended	RandomDelete	68.35 ± 14.36	68.14 ± 14.34	75.14 ± 19.14
Extended	AdvDelete	70.25 ± 5.71	69.56 ± 6.08	84.74 ± 1.31
Extended	MaDRO-1	74.05 ± 5.61	73.56 ± 6.05	83.95 ± 0.98
Extended	Dropout	69.62 ± 10.49	69.53 ± 10.23	75.63 ± 19.54
Extended	SAdD	77.53 ± 3.64	77.25 ± 3.55	86.01 ± 1.60

L.7 ACCURACY ACROSS MISSINGNESS RATES

These sensitivity curves use one common test-mask realization per example and rate; they complement the 100-replicate numerical primary estimates.

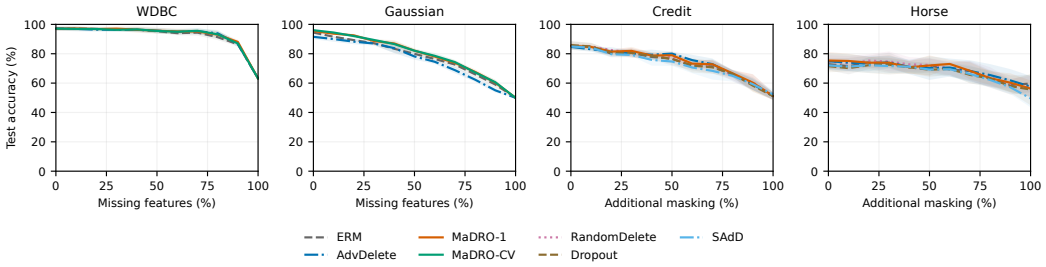

Figure 4: Short-horizon test accuracy across all eleven masking rates. The configuration and checkpoint are held fixed across rates. Lines and bands show means and one descriptive SD over five partition means, each averaging two optimization seeds. Credit/Horse retain intrinsic missingness before the plotted additional deletions; zero additional deletion is their primary condition.

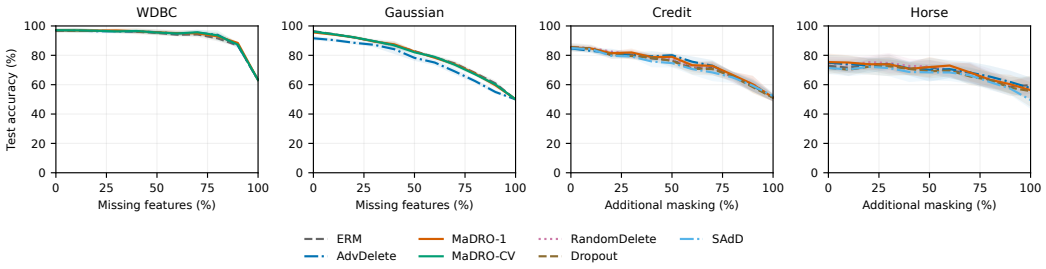

Figure 5: Equal time-horizon test accuracy across all eleven masking rates. The configuration and checkpoint are held fixed across rates. Lines and bands show means and one descriptive SD over five partition means, each averaging two optimization seeds. Credit/Horse retain intrinsic missingness before the plotted additional deletions; zero additional deletion is their primary condition.

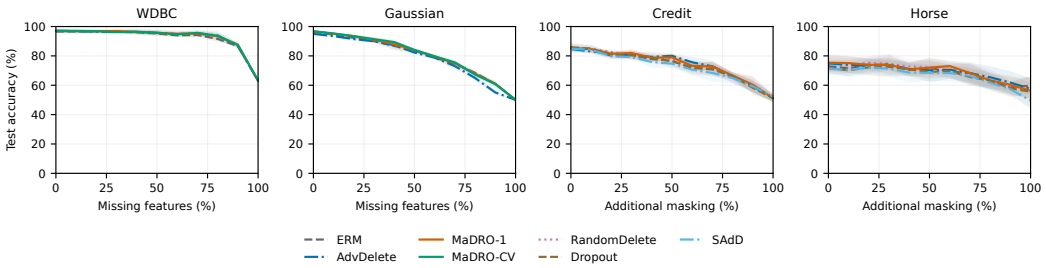

Figure 6: Extended-horizon test accuracy across all eleven masking rates. The configuration and checkpoint are held fixed across rates. Lines and bands show means and one descriptive SD over five partition means, each averaging two optimization seeds. Credit/Horse retain intrinsic missingness before the plotted additional deletions; zero additional deletion is their primary condition.

L.8 FIXED-MIXTURE CONTROL FOR ADVERSARIAL-LOSS STRENGTH

We compare cost thresholding against a fixed convex mixture of the nominal and zero-cost no-op/delete-one losses,

$$L_\lambda = (1-\lambda)\ell_0 + \lambda \max\{\ell_0, \max_{j \in V(m)} \ell_j\} = \ell_0 + \lambda(\Delta)_+.$$

The no-op branch is eligible, and the all-missing case reduces to ℓ_0. For interior weights, gradients flow through both the nominal and selected adversarial loss, with detached branch selection. Unlike MaDRO's threshold, a fixed λ scales every positive raw deletion penalty. This control compares the predictive performance of these two ways of moderating adversarial training.

Matched design and selection. WDBC and Gaussian use the same outer splits 2027–2031, inner split seed 0, two optimization seeds, training-only preprocessing, predictor, initialization, data order and mask random generators as the numerical study. Adam uses learning rate .001, zero weight decay, batch size 64, and T4 hardware. We select among $\lambda \in \{0, .25, .5, .75, 1\}$ separately for each horizon, matching each five-setting MaDRO family in search cardinality. The exact ERM and AdvDelete endpoints reuse the corresponding trajectories and predictions; the three interior weights are trained from the original initialization. Short uses 3,000 updates, time uses 300 synchronized training seconds per candidate, and extended uses the same 15,000–60,000-update bounds and validation-loss patience rule. Checkpoint eligibility and selection are unchanged: minimize validation cross-entropy within each horizon, then maximize validation accuracy across configurations, breaking ties by loss and configuration name. The full five-setting choice is fixed before evaluating its selected test predictor. Endpoints remain eligible and are retained when selected.

All 100 shared test-mask replicates and both seed outcomes are averaged within each partition before computing the five-partition mean and descriptive SD defined at the start of this section. The comparison covers nominal-mask performance on WDBC and Gaussian data. Every horizon and comparator is reported below. Full-family cost includes the reused ERM and AdvDelete endpoints as well as the three newly trained interior weights. Each candidate receives the stated time budget, so full search cost accounts for all five candidates.

Table 51: Fixed-mixture performance (percent units), mean ± descriptive SD across five partition means.

Dataset	Horizon	Accuracy	Balanced acc.	AUROC
WDBC	Short	92.16 ± 1.04	90.51 ± 1.44	94.41 ± 0.29
WDBC	Time	92.07 ± 0.95	90.42 ± 1.36	94.32 ± 0.25
WDBC	Extended	92.12 ± 0.88	90.59 ± 1.33	94.36 ± 0.21
Gaussian	Short	78.36 ± 0.39	78.36 ± 0.38	84.75 ± 0.30
Gaussian	Time	79.29 ± 0.56	79.29 ± 0.56	85.34 ± 0.48
Gaussian	Extended	79.84 ± 0.34	79.84 ± 0.34	85.83 ± 0.28

Table 52: WDBC: fixed mixture minus comparator (percentage points), paired partition mean ± descriptive SD. Positive favors the mixture.

Horizon	Comparator	Accuracy	Balanced acc.	AUROC
Short	ERM	0.21 ± 0.17	0.19 ± 0.15	0.15 ± 0.08
Short	AdvDelete	-0.08 ± 0.29	-0.08 ± 0.20	-0.10 ± 0.20
Short	MaDRO-1	-0.19 ± 0.17	-0.07 ± 0.28	-0.10 ± 0.16
Short	MaDRO-CV	-0.13 ± 0.26	-0.17 ± 0.33	-0.08 ± 0.17
Time	ERM	0.29 ± 0.31	0.20 ± 0.16	0.15 ± 0.13
Time	AdvDelete	-0.17 ± 0.33	-0.17 ± 0.26	-0.18 ± 0.15
Time	MaDRO-1	-0.23 ± 0.39	-0.05 ± 0.49	-0.13 ± 0.20
Time	MaDRO-CV	-0.24 ± 0.36	-0.32 ± 0.38	-0.19 ± 0.10
Extended	ERM	0.34 ± 0.31	0.37 ± 0.43	0.18 ± 0.12
Extended	AdvDelete	-0.12 ± 0.45	-0.00 ± 0.56	-0.15 ± 0.20
Extended	MaDRO-1	-0.12 ± 0.46	0.07 ± 0.53	-0.10 ± 0.22
Extended	MaDRO-CV	-0.16 ± 0.38	-0.10 ± 0.56	-0.14 ± 0.10

Table 53: Gaussian: fixed mixture minus comparator (percentage points), paired partition mean ± descriptive SD. Positive favors the mixture.

Horizon	Comparator	Accuracy	Balanced acc.	AUROC
Short	ERM	1.27 ± 0.40	1.27 ± 0.38	0.87 ± 0.21
Short	AdvDelete	3.16 ± 1.22	3.16 ± 1.23	1.74 ± 0.61
Short	MaDRO-1	-0.65 ± 0.16	-0.66 ± 0.15	-0.45 ± 0.20
Short	MaDRO-CV	-0.72 ± 0.30	-0.73 ± 0.28	-0.38 ± 0.15
Time	ERM	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Time	AdvDelete	3.91 ± 1.16	3.91 ± 1.18	2.26 ± 0.62
Time	MaDRO-1	0.09 ± 0.28	0.10 ± 0.28	0.01 ± 0.19
Time	MaDRO-CV	0.18 ± 0.48	0.17 ± 0.48	0.16 ± 0.40
Extended	ERM	0.36 ± 0.42	0.36 ± 0.42	0.28 ± 0.29
Extended	AdvDelete	1.59 ± 0.48	1.59 ± 0.48	0.97 ± 0.34
Extended	MaDRO-1	-0.32 ± 0.44	-0.32 ± 0.43	-0.24 ± 0.25
Extended	MaDRO-CV	-0.44 ± 0.36	-0.44 ± 0.35	-0.33 ± 0.23

Table 54: Selected mixture weights across ten split–seed runs per horizon. Zero and one are ERM and AdvDelete.

Dataset	Horizon	0	.25	.5	.75	1
WDBC	Short	4	1	0	3	2
WDBC	Time	4	0	1	3	2
WDBC	Extended	4	1	1	3	1
Gaussian	Short	0	5	3	2	0
Gaussian	Time	10	0	0	0	0
Gaussian	Extended	3	6	0	1	0

Table 55: Mean synchronized training seconds per split–seed search. Full cost counts all five settings; the endpoint and interior columns partition that cost. Validation and I/O are excluded.

