AN IDENTIFIABILITY THEORY OF MASKED PREDICTION: MODE BLINDNESS AND MASK SCHEDULES

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked prediction learns to infer missing variables from visible context selected by a mask schedule. When does small masked-prediction excess risk guarantee recovery of the true joint distribution? We study this question on finite product spaces under the joint masked-block log loss, with conditionals induced by a single joint distribution. To quantify recovery, we introduce an ε-identifiability modulus that measures the worst-case error among joint distributions with population excess risk at most ε. For data with separated modes pinned down by sufficient context, mask schedules supported only on large contexts can permit nonvanishing mode-weight errors at an exponentially small excess risk. An information decomposition explains why: the loss captures mode-weight mismatch only where context leaves the mode uncertain. We prove two-sided bounds showing that, over a fixed range of mode weights, masked-prediction sensitivity to mode reweighting is governed by schedule-averaged residual mode uncertainty. Assigning schedule mass to low-visibility masks that retain this uncertainty yields recovery bounds within the reweighting family. Moreover, positive full-mask probability characterizes uniform control of joint KL divergence by excess risk. Exact calculations and controlled gradient optimization validate these predictions.

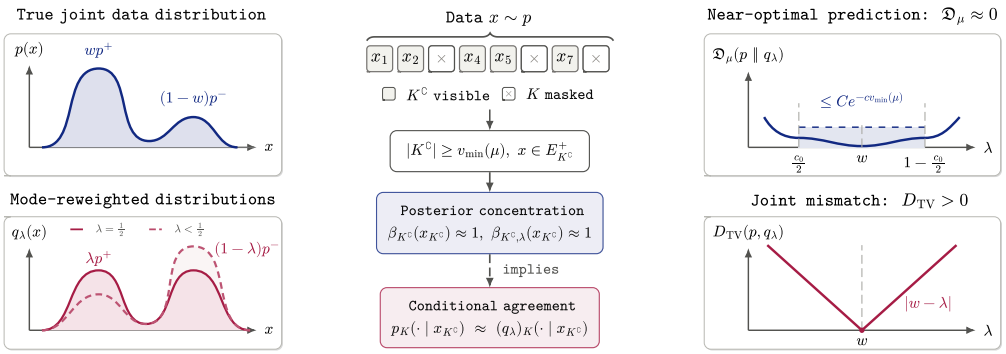
1 INTRODUCTION

Masked prediction replaces direct joint modeling with predicting missing variables from partial observations, with a *mask schedule* governing the distribution of revealed coordinates. This training paradigm underlies BERT-style pretraining (Devlin et al., 2019; Liu et al., 2019), masked autoencoding (He et al., 2022), absorbing-state discrete diffusion (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024; Ou et al., 2025), and masked latent prediction (Assran et al., 2023; Bardes et al., 2024; Nam et al., 2026). Despite architectural differences, these methods raise a fundamental question: when does small masked-prediction excess risk guarantee recovery of the true joint distribution?

A simple thought experiment shows this cannot hold unconditionally. Suppose the data contain two well-separated regimes, such as code and prose, and we perturb only their relative mixture weights while keeping distributions fixed within regimes. With little visible context, a model must hedge its predictions across regimes, scaling completion probabilities directly by global mode weights. Once the context pins down the active regime, however, its global weight has little effect on the masked conditional. Increasing visible context thus makes conditional inference easier while eroding the supervisory signal required to recover global weights. Consequently, a model can achieve near-optimal loss while incurring substantial errors in mode weights, a pathology we term *mode blindness*.

To quantify this gap, we introduce an *ε-identifiability modulus*, which captures the worst-case recovery error over joint distributions with excess risk at most ε. Specifically, the modulus vanishes at $\varepsilon = 0$ when the objective identifies the true distribution; for $\varepsilon > 0$, it measures the stability of recovery. Prior work relies on functional inequalities to upper-bound this modulus, translating training objective discrepancies into joint recovery error. Koehler et al. (2023) analyze score matching and pseudolikelihood from this perspective, showing that score matching fails to recover mixture weights. For masked prediction, Li et al. (2024) show that larger masked blocks improve asymptotic statistical efficiency and derive joint recovery bounds through block approximate tensorization. Complementing these upper bounds, we show how the mask schedule governs the excess-risk scale at which nonvanishing mode-weight errors can persist.

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



(a) Mode-reweighting family (b) Context ↑, mode uncertainty ↓ (c) Mode blindness

Figure 1: Mode blindness in masked prediction. (a) The true data distribution p and a mode-reweighting distribution q_λ share mode conditionals but differ in their global mode weights. (b) Once the visible context leaves little uncertainty about the mode, the masked conditionals under p and q_λ become nearly indistinguishable. (c) The masked discrepancy can therefore decay exponentially with visible-context size while the joint distributions remain a constant total-variation distance apart. See Thm. 4.1 for the statement and Appx. E for the extension to $M \geq 2$ modes.

We characterize this schedule dependence under joint masked-block log loss, where model conditionals derive from a single joint distribution. The analysis focuses on the mode-reweighting family, which changes only the mixture weights while keeping the distribution within each mode fixed. By the KL chain rule, the joint mismatch separates into a marginal term for the visible variables and a conditional term for the masked variables; only the latter contributes to the learning