Dataset	Horizon	Full family	Endpoints	Interiors
WDBC	Short	614.4	160.0	454.4
WDBC	Time	1500.1	600.0	900.1
WDBC	Extended	3124.6	803.9	2320.7
Gaussian	Short	1286.0	331.1	954.9
Gaussian	Time	1500.2	600.1	900.1
Gaussian	Extended	9389.3	2832.6	6556.7

Interpretation. Cost thresholding improves mean accuracy over fixed loss scaling at the extended horizon: Gaussian MaDRO-1/CV reach 80.16%/80.29%, compared with 79.84% for the fixed mixture. The paired MaDRO-minus-mixture differences are 0.32 ± 0.44/0.44 ± 0.36 points (mean $\pm$ descriptive partition SD); the unit-cost difference ranges from -0.23 to 0.96, and the CV difference from 0.18 to 1.05. Their conditional mask Monte Carlo SEs are 0.0147/0.0153 points. Extended WDBC differences are 0.12 ± 0.46/0.16 ± 0.38 points and change sign across partitions. Under equal per-candidate time on Gaussian data, the mixture selects ERM in all runs and leads MaDRO-1/CV by 0.09/0.18 points. Its extended full-family search costs 9,389.3 seconds, versus 10,459.8/10,115.7 for MaDRO-1/CV. The comparison identifies a budget-dependent predictive benefit of thresholding, with its magnitude varying across partitions and cost choices. Appendix L.10 separates numerical integration precision from partition variability for every metric and pairwise comparison.

L.9 Single-realization mask sensitivity

The same selected predictors are evaluated with one shared mask per example and rate (PyTorch seed 202709). Comparison with Table 1 isolates numerical integration sensitivity at fixed models and partitions.

Table 56: Single-realization nominal-mask accuracy (%), mean $\pm$ descriptive partition SD. The selected predictors and training horizons match Table 1.

	WDBC			Gaussian		
Method	Short	Time	Extended	Short	Time	Extended
ERM	91.71 ± 0.98	91.59 ± 0.81	91.59 ± 0.81	77.27 ± 0.68	79.50 ± 0.73	79.66 ± 0.79
AdvDelete	92.11 ± 1.07	92.11 ± 1.07	92.11 ± 1.07	75.40 ± 1.29	75.60 ± 1.24	78.29 ± 0.59
MaDRO-1	92.26 ± 1.28	92.22 ± 1.30	92.19 ± 1.33	79.02 ± 0.99	79.18 ± 0.77	80.13 ± 0.79
MaDRO-CV	92.12 ± 1.11	92.18 ± 1.22	92.16 ± 1.22	79.21 ± 0.40	79.18 ± 0.33	80.33 ± 0.31
Fixed mixture	92.07 ± 1.10	91.98 ± 0.93	91.93 ± 0.94	78.37 ± 0.75	79.50 ± 0.73	79.80 ± 0.62

L.10 PRECISION UNDER REPEATED EVALUATION MASKS

All numerical primary comparisons use 100 independent common mask replicates at the same frozen, validation-selected predictors. Each replicate assigns one exact-size mask to every test example at each of the eleven rates. Masks are shared across methods, optimization seeds and horizons within a partition. The resulting mean integrates evaluation-mask variation; the underlying five partitions and two optimization seeds remain fixed.

For each replicate, we average the two optimization seeds within each partition and then the five partitions. The conditional mask Monte Carlo standard error (MCSE) is the sample SD of these 100 replicate means divided by $\sqrt{100}$. Paired MCSE uses the shared-mask method difference before averaging. Separately, partition SD is computed across five partition means after averaging all masks and both optimization seeds. MCSE measures integration precision conditional on the selected models and data; partition SD describes sensitivity across overlapping partitions.

Table 57: WDBC: mask-averaged metrics $\pm$ conditional mask MCSE. Accuracy, balanced accuracy and AUROC are in percent; BCE is in nats.

Horizon	Method	Accuracy	Balanced acc.	AUROC	BCE
Short	ERM	91.955 ± 0.015	90.322 ± 0.017	94.257 ± 0.008	0.1806 ± 0.0003
Short	AdvDelete	92.242 ± 0.014	90.591 ± 0.016	94.507 ± 0.007	0.1752 ± 0.0003
Short	MaDRO-1	92.357 ± 0.015	90.584 ± 0.018	94.508 ± 0.007	0.1707 ± 0.0002
Short	MaDRO-CV	92.289 ± 0.015	90.686 ± 0.017	94.488 ± 0.007	0.1740 ± 0.0003
Short	Fixed mixture	92.163 ± 0.015	90.512 ± 0.017	94.406 ± 0.007	0.1759 ± 0.0003
Time	ERM	91.785 ± 0.016	90.221 ± 0.018	94.176 ± 0.008	0.1884 ± 0.0004
Time	AdvDelete	92.242 ± 0.014	90.591 ± 0.016	94.507 ± 0.007	0.1752 ± 0.0003
Time	MaDRO-1	92.304 ± 0.015	90.462 ± 0.017	94.459 ± 0.008	0.1728 ± 0.0003
Time	MaDRO-CV	92.315 ± 0.014	90.731 ± 0.016	94.511 ± 0.007	0.1725 ± 0.0003
Time	Fixed mixture	92.071 ± 0.015	90.416 ± 0.017	94.325 ± 0.007	0.1834 ± 0.0003
Extended	ERM	91.785 ± 0.016	90.221 ± 0.018	94.176 ± 0.008	0.1884 ± 0.0004
Extended	AdvDelete	92.242 ± 0.014	90.591 ± 0.016	94.507 ± 0.007	0.1752 ± 0.0003
Extended	MaDRO-1	92.243 ± 0.015	90.518 ± 0.017	94.462 ± 0.008	0.1793 ± 0.0003
Extended	MaDRO-CV	92.283 ± 0.015	90.688 ± 0.017	94.504 ± 0.007	0.1753 ± 0.0003
Extended	Fixed mixture	92.124 ± 0.016	90.587 ± 0.018	94.360 ± 0.007	0.1857 ± 0.0004

Table 58: Gaussian: mask-averaged metrics $\pm$ conditional mask MCSE. Accuracy, balanced accuracy and AUROC are in percent; BCE is in nats.

Horizon	Method	Accuracy	Balanced acc.	AUROC	BCE
Short	ERM	77.087 ± 0.019	77.089 ± 0.019	83.883 ± 0.019	0.4211 ± 0.0003
Short	AdvDelete	75.201 ± 0.018	75.196 ± 0.018	83.011 ± 0.019	0.4787 ± 0.0002
Short	MaDRO-1	79.009 ± 0.019	79.013 ± 0.019	85.204 ± 0.019	0.3942 ± 0.0002
Short	MaDRO-CV	79.075 ± 0.019	79.084 ± 0.019	85.137 ± 0.019	0.3946 ± 0.0002
Short	Fixed mixture	78.358 ± 0.018	78.357 ± 0.018	84.752 ± 0.018	0.4050 ± 0.0003
Time	ERM	79.292 ± 0.020	79.292 ± 0.020	85.342 ± 0.018	0.3929 ± 0.0003
Time	AdvDelete	75.383 ± 0.018	75.379 ± 0.018	83.082 ± 0.020	0.4785 ± 0.0002
Time	MaDRO-1	79.200 ± 0.019	79.196 ± 0.019	85.331 ± 0.019	0.3903 ± 0.0002
Time	MaDRO-CV	79.115 ± 0.019	79.122 ± 0.019	85.180 ± 0.019	0.3924 ± 0.0002
Time	Fixed mixture	79.292 ± 0.020	79.292 ± 0.020	85.342 ± 0.018	0.3929 ± 0.0003
Extended	ERM	79.479 ± 0.021	79.478 ± 0.021	85.550 ± 0.018	0.3908 ± 0.0003
Extended	AdvDelete	78.251 ± 0.016	78.253 ± 0.016	84.857 ± 0.017	0.4298 ± 0.0003
Extended	MaDRO-1	80.162 ± 0.018	80.159 ± 0.018	86.068 ± 0.018	0.3765 ± 0.0003
Extended	MaDRO-CV	80.287 ± 0.020	80.286 ± 0.020	86.152 ± 0.018	0.3689 ± 0.0003
Extended	Fixed mixture	79.842 ± 0.018	79.842 ± 0.018	85.827 ± 0.017	0.3890 ± 0.0004

Table 59: WDBC, short: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	0.287 ± 0.014	0.269 ± 0.015	0.250 ± 0.006	-0.0054 ± 0.0002
MaDRO-1 – ERM	0.402 ± 0.013	0.261 ± 0.015	0.250 ± 0.005	-0.0099 ± 0.0002
MaDRO-1 – AdvDelete	0.115 ± 0.013	-0.007 ± 0.014	0.000 ± 0.005	-0.0045 ± 0.0002
MaDRO-CV – ERM	0.334 ± 0.013	0.364 ± 0.015	0.231 ± 0.005	-0.0066 ± 0.0002
MaDRO-CV – AdvDelete	0.047 ± 0.013	0.095 ± 0.014	-0.019 ± 0.004	-0.0012 ± 0.0002
MaDRO-CV – MaDRO-1	-0.068 ± 0.014	0.102 ± 0.015	-0.019 ± 0.004	0.0033 ± 0.0002
Fixed mixture – ERM	0.209 ± 0.011	0.190 ± 0.011	0.149 ± 0.004	-0.0046 ± 0.0002
Fixed mixture – AdvDelete	-0.079 ± 0.012	-0.079 ± 0.014	-0.101 ± 0.005	0.0007 ± 0.0002
Fixed mixture – MaDRO-1	-0.194 ± 0.013	-0.071 ± 0.014	-0.101 ± 0.004	0.0053 ± 0.0002
Fixed mixture – MaDRO-CV	-0.126 ± 0.013	-0.174 ± 0.015	-0.082 ± 0.004	0.0020 ± 0.0002

Table 60: WDBC, time: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	0.457 ± 0.014	0.370 ± 0.015	0.331 ± 0.006	-0.0132 ± 0.0003
MaDRO-1 – ERM	0.519 ± 0.015	0.241 ± 0.016	0.283 ± 0.005	-0.0156 ± 0.0002
MaDRO-1 – AdvDelete	0.061 ± 0.012	-0.129 ± 0.013	-0.049 ± 0.005	-0.0024 ± 0.0002
MaDRO-CV – ERM	0.530 ± 0.012	0.511 ± 0.014	0.335 ± 0.005	-0.0159 ± 0.0002
MaDRO-CV – AdvDelete	0.073 ± 0.012	0.141 ± 0.014	0.003 ± 0.005	-0.0027 ± 0.0002
MaDRO-CV – MaDRO-1	0.011 ± 0.012	0.270 ± 0.014	0.052 ± 0.004	-0.0003 ± 0.0002
Fixed mixture – ERM	0.286 ± 0.012	0.195 ± 0.012	0.149 ± 0.005	-0.0050 ± 0.0002
Fixed mixture – AdvDelete	-0.171 ± 0.011	-0.175 ± 0.012	-0.183 ± 0.005	0.0082 ± 0.0002
Fixed mixture – MaDRO-1	-0.232 ± 0.012	-0.046 ± 0.013	-0.134 ± 0.005	0.0106 ± 0.0002
Fixed mixture – MaDRO-CV	-0.243 ± 0.012	-0.315 ± 0.013	-0.186 ± 0.005	0.0109 ± 0.0002

Table 61: WDBC, extended: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	0.457 ± 0.014	0.370 ± 0.015	0.331 ± 0.006	-0.0132 ± 0.0003
MaDRO-1 – ERM	0.458 ± 0.014	0.298 ± 0.015	0.286 ± 0.005	-0.0091 ± 0.0003
MaDRO-1 – AdvDelete	0.001 ± 0.012	-0.072 ± 0.014	-0.046 ± 0.005	0.0041 ± 0.0002
MaDRO-CV – ERM	0.499 ± 0.013	0.467 ± 0.014	0.328 ± 0.005	-0.0131 ± 0.0002
MaDRO-CV – AdvDelete	0.041 ± 0.012	0.097 ± 0.014	-0.003 ± 0.005	0.0001 ± 0.0002
MaDRO-CV – MaDRO-1	0.041 ± 0.013	0.169 ± 0.014	0.043 ± 0.004	-0.0040 ± 0.0002
Fixed mixture – ERM	0.339 ± 0.011	0.367 ± 0.012	0.184 ± 0.005	-0.0027 ± 0.0002
Fixed mixture – AdvDelete	-0.118 ± 0.012	-0.003 ± 0.014	-0.147 ± 0.005	0.0105 ± 0.0003
Fixed mixture – MaDRO-1	-0.119 ± 0.013	0.069 ± 0.014	-0.102 ± 0.005	0.0064 ± 0.0003
Fixed mixture – MaDRO-CV	-0.159 ± 0.012	-0.100 ± 0.013	-0.144 ± 0.004	0.0104 ± 0.0003

Table 62: Gaussian, short: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	-1.886 ± 0.018	-1.892 ± 0.018	-0.873 ± 0.013	0.0575 ± 0.0002
MaDRO-1 – ERM	1.922 ± 0.015	1.924 ± 0.015	1.320 ± 0.012	-0.0270 ± 0.0002
MaDRO-1 – AdvDelete	3.808 ± 0.016	3.817 ± 0.016	2.193 ± 0.012	-0.0845 ± 0.0002
MaDRO-CV – ERM	1.988 ± 0.018	1.996 ± 0.018	1.254 ± 0.013	-0.0266 ± 0.0002
MaDRO-CV – AdvDelete	3.874 ± 0.018	3.888 ± 0.018	2.126 ± 0.014	-0.0841 ± 0.0002
MaDRO-CV – MaDRO-1	0.065 ± 0.016	0.071 ± 0.016	-0.067 ± 0.011	0.0004 ± 0.0001
Fixed mixture – ERM	1.271 ± 0.015	1.269 ± 0.015	0.869 ± 0.010	-0.0161 ± 0.0002
Fixed mixture – AdvDelete	3.157 ± 0.017	3.161 ± 0.017	1.742 ± 0.012	-0.0736 ± 0.0002
Fixed mixture – MaDRO-1	-0.651 ± 0.015	-0.655 ± 0.015	-0.451 ± 0.010	0.0109 ± 0.0002
Fixed mixture – MaDRO-CV	-0.716 ± 0.019	-0.727 ± 0.019	-0.385 ± 0.012	0.0105 ± 0.0002

Table 63: Gaussian, time: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	-3.909 ± 0.018	-3.913 ± 0.018	-2.259 ± 0.013	0.0856 ± 0.0002
MaDRO-1 – ERM	-0.092 ± 0.015	-0.097 ± 0.015	-0.010 ± 0.010	-0.0026 ± 0.0002
MaDRO-1 – AdvDelete	3.817 ± 0.016	3.816 ± 0.016	2.249 ± 0.013	-0.0882 ± 0.0002
MaDRO-CV – ERM	-0.176 ± 0.016	-0.170 ± 0.016	-0.162 ± 0.011	-0.0005 ± 0.0002
MaDRO-CV – AdvDelete	3.733 ± 0.016	3.743 ± 0.016	2.098 ± 0.013	-0.0861 ± 0.0002
MaDRO-CV – MaDRO-1	-0.084 ± 0.015	-0.073 ± 0.015	-0.152 ± 0.011	0.0021 ± 0.0001
Fixed mixture – ERM	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.0000 ± 0.0000
Fixed mixture – AdvDelete	3.909 ± 0.018	3.913 ± 0.018	2.259 ± 0.013	-0.0856 ± 0.0002
Fixed mixture – MaDRO-1	0.092 ± 0.015	0.097 ± 0.015	0.010 ± 0.010	0.0026 ± 0.0002
Fixed mixture – MaDRO-CV	0.176 ± 0.016	0.170 ± 0.016	0.162 ± 0.011	0.0005 ± 0.0002

Table 64: Gaussian, extended: all paired differences ± conditional mask MCSE. First method minus second; positive favors the first for accuracy, balanced accuracy and AUROC, while negative favors it for BCE. Classification metrics use percentage points; BCE uses nats.

Comparison	Accuracy	Balanced acc.	AUROC	BCE
AdvDelete – ERM	-1.228 ± 0.018	-1.225 ± 0.018	-0.693 ± 0.014	0.0390 ± 0.0002
MaDRO-1 – ERM	0.683 ± 0.015	0.681 ± 0.015	0.518 ± 0.011	-0.0143 ± 0.0002
MaDRO-1 – AdvDelete	1.912 ± 0.017	1.905 ± 0.017	1.211 ± 0.012	-0.0533 ± 0.0002
MaDRO-CV – ERM	0.807 ± 0.017	0.808 ± 0.017	0.602 ± 0.011	-0.0219 ± 0.0002
MaDRO-CV – AdvDelete	2.036 ± 0.017	2.032 ± 0.017	1.295 ± 0.014	-0.0608 ± 0.0003
MaDRO-CV – MaDRO-1	0.124 ± 0.017	0.127 ± 0.017	0.084 ± 0.011	-0.0076 ± 0.0002
Fixed mixture – ERM	0.363 ± 0.013	0.363 ± 0.013	0.276 ± 0.009	-0.0018 ± 0.0002
Fixed mixture – AdvDelete	1.591 ± 0.016	1.588 ± 0.016	0.969 ± 0.014	-0.0408 ± 0.0003
Fixed mixture – MaDRO-1	-0.320 ± 0.015	-0.317 ± 0.015	-0.242 ± 0.010	0.0125 ± 0.0002
Fixed mixture – MaDRO-CV	-0.444 ± 0.015	-0.444 ± 0.015	-0.326 ± 0.011	0.0200 ± 0.0002

L.11 Common-cost deletion-sensitivity profiles

We evaluate all five method families at their fixed, validation-selected extended-horizon checkpoints on WDBC and the Gaussian benchmark. The five partitions and two optimization seeds give 100 selected predictors, represented by 92 distinct checkpoint contexts because some fixed mixtures select an endpoint. Every predictor uses the same 100 test-mask replicates and eleven equally weighted missing rates as Appendix L.10. The additional evaluation enumerates every deletion of a visible feature; it does not change training, checkpoint selection or the mask banks.

For a fixed predictor and masked example, let L_0 be its nominal binary cross-entropy and let L_j be the loss after deleting visible feature j. We measure

$$T = \max\{0, \max_{j \text{ visible}} (L_j - L_0)\}, \qquad H(\gamma) = \mathbb{E}(T - \gamma)_+, \qquad A(\gamma) = \Pr(T > \gamma).$$

Here $T = 0$ for an all-missing example, and the expectation averages test rows, missing rates, mask replicates, seeds and partitions with the same weights as the primary numerical comparison. Activation uses a strict inequality, including at ties; float32 BCE losses are subtracted and reduced in float64. All methods are evaluated at the common unit-cost grid $\gamma \in \{0, 0.05, 0.1, 0.2, 0.5, 1\}$, including MaDRO-CV, whose feature-dependent costs are used in training. This common evaluation scale separates the learned predictor's deletion sensitivity from its training cost convention. Nominal BCE is reported alongside the sensitivity profiles.

The two profiles have complementary interpretations: $A(\gamma)$ measures the fraction of examples whose largest deletion-induced loss increase exceeds the cost, while $H(\gamma)$ also measures the size of that excess. In particular, $H(\gamma) = \int_\gamma^\infty A(t)\,dt$ and its right derivative is $-A(\gamma)$. Thus the profiles characterize the loss-increase distribution at fixed predictors, and the common-cost objective is $\mathbb{E}L_0 + H(\gamma)$. Tables 65–66 report every method and threshold. The accompanying evaluation tables provide all signed paired differences, rate-level profiles, descriptive SD across the five overlapping partitions and conditional mask Monte Carlo SE separately.

Predictive risk and sensitivity. In the aggregate extended-horizon comparison, both MaDRO variants have lower nominal BCE and mean T than ERM and the validation-selected fixed mixture on both datasets (Table 65). Gaussian AdvDelete has the smallest mean T (0.3095), compared with 0.3364/0.3484 for MaDRO-1/CV, while their nominal BCE is 0.3765/0.3689 versus AdvDelete's 0.4298. The hinge profiles cross: AdvDelete has lower $H(\gamma)$ than both MaDRO variants at $\gamma \in \{0, 0.05, 0.1\}$, and higher values at $\gamma \in \{0.2, 0.5, 1\}$. The latter contrast holds in all five partition means at costs 0.5 and 1; at cost 0.2 it holds in three partitions for MaDRO-1 and two for CV. Thus a smaller average deletion-induced increase can coexist with greater excess above a larger cost. On WDBC, AdvDelete has the lowest mean nominal BCE (0.1752) and mean T (0.1003), with CV close at 0.1753 and 0.1040. These fixed-predictor profiles distinguish nominal risk, average deletion sensitivity and tail severity; they complement the predictive comparisons and characterize the measured tradeoff.

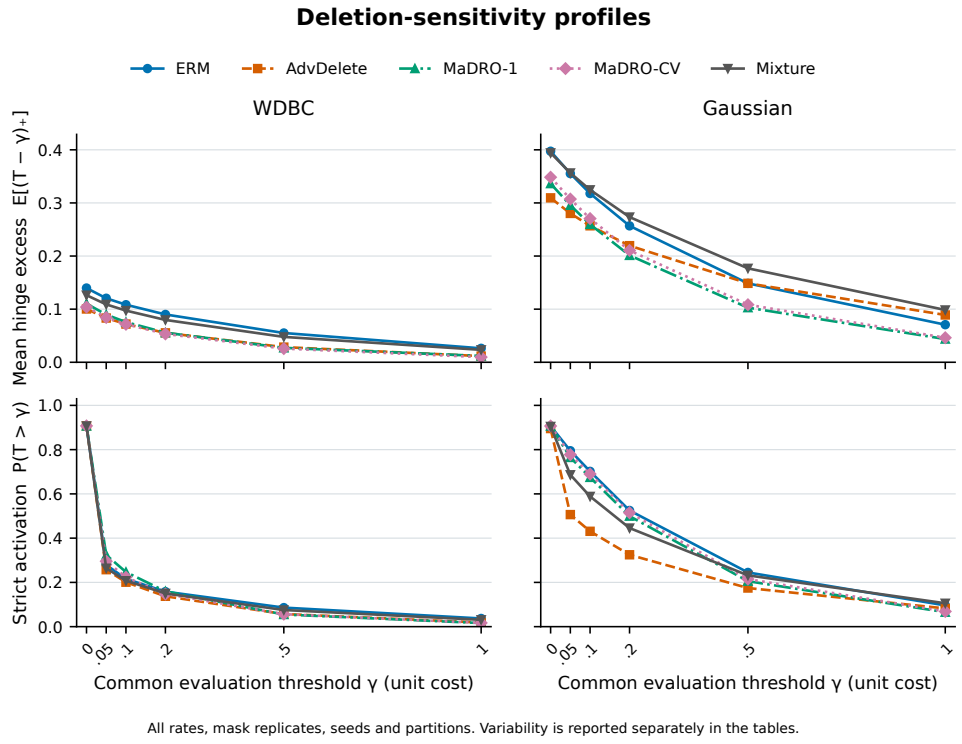

Figure 7: Common-cost deletion-sensitivity profiles for all five families at validation-selected extended checkpoints. Lines join all six thresholds, averaging eleven rates, 100 shared mask replicates, two seeds and five partitions. Tables 65–66 accompany the profiles.

Dataset	Family	Nominal BCE Mean	Nominal BCE SD	Nominal BCE MCSE	Mean T Mean	Mean T SD	Mean T MCSE
WDBC	ERM	0.1884	0.01562	0.0003544	0.1398	0.01498	0.0002891
WDBC	AdvDelete	0.1752	0.02145	0.0002692	0.1003	0.01619	0.0002079
WDBC	MaDRO-1	0.1793	0.02751	0.0002993	0.1104	0.02282	0.0001726
WDBC	MaDRO-CV	0.1753	0.0177	0.0002769	0.104	0.01006	0.0001857
WDBC	Mixture	0.1857	0.02202	0.0003732	0.1263	0.01693	0.0002849
Gaussian	ERM	0.3908	0.01797	0.0003232	0.3973	0.01453	0.0002043
Gaussian	AdvDelete	0.4298	0.02095	0.0002967	0.3095	0.03995	0.0002305
Gaussian	MaDRO-1	0.3765	0.01326	0.000302	0.3364	0.01243	0.0001868
Gaussian	MaDRO-CV	0.3689	0.008522	0.0002852	0.3484	0.01683	0.0002004
Gaussian	Mixture	0.389	0.01238	0.0003547	0.3942	0.01633	0.0002331

Table 65: Nominal predictive risk and deletion sensitivity for all families. SD denotes descriptive partition SD; MCSE denotes conditional mask Monte Carlo standard error. They quantify different variation and are not confidence intervals.

Dataset	Family	Profile	$\gamma = 0$	.05	.1	.2	.5	1
WDBC	ERM	Hinge	0.1398	0.1205	0.1084	0.09002	0.05511	0.02624
WDBC	ERM	Activation	0.9069	0.2791	0.216	0.1582	0.086	0.03696
WDBC	AdvDelete	Hinge	0.1003	0.08327	0.07196	0.05536	0.02852	0.01165
WDBC	AdvDelete	Activation	0.9071	0.2574	0.2009	0.1373	0.05703	0.01873
WDBC	MaDRO-1	Hinge	0.1104	0.08973	0.07563	0.05573	0.02717	0.01193
WDBC	MaDRO-1	Activation	0.9068	0.3239	0.2468	0.1595	0.05489	0.01623
WDBC	MaDRO-CV	Hinge	0.104	0.08433	0.07156	0.05339	0.02574	0.009701
WDBC	MaDRO-CV	Activation	0.9073	0.2956	0.2234	0.1475	0.05646	0.01694
WDBC	Mixture	Hinge	0.1263	0.1089	0.0973	0.07972	0.04778	0.02337
WDBC	Mixture	Activation	0.9069	0.2641	0.2075	0.1503	0.07522	0.03087
Gaussian	ERM	Hinge	0.3973	0.3553	0.3178	0.2569	0.1485	0.07077
Gaussian	ERM	Activation	0.9066	0.795	0.7017	0.525	0.245	0.09608
Gaussian	AdvDelete	Hinge	0.3095	0.2803	0.2569	0.2195	0.1486	0.08909
Gaussian	AdvDelete	Activation	0.8953	0.5064	0.4309	0.3248	0.1746	0.08235
Gaussian	MaDRO-1	Hinge	0.3364	0.2956	0.2596	0.2012	0.1028	0.04369
Gaussian	MaDRO-1	Activation	0.9063	0.7647	0.6746	0.4994	0.2056	0.06531
Gaussian	MaDRO-CV	Hinge	0.3484	0.3073	0.2705	0.2106	0.1084	0.04623
Gaussian	MaDRO-CV	Activation	0.9067	0.7773	0.6898	0.5148	0.215	0.06874
Gaussian	Mixture	Hinge	0.3942	0.3564	0.3246	0.2734	0.1769	0.09848
Gaussian	Mixture	Activation	0.9048	0.6863	0.5886	0.4456	0.2318	0.1057

Table 66: All common evaluation thresholds under unit deletion cost, independent of selected training parameters. Hinge is $H_\gamma = \mathbb{E}(T - \gamma)_+$ and activation is $A_\gamma = \mathbb{P}(T > \gamma)$. Thus $H_0 = \mathbb{E}T$, while $A_0 = \mathbb{P}(T > 0)$. Complete uncertainty and all signed paired differences accompany the standalone tables.

L.12 TRAINING AND VALIDATION TRAJECTORIES

The following diagnostics use the extended-horizon validation-selected configuration for each partition, seed, and method. Every trajectory remains separate, avoiding changing-composition averages after early stopping. Both panels evaluate nominal-mask binary cross-entropy (BCE), not the robust training objective. The fixed training probe contains at most 256 training examples; validation uses the full validation partition.

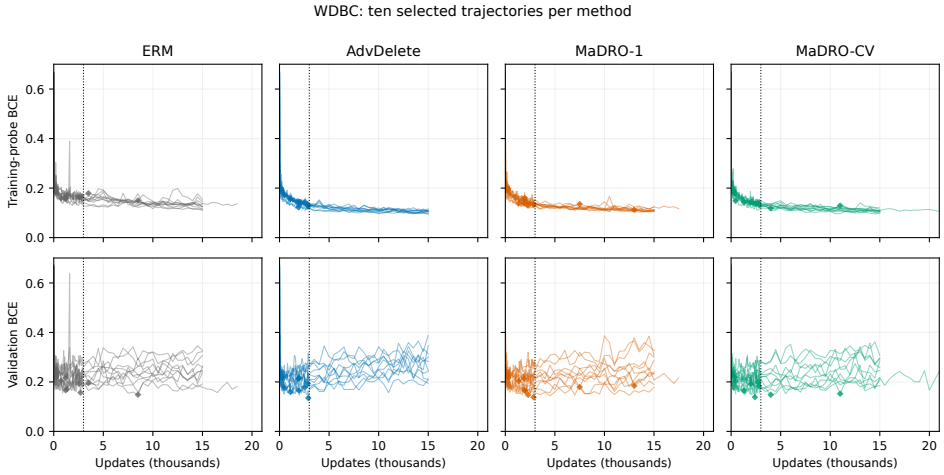

Figure 8: WDBC: training-probe and validation BCE for ERM, AdvDelete, MaDRO-1 and MaDRO-CV. Each column contains all ten trajectories (five partitions, two seeds), ending at the extended stopping update. Diamonds mark minimum-validation-loss checkpoints; dotted lines mark 3,000 updates. Curves use the common checkpoint grid and fixed nominal evaluation masks. Axes are shared across methods. Together, the panels show training fit, validation fit, and checkpoint selection along each trajectory.

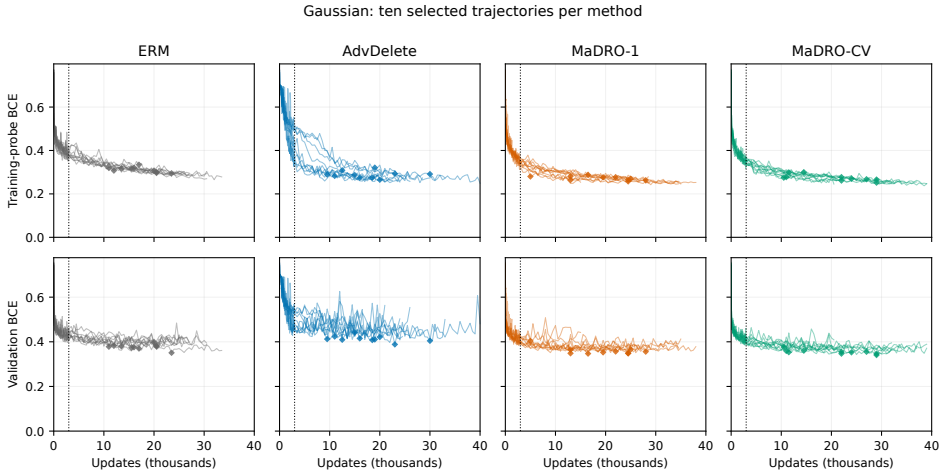

Figure 9: Gaussian: training-probe and validation BCE for ERM, AdvDelete, MaDRO-1 and MaDRO-CV. Each column contains all ten trajectories (five partitions, two seeds), ending at the extended stopping update. Diamonds mark minimum-validation-loss checkpoints; dotted lines mark 3,000 updates. Curves use the common checkpoint grid and fixed nominal evaluation masks. Axes are shared across methods. Together, the panels show training fit, validation fit, and checkpoint selection along each trajectory.

L.12.1 DESCRIPTIVE LOSS-GAP DIAGNOSTICS

To distinguish prediction quality from a smaller training–validation loss gap, define

$$G_h = L_{\mathrm{val}}(\widehat{\theta}_h) - L_{\mathrm{probe}}(\widehat{\theta}_h), \qquad D_{\mathrm{ext}} = L_{\mathrm{val}}(\theta_{T_{\mathrm{ext}}}) - L_{\mathrm{val}}(\widehat{\theta}_{\mathrm{ext}}).$$

Here $\widehat{\theta}_h$ is the configuration and checkpoint selected under the validation rule for horizon h, and T_{ext} is that configuration's frozen extended endpoint. Both losses are nominal-mask BCE in nats. The probe is the first $\min(256, n_{\mathrm{train}})$ rows in the training-index order, with mask seed 702; full validation uses seed 701. Every example receives one mask per rate. We retain signed gaps, including negative values, and zero endpoint deterioration. All five methods, three horizons, five partitions and two optimization seeds are included; Tables 67–68 average seeds within partitions before reporting descriptive partition SD.

The gaps describe the selected models conditional on a training subset and distinct fixed training/validation masks. They are descriptive rather than unbiased population generalization-gap estimates. Endpoint differences use each configuration's stopping step, so they summarize its subsequent trajectory rather than a common-duration causal comparison.

Table 67: Signed nominal validation-minus-training-probe BCE at the selected checkpoints. Mean $\pm$ descriptive partition SD; corresponding loss levels appear in Table 68.

Dataset	Method	Short	Equal time	Extended
WDBC	ERM	0.022 ± 0.026	0.022 ± 0.024	0.022 ± 0.024
	AdvDelete	0.037 ± 0.032	0.037 ± 0.032	0.037 ± 0.032
	MaDRO-1	0.029 ± 0.024	0.035 ± 0.027	0.035 ± 0.027
	MaDRO-CV	0.029 ± 0.028	0.034 ± 0.021	0.037 ± 0.018
	Fixed mixture	0.032 ± 0.032	0.040 ± 0.037	0.040 ± 0.036
Gaussian	ERM	0.033 ± 0.026	0.063 ± 0.013	0.065 ± 0.016
	AdvDelete	0.049 ± 0.019	0.051 ± 0.018	0.125 ± 0.007
	MaDRO-1	0.065 ± 0.008	0.074 ± 0.013	0.087 ± 0.012
	MaDRO-CV	0.066 ± 0.013	0.072 ± 0.017	0.080 ± 0.010
	Fixed mixture	0.064 ± 0.010	0.063 ± 0.013	0.098 ± 0.012

Table 68: Extended-horizon nominal BCE at the validation-selected checkpoint and subsequent validation deterioration at that configuration's extended endpoint. Mean $\pm$ descriptive partition SD.

Dataset	Method	Probe BCE	Validation BCE	D_{ext}
WDBC	ERM	0.167 ± 0.011	0.190 ± 0.029	0.068 ± 0.027
	AdvDelete	0.150 ± 0.023	0.187 ± 0.030	0.086 ± 0.035
	MaDRO-1	0.149 ± 0.028	0.184 ± 0.032	0.067 ± 0.024
	MaDRO-CV	0.146 ± 0.022	0.183 ± 0.032	0.071 ± 0.016
	Fixed mixture	0.145 ± 0.026	0.185 ± 0.032	0.077 ± 0.023
Gaussian	ERM	0.312 ± 0.011	0.377 ± 0.010	0.015 ± 0.006
	AdvDelete	0.289 ± 0.011	0.413 ± 0.009	0.067 ± 0.045
	MaDRO-1	0.272 ± 0.007	0.359 ± 0.015	0.025 ± 0.009
	MaDRO-CV	0.275 ± 0.011	0.355 ± 0.008	0.022 ± 0.007
	Fixed mixture	0.278 ± 0.015	0.376 ± 0.011	0.031 ± 0.023

At the extended horizon, Gaussian MaDRO-1/CV reduce selected validation BCE to 0.359/0.355 from ERM's 0.377. Training-probe BCE also decreases, by a larger amount, yielding mean signed gaps of 0.087/0.080 versus 0.065 for ERM. Both variants have larger mean signed gaps than ERM at each horizon on both datasets. Endpoint validation increases are 0.025/0.022 versus ERM's 0.015 on Gaussian, and 0.067/0.071 versus 0.068 on WDBC. The predictive improvement is therefore expressed in lower selected validation loss and higher test accuracy, alongside a different balance between training fit, validation fit, and subsequent trajectory behavior.

M SINGLE-PARTITION CONTROLS AND EXACT INNER MAXIMIZATION

This section specifies separate four-seed, single-partition comparisons on L4 GPUs, including the exact eight-feature adversary study. Their checkpoint and timing protocols differ from the repeated-partition T4 comparisons in Appendix K; results are not pooled across these protocols. All methods within each comparison share the partitions and preprocessing below and use validation-based selection.

Data and splitting. We use WDBC and the 50-dimensional Gaussian semi-synthetic benchmark, together with UCI Credit Approval and Horse Colic (Quinlan, 1987; McLeish & Cecile, 1989). Credit Approval retains all 15 predictors and its original missing values. For Horse Colic, we pool the two source files, discard examples with unknown outcome, and predict survival versus death/euthanasia. Predictors are age and original fields 4–22; surgery status, hospital identifier, outcome, and retrospective lesion/pathology fields are excluded. Repeated hospital identifiers stay within one split. Other datasets group identical predictor rows before splitting. A fixed outer seed 2027 and inner seed 0 give approximately 60/20/20 training/validation/test partitions. Where all groups are singletons we stratify by label; otherwise we use disjoint group splits. Numerical means and standard deviations and categorical vocabularies are fitted on training rows only. Missing values remain explicitly masked; unseen categories receive a separate unknown code. Four optimization seeds $\{1, 2, 3, 4\}$ share this fixed split. Their standard deviation measures optimization variability conditional on this split, not uncertainty over independent datasets.

Mask distributions. For WDBC and semi-synthetic data, training draws an index uniformly from the eleven rates $p \in \{0, 0.1, \ldots, 1\}$ and then chooses exactly $D - \mathrm{round}(D(1-p))$ missing features uniformly. Evaluation averages equally over the same eleven rates, including the all-missing condition. Distinct random streams for data order, base masks, and method-specific augmentation give all methods the same training examples and base masks. Fixed evaluation random streams are shared across methods. On the naturally missing datasets, the base mask is the observed missingness pattern. Primary validation and test metrics introduce no additional masks; artificial deletions form a separate secondary stress test.

Models and selection. All neural methods use the same Transformer feature encoder, dimension 128, two Transformer layers, two heads, no internal dropout, and a binary linear output head. Training uses Adam with learning rate 10^{-3}, batch size 64, and zero weight decay. Equal-step comparisons use 3,000 updates on WDBC and the semi-synthetic data, 2,000 on the natural datasets, and 1,000 in the eight-feature experiment below. Each run is evaluated at 20 checkpoints, and the checkpoint with minimum validation cross-entropy is selected. For each seed, hyperparameters maximize validation accuracy at the corresponding loss-selected checkpoint, with ties resolved by lower validation loss and then a fixed configuration name. Positive costs use $\gamma \in \{0.05, 0.1, 0.2, 0.5, 1\}$; $\gamma = 0$ is a separate AdvDelete control. Test evaluation uses the selected configuration and checkpoint without further tuning.

Controls and adversarial dropout. ERM uses the base mask without an additional deletion. RandomDelete deletes one uniformly sampled visible feature, doing nothing if none remains. AdvDelete maximizes over no-op and single-feature deletions without a cost. MaDRO-1 uses unit deletion costs. MaDRO-CV is restricted to numerical datasets: a training-only covariance with $10^{-4}I$ regularization gives costs $v_j = 1/(\Sigma^{-1})_{jj}$, which are normalized to mean one, clipped to $[0.5, 2]$, and renormalized.

We adapt supervised adversarial dropout (SAdD) (Park et al., 2018) to the final 128-dimensional hidden vector of the same classifier. A random mask $\epsilon \sim \mathrm{Bernoulli}(0.5)^{128}$ defines the nominal loss. Its adversary may flip at most $k \in \{4, 16\}$ mask bits, and training minimizes the nominal cross-entropy plus $\lambda \in \{0.1, 1\}$ times adversarial cross-entropy (four configurations). For a binary linear head with weight w, the signed logit changes are $(1-2y)h_j w_j(1-2\epsilon_j)$; flipping the largest positive changes exactly solves this hidden-layer Hamming-ball problem. Deterministic inference uses $0.5h$. Ordinary hidden dropout with the same mask probability and inference rule is an additional control.

Compute matching. Separate matched-time runs of ERM, RandomDelete, AdvDelete, validation-selected MaDRO-1, and validation-selected SAdD receive the training time measured for the $\gamma = 0.2$

MaDRO-1 equal-step run on the same NVIDIA L4 GPU. Timing includes synchronized training computation and excludes validation and checkpoint/storage I/O uniformly. Positive-γ and SAdD configurations are selected in the equal-step runs and reused without retuning at the time budget. Twenty validation checks, evenly spaced in elapsed training time, select the checkpoint; these occur at method-dependent update indices and need not include the early checkpoints from equal-step runs. Thus budget and checkpoint-selection effects are coupled. The last update can overshoot the budget by one step. Hyperparameter-search time is reported separately from this per-configuration comparison: five positive-γ configurations for each MaDRO family and four for SAdD. The per-configuration training budgets are matched, while the total hyperparameter-search budgets differ.

Native missing-value tree baseline. On each natural dataset, HistGradientBoostingClassifier uses the same split and training-only preprocessing, retains numerical NaNs, and marks categorical columns explicitly. The grid is learning rate $\{0.03, 0.1\}$ times maximum leaves $\{7, 15, 31\}$, with 300 iterations and early stopping disabled. Validation accuracy, then cross-entropy, selects the setting. This deterministic baseline is fitted once per configuration and reported without a seed-based standard deviation.

Multi-feature inner maximization. Using the first eight WDBC predictors, we enumerate every subset of at most $k \in \{1, 2, 4, 8\}$ further deletions, including no-op, with fixed $\gamma = 0.2$ and unit costs. The penalty scale and base masks are fixed across budgets; the admissible deletion set grows with k. This exact small-dimensional study measures the effect of enlarging the uncertainty set.

Evaluation metrics and deletion sensitivity. We report seed-level paired differences as well as means and standard deviations. Held-out diagnostics measure the maximum nonnegative loss increase from further deletion and the activation rate of the cost-penalized branch. For comparability, all methods use the same unit-cost $\gamma = 0.2$ diagnostic, including models trained with CV costs. The diagnostic uses up to the first 128 test examples and a fixed 50% base mask for complete datasets, or the intrinsic mask for natural datasets. Training and validation curves use fixed evaluation masks.

M.1 Results for single-partition controls

Under the L4 protocol of Appendix M, neural results are reported as mean $\pm$ sample standard deviation over four optimization seeds on one fixed split. These standard deviations quantify optimization variability conditional on that split.

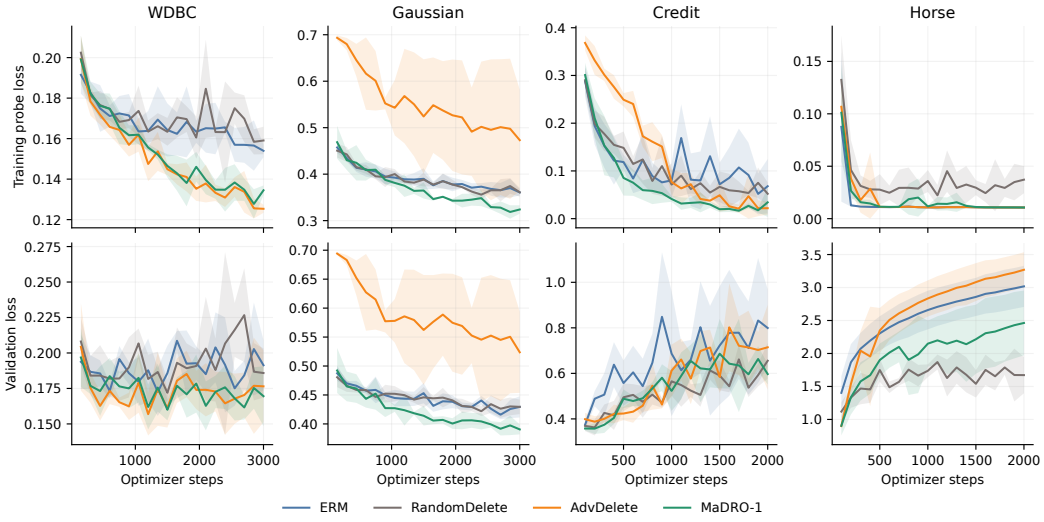

Figure 10: Equal-step training-probe (top) and validation (bottom) cross-entropy. Lines and bands show means and one sample SD over four optimization seeds. Each method uses its validation-selected configuration. The fixed training probe contains at most 256 training examples; validation uses the entire validation split. Complete datasets average the aligned eleven-rate mask mixture; natural datasets retain intrinsic missingness.

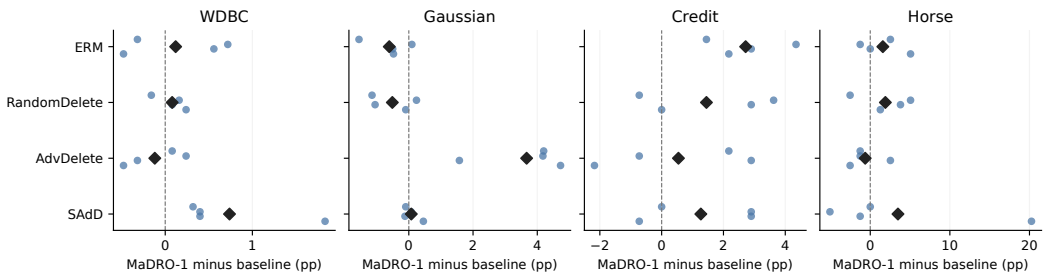

Figure 11: Paired accuracy differences under matched training time. Blue points show seeds 1–4 and black diamonds their mean. Positive differences favor MaDRO-1. Each panel uses its own horizontal scale.

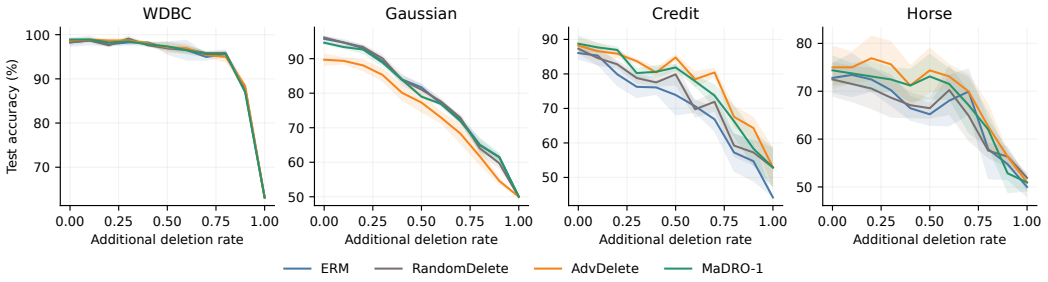

Figure 12: Matched-time models under additional test-time deletion. Lines and bands are means and one sample SD over four seeds. At each rate, the additional mask is drawn over all feature indices and combined with intrinsic missingness. Credit and Horse are trained and selected using intrinsic missingness only; the zero-additional-deletion endpoint is their primary test metric, and the remaining curve is a secondary stress test.

Table 69: Data sizes and intrinsic missingness after the filtering described above. Missingness is the fraction of predictor cells, before artificial masks.

Dataset	Rows	Features	Train/valid/test	Missing
WDBC	569	30	341/114/114	0.00%
Gaussian	1904	50	1142/381/381	0.00%
Credit	690	15	414/138/138	0.65%
Horse	366	20	214/73/79	26.11%
WDBC-8	569	8	341/114/114	0.00%

Table 70: Equal training steps. Primary test accuracy (%). The complete datasets use the aligned rate mixture; natural datasets retain only their intrinsic missingness.

Method	WDBC	Gaussian	Credit	Horse
ERM	93.18 ± 0.58	76.71 ± 0.74	89.67 ± 1.49	74.68 ± 5.94
RandomDelete	93.12 ± 0.56	76.88 ± 0.54	87.14 ± 2.86	73.42 ± 3.73
AdvDelete	93.82 ± 0.33	72.01 ± 6.35	86.41 ± 1.49	75.63 ± 4.05
Hidden dropout	92.78 ± 0.31	76.26 ± 0.36	89.31 ± 1.24	75.32 ± 3.01
SAdD	92.80 ± 0.55	76.76 ± 0.60	87.14 ± 1.49	72.78 ± 8.30
MaDRO-1	93.88 ± 0.25	78.31 ± 0.67	88.41 ± 1.32	74.05 ± 2.19
MaDRO-CV	93.98 ± 0.29	78.60 ± 0.55	–	–

Table 71: Equal training steps: measured training seconds, excluding validation and checkpoint/storage I/O. Neural experiments in this section use NVIDIA L4 GPUs.

Method	WDBC	Gaussian	Credit	Horse
ERM	31.64 ± 0.30	43.27 ± 0.64	13.64 ± 0.13	15.09 ± 0.36
RandomDelete	32.69 ± 0.51	43.71 ± 0.54	14.55 ± 0.19	15.56 ± 0.19
AdvDelete	80.44 ± 0.81	181.33 ± 1.43	25.16 ± 0.26	32.10 ± 0.58
Hidden dropout	32.34 ± 0.27	44.26 ± 0.52	14.23 ± 0.11	15.75 ± 0.26
SAdD	33.75 ± 0.61	45.73 ± 0.66	15.27 ± 0.30	16.53 ± 0.35
MaDRO-1	82.99 ± 5.07	182.43 ± 1.55	25.89 ± 0.82	32.25 ± 0.79
MaDRO-CV	80.22 ± 0.66	182.35 ± 1.53	–	–

Table 72: Matched training time. Primary test accuracy (%). The complete datasets use the aligned rate mixture; natural datasets retain only their intrinsic missingness.

Method	WDBC	Gaussian	Credit	Horse
ERM	93.30 ± 0.37	78.59 ± 0.42	86.05 ± 1.91	72.78 ± 3.19
RandomDelete	93.34 ± 0.25	78.49 ± 0.57	87.32 ± 1.51	72.47 ± 3.64
AdvDelete	93.54 ± 0.31	74.31 ± 1.56	88.22 ± 2.16	75.00 ± 4.55
SAdD	92.68 ± 0.88	77.90 ± 0.52	87.50 ± 0.91	70.89 ± 9.19
MaDRO-1	93.42 ± 0.24	77.98 ± 0.40	88.77 ± 2.40	74.37 ± 3.16

Table 73: Matched training time: measured training seconds, excluding validation and checkpoint/storage I/O. Neural experiments in this section use NVIDIA L4 GPUs.

Method	WDBC	Gaussian	Credit	Horse
ERM	90.01 ± 0.65	182.58 ± 1.62	35.46 ± 0.36	32.82 ± 0.78
RandomDelete	90.02 ± 0.65	182.58 ± 1.62	35.46 ± 0.36	32.82 ± 0.78
AdvDelete	90.02 ± 0.65	182.59 ± 1.62	35.46 ± 0.36	32.82 ± 0.78
SAdD	90.01 ± 0.65	182.58 ± 1.62	35.46 ± 0.36	32.82 ± 0.78
MaDRO-1	90.02 ± 0.65	182.60 ± 1.62	35.46 ± 0.36	32.83 ± 0.78

Table 74: Optimizer updates completed under matched training-time budgets, mean and sample SD over four seeds. Each method selects a checkpoint within this run using validation loss.

Method	WDBC	Gaussian	Credit	Horse
ERM	8492 ± 144	12646 ± 67	5214 ± 57	4352 ± 30
RandomDelete	8336 ± 142	12350 ± 62	5022 ± 97	4223 ± 47
AdvDelete	3367 ± 38	3027 ± 22	2797 ± 41	2047 ± 23
SAdD	7934 ± 201	11911 ± 85	4688 ± 79	3979 ± 24
MaDRO-1	3368 ± 40	3008 ± 15	2783 ± 40	2042 ± 25

Table 75: Natural missingness: balanced accuracy and AUROC (both %). HGB is a separate deterministic CPU baseline, outside the neural time-matching protocol. Its test accuracy is 84.78% on Credit and 79.75% on Horse. HGB entries correspond to a single deterministic fit.

Dataset	Budget	Method	Balanced acc.	AUROC
Credit	steps	ERM	89.68 ± 1.58	93.93 ± 0.91
Credit	steps	RandomDelete	87.20 ± 2.53	93.08 ± 1.16
Credit	steps	AdvDelete	85.40 ± 1.57	93.27 ± 2.03
Credit	steps	SAdD	87.32 ± 1.69	94.38 ± 0.93
Credit	steps	MaDRO-1	87.91 ± 1.64	94.27 ± 1.29
Credit	time	ERM	86.31 ± 1.71	92.35 ± 1.22
Credit	time	RandomDelete	87.27 ± 1.40	93.46 ± 1.07
Credit	time	AdvDelete	87.40 ± 2.36	92.90 ± 2.87
Credit	time	SAdD	87.86 ± 1.16	92.02 ± 2.25
Credit	time	MaDRO-1	88.53 ± 2.83	94.26 ± 2.36
Credit	CPU	HGB	84.83	94.02
Horse	steps	ERM	74.67 ± 5.74	84.35 ± 6.76
Horse	steps	RandomDelete	73.43 ± 3.85	83.36 ± 3.20
Horse	steps	AdvDelete	75.42 ± 4.05	81.93 ± 2.14
Horse	steps	SAdD	72.79 ± 8.19	75.85 ± 6.84
Horse	steps	MaDRO-1	74.01 ± 2.03	83.12 ± 2.24
Horse	time	ERM	72.70 ± 3.29	83.30 ± 4.61
Horse	time	RandomDelete	72.22 ± 3.71	76.75 ± 4.06
Horse	time	AdvDelete	74.90 ± 4.54	82.57 ± 3.36
Horse	time	SAdD	70.63 ± 9.43	76.67 ± 7.72
Horse	time	MaDRO-1	74.34 ± 3.19	82.67 ± 1.38
Horse	CPU	HGB	79.62	87.42

Table 76: Paired MaDRO-1 accuracy differences (percentage points), equal steps. Positive favors MaDRO-1. Summary entries give the mean and sample SD across four optimization seeds; individual differences appear in seed order 1–4.

Dataset	Budget	Comparator	Mean ± SD	Per-seed differences
WDBC	steps	vs. ERM	+0.70 ± 0.62	+1.59, +0.32, +0.24, +0.64
WDBC	steps	vs. AdvDelete	+0.06 ± 0.29	-0.08, -0.16, +0.48, -0.00
WDBC	steps	vs. RandomDelete	+0.76 ± 0.57	+0.24, +0.88, +1.52, +0.40
WDBC	steps	vs. SAdD	+1.08 ± 0.38	+1.28, +0.80, +0.72, +1.52
Gaussian	steps	vs. ERM	+1.60 ± 1.02	+0.91, +2.58, +0.55, +2.36
Gaussian	steps	vs. AdvDelete	+6.30 ± 6.32	+4.18, +5.56, +0.29, +15.18
Gaussian	steps	vs. RandomDelete	+1.43 ± 1.08	+1.31, +2.74, +0.12, +1.53
Gaussian	steps	vs. SAdD	+1.55 ± 0.87	+2.24, +2.29, +0.50, +1.17
Credit	steps	vs. ERM	−1.27 ± 2.60	-3.62, -0.72, -2.90, +2.17
Credit	steps	vs. AdvDelete	+1.99 ± 2.16	+2.90, +1.45, -0.72, +4.35
Credit	steps	vs. RandomDelete	+1.27 ± 3.71	-3.62, +2.90, +0.72, +5.07
Credit	steps	vs. SAdD	+1.27 ± 2.16	+2.17, +0.72, -1.45, +3.62
Horse	steps	vs. ERM	−0.63 ± 6.66	+8.86, -3.80, -6.33, -1.27
Horse	steps	vs. AdvDelete	−1.58 ± 2.99	+2.53, -3.80, -1.27, -3.80
Horse	steps	vs. RandomDelete	+0.63 ± 5.61	+2.53, +5.06, -7.59, +2.53
Horse	steps	vs. SAdD	+1.27 ± 9.42	-1.27, -5.06, -3.80, +15.19

Table 77: Paired MaDRO-1 accuracy differences (percentage points), matched training time. Positive favors MaDRO-1. Summary entries give the mean and sample SD across four optimization seeds; individual differences appear in seed order 1–4.

Dataset	Budget	Comparator	Mean ± SD	Per-seed differences
WDBC	time	vs. ERM	+0.12 ± 0.61	-0.32, +0.72, +0.56, -0.48
WDBC	time	vs. AdvDelete	−0.12 ± 0.34	+0.08, +0.24, -0.32, -0.48
WDBC	time	vs. RandomDelete	+0.08 ± 0.17	-0.16, +0.16, +0.08, +0.24
WDBC	time	vs. SAdD	+0.74 ± 0.73	+0.32, +0.40, +0.40, +1.83
Gaussian	time	vs. ERM	−0.61 ± 0.69	-1.55, +0.10, -0.50, -0.48
Gaussian	time	vs. AdvDelete	+3.67 ± 1.42	+4.20, +4.18, +1.57, +4.72
Gaussian	time	vs. RandomDelete	−0.51 ± 0.69	-1.15, +0.24, -1.05, -0.10
Gaussian	time	vs. SAdD	+0.08 ± 0.26	-0.10, +0.07, -0.12, +0.45
Credit	time	vs. ERM	+2.72 ± 1.24	+1.45, +4.35, +2.90, +2.17
Credit	time	vs. AdvDelete	+0.54 ± 2.39	+2.17, -0.72, +2.90, -2.17
Credit	time	vs. RandomDelete	+1.45 ± 2.13	-0.72, +3.62, +2.90, +0.00
Credit	time	vs. SAdD	+1.27 ± 1.91	+0.00, +2.90, +2.90, -0.72
Horse	time	vs. ERM	+1.58 ± 2.81	+2.53, -1.27, +0.00, +5.06
Horse	time	vs. AdvDelete	−0.63 ± 2.19	-1.27, -1.27, +2.53, -2.53
Horse	time	vs. RandomDelete	+1.90 ± 3.35	-2.53, +5.06, +3.80, +1.27
Horse	time	vs. SAdD	+3.48 ± 11.39	+0.00, -5.06, -1.27, +20.25

Table 78: Validation-only selections in seed order 1–4. For MaDRO the setting is γ; for SAdD it is hidden flip budget k/weight λ. Checkpoint steps minimize validation loss within that configuration.

Dataset	Method	Selected settings	Best steps
WDBC	MaDRO-1	0.1, 0.2, 0.05, 0.05	2100, 2100, 1500, 1200
WDBC	MaDRO-CV	0.05, 0.2, 0.05, 0.1	2100, 2400, 2550, 1200
WDBC	SAdD	16/1.0, 16/1.0, 4/0.1, 16/1.0	600, 1500, 1950, 3000
Gaussian	MaDRO-1	0.5, 0.2, 0.5, 0.5	3000, 3000, 2700, 2850
Gaussian	MaDRO-CV	0.2, 0.2, 0.2, 0.5	3000, 2550, 3000, 3000
Gaussian	SAdD	4/0.1, 16/0.1, 4/0.1, 4/0.1	2700, 2850, 3000, 2100
Credit	MaDRO-1	0.5, 1.0, 1.0, 0.05	100, 100, 200, 300
Credit	SAdD	16/1.0, 4/0.1, 4/0.1, 16/0.1	1300, 800, 100, 100
Horse	MaDRO-1	0.5, 0.1, 1.0, 0.1	100, 100, 100, 100
Horse	SAdD	4/1.0, 16/0.1, 4/1.0, 4/1.0	100, 100, 1600, 600

Table 79: Measured equal-step configuration-search cost, averaged over seeds. Training excludes validation. Run wall time starts after data/model setup and includes validation, checkpoint writing, and background storage I/O. Separate matched-time runs incur additional computation.

Dataset	Method	Configs	Train s	Wall s
WDBC	ERM	1	31.6	34.9
WDBC	RandomDelete	1	32.7	35.9
WDBC	AdvDelete	1	80.4	83.7
WDBC	Hidden dropout	1	32.3	35.6
WDBC	SAdD	4	135.8	148.9
WDBC	MaDRO-1	5	410.9	427.3
WDBC	MaDRO-CV	5	401.6	417.9
Gaussian	ERM	1	43.3	48.9
Gaussian	RandomDelete	1	43.7	49.3
Gaussian	AdvDelete	1	181.3	187.0
Gaussian	Hidden dropout	1	44.3	49.9
Gaussian	SAdD	4	182.6	205.4
Gaussian	MaDRO-1	5	910.9	939.3
Gaussian	MaDRO-CV	5	910.9	939.4
Credit	ERM	1	13.6	14.2
Credit	RandomDelete	1	14.6	15.2
Credit	AdvDelete	1	25.2	25.8
Credit	Hidden dropout	1	14.2	14.8
Credit	SAdD	4	60.6	63.1
Credit	MaDRO-1	5	137.6	141.4
Horse	ERM	1	15.1	15.8
Horse	RandomDelete	1	15.6	16.3
Horse	AdvDelete	1	32.1	32.8
Horse	Hidden dropout	1	15.8	16.5
Horse	SAdD	4	66.1	68.9
Horse	MaDRO-1	5	162.0	165.6

Table 80: Eight-feature WDBC, exact inner maximization over at most k extra deletions. Positive-cost methods fix $\gamma = 0.2$. Accuracy averages the aligned eleven-rate mixture. Full-8 enumerates all 256 subsets (with duplicates induced by already-missing features).

Method	Accuracy (%)	Training seconds
ERM	90.47 ± 0.88	6.60 ± 0.14
Full-1	90.45 ± 0.32	9.86 ± 0.16
Full-2	90.25 ± 0.18	15.09 ± 0.19
Full-4	90.63 ± 0.29	36.75 ± 0.69
Full-8	90.57 ± 0.55	52.08 ± 0.78

Table 81: Held-out cost-penalized loss increase in the eight-feature study, evaluating each trained model under every probe budget. Each entry is the mean over the fixed test probe, then mean $\pm$ SD over seeds. The 50% base mask leaves four visible features, so probe budgets 4 and 8 have the same feasible set.

Training	Probe $k = 1$	$k = 2$	$k = 4$	$k = 8$
ERM	0.0844 ± 0.0141	0.2144 ± 0.0383	0.3636 ± 0.0800	0.3636 ± 0.0800
Full-1	0.0355 ± 0.0081	0.0944 ± 0.0234	0.1662 ± 0.0250	0.1662 ± 0.0250
Full-2	0.0337 ± 0.0060	0.0742 ± 0.0047	0.1107 ± 0.0135	0.1107 ± 0.0135
Full-4	0.0427 ± 0.0041	0.0715 ± 0.0031	0.0886 ± 0.0032	0.0886 ± 0.0032
Full-8	0.0427 ± 0.0063	0.0720 ± 0.0064	0.0882 ± 0.0044	0.0882 ± 0.0044

Table 82: Equal-step held-out deletion diagnostics using the common unit-cost $\gamma = 0.2$ probe. Maximum loss increase includes no-op; hinge subtracts deletion cost. Activation is the fraction (%) selecting a deletion. Diagnostics are evaluated after validation-based model selection.

Dataset	Method	Max loss increase	Hinge	Activation (%)
WDBC	ERM	0.109 ± 0.039	0.065 ± 0.044	10.526 ± 1.241
WDBC	RandomDelete	0.100 ± 0.023	0.061 ± 0.020	9.868 ± 3.233
WDBC	AdvDelete	0.061 ± 0.008	0.033 ± 0.005	8.114 ± 0.840
WDBC	SAdD	0.089 ± 0.017	0.045 ± 0.017	13.158 ± 2.582
WDBC	MaDRO-1	0.076 ± 0.018	0.037 ± 0.016	10.088 ± 2.532
WDBC	MaDRO-CV	0.073 ± 0.015	0.032 ± 0.012	10.307 ± 1.316
Gaussian	ERM	0.426 ± 0.039	0.251 ± 0.034	70.312 ± 3.932
Gaussian	RandomDelete	0.437 ± 0.046	0.260 ± 0.046	72.461 ± 3.338
Gaussian	AdvDelete	0.196 ± 0.131	0.099 ± 0.081	33.203 ± 25.020
Gaussian	SAdD	0.402 ± 0.051	0.227 ± 0.044	68.750 ± 6.411
Gaussian	MaDRO-1	0.387 ± 0.036	0.212 ± 0.036	67.383 ± 1.732
Gaussian	MaDRO-CV	0.369 ± 0.058	0.192 ± 0.055	67.383 ± 5.038
Credit	ERM	0.497 ± 0.097	0.349 ± 0.102	57.617 ± 6.486
Credit	RandomDelete	0.482 ± 0.075	0.350 ± 0.080	49.414 ± 1.732
Credit	AdvDelete	0.319 ± 0.113	0.227 ± 0.108	29.883 ± 3.900
Credit	SAdD	0.448 ± 0.103	0.323 ± 0.083	49.023 ± 16.845
Credit	MaDRO-1	0.373 ± 0.091	0.210 ± 0.074	64.258 ± 21.503
Horse	ERM	1.132 ± 0.288	1.012 ± 0.291	52.532 ± 5.708
Horse	RandomDelete	0.722 ± 0.174	0.593 ± 0.160	55.063 ± 7.489
Horse	AdvDelete	0.541 ± 0.155	0.412 ± 0.154	49.051 ± 0.633
Horse	SAdD	0.660 ± 0.164	0.526 ± 0.154	52.848 ± 6.319
Horse	MaDRO-1	0.492 ± 0.140	0.366 ± 0.151	50.633 ± 5.168

Table 83: Effect of the training horizon on nominal-mask accuracy (%, mean ± sample SD across four optimization seeds on one fixed split). Each horizon uses minimum-validation-loss checkpoints and independently selects positive γ by validation accuracy. Extended runs follow the validation-stability rule in Appendix N, with method-dependent training times.

	WDBC		Gaussian	
Method	3,000 steps	Extended	3,000 steps	Extended
ERM	93.18 ± 0.58	93.36 ± 0.74	76.71 ± 0.74	79.07 ± 1.26
AdvDelete	93.82 ± 0.33	93.82 ± 0.33	72.01 ± 6.35	76.94 ± 2.04
MaDRO-1	93.88 ± 0.25	93.78 ± 0.20	78.31 ± 0.67	79.69 ± 0.75

N SINGLE-PARTITION TRAINING-HORIZON COMPARISON

Protocol. We compare ERM, AdvDelete, and MaDRO-1 on WDBC and Gaussian data using the fixed splits, preprocessing, TaskNet architecture, aligned eleven-rate mask mixture, Adam learning rate 10^{-3}, batch size 64, and four optimization seeds of Appendix M. Each seed includes ERM, AdvDelete ($\gamma = 0$), and five positive-cost candidates with $\gamma \in \{0.05, 0.1, 0.2, 0.5, 1\}$, giving 56 trajectories. We compare checkpoints from the first 3,000 updates with those from an extended horizon on the same trajectory. Validation is evaluated every 150 updates through update 3,000 and every 500 thereafter. Training runs for at least 15,000 and at most 60,000 updates. An anchor stores the running minimum validation loss and its update index; it resets whenever that minimum improves by more than 0.001 relative to the anchor. After the minimum training horizon, training stops when the anchor has not reset for 10,000 updates. This rule is fixed independently of test performance.

At each horizon, we retain the exact minimum-validation-loss checkpoint, including improvements smaller than the stopping tolerance. Positive γ is selected independently for each horizon and seed by validation accuracy at that checkpoint, breaking ties by lower validation loss and then configuration name. Test evaluation follows completion of all seven candidates for a dataset and seed. All 56 trajectories satisfy the validation-plateau criterion before reaching the 60,000-update cap. This finite stopping rule measures empirical validation stability over the specified patience window. Training time and full configuration-search cost vary across methods.

This single-partition L4 comparison uses a separate checkpoint grid from Appendix K.

Accuracy and variability. On Gaussian data, extending training raises ERM accuracy from 76.71% to 79.07%, AdvDelete from 72.01% to 76.94%, and MaDRO-1 from 78.31% to 79.69% (Table 83). The mean MaDRO-1 advantage over ERM decreases from 1.60 to 0.62 percentage points. Its paired extended-horizon differences across seeds 1–4 are $+2.98, +0.62, +0.67, -1.79$ points, quantifying the optimization variability around the mean improvement. The corresponding differences from AdvDelete are $+1.96, +1.72, +2.58, +4.72$ points, with a mean of $+2.74$. On WDBC, extended-horizon MaDRO-1 exceeds ERM by 0.42 points on average, with paired differences $-0.64, +1.20, +0.24, +0.88$, and differs from AdvDelete by -0.04 points. The clearest effect in this comparison is the Gaussian improvement over AdvDelete across all four seeds; gains over ERM vary with optimization seed, while the WDBC mean is close to AdvDelete. All SDs describe optimization variability conditional on the fixed data split.

Checkpoints and computational cost. The extended-horizon MaDRO-1 selections for seeds 1–4 are $\gamma = (0.05, 0.5, 0.05, 0.1)$ on WDBC and $(0.5, 0.2, 0.5, 1)$ on Gaussian data. Table 84 separates the selected checkpoint from the stopping update. In particular, early selected WDBC checkpoints remain eligible even though every trajectory is trained for at least 15,000 updates. Training time includes the selected configuration's full trajectory and excludes validation, checkpoint writing, and archive I/O; searching the other positive-cost candidates incurs additional cost. All workers use an NVIDIA L4 GPU; one WDBC worker has eight virtual CPUs and the others four.

Interpretation across budgets. The extended-horizon comparison complements the equal-training-time results in Table 72: the latter measure performance within a common compute budget, whereas the former allows each candidate to reach a prespecified validation plateau. Gaussian MaDRO-1 is 0.61 points below ERM under the matched-time protocol and 0.62 points above it on average under

Table 84: Extended-horizon checkpoint and stopping updates (seeds 1–4), with selected-configuration training time in seconds (mean ± sample SD). Updates are in thousands. Search over the five positive-cost configurations is additional.

Dataset	Method	Selected updates	Stopping updates	Time (s)
WDBC	ERM	4.5, 5, 1.35, 1.5	15, 15, 15, 15	165.9 ± 3.9
	AdvDelete	2.25, 2.4, 1.2, 1.2	15, 15, 15, 15	403.9 ± 3.8
	MaDRO-1	23.5, 19.5, 1.5, 9	33.5, 29.5, 15, 19	654.3 ± 239.3
Gaussian	ERM	7, 15, 13.5, 30.5	17, 25, 23.5, 36	371.9 ± 113.1
	AdvDelete	15.5, 14.5, 9.5, 12.5	25.5, 24.5, 19.5, 22.5	1398.3 ± 173.1
	MaDRO-1	11.5, 9.5, 13.5, 19	21.5, 19.5, 23.5, 29	1416.9 ± 235.8

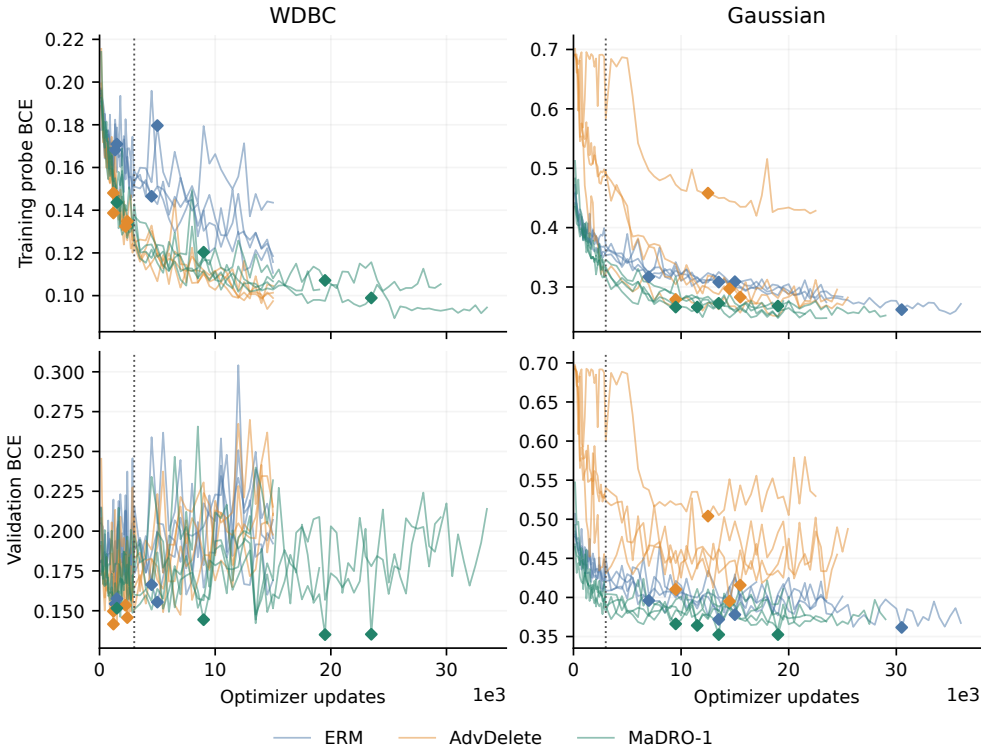

Figure 13: Training-horizon diagnostics for ERM, AdvDelete, and the extended-horizon validation-selected MaDRO-1 configuration in each seed. Each line is one of four optimization seeds and ends at its stopping update. The upper panels show nominal-mask binary cross-entropy (BCE) on a fixed probe of up to 256 training examples; the lower panels show nominal-mask validation BCE. The training probe is distinct from the robust training objective. Both diagnostics use fixed evaluation masks. Diamonds mark the minimum-validation-loss checkpoint; the dotted line marks 3,000 updates. The panels characterize training fit, validation fit, and checkpoint selection.

the extended protocol. This reversal makes the training budget part of the empirical comparison. Figure 13 shows substantial Gaussian improvements after 3,000 updates, motivating assessment at both fixed-compute and extended-training horizons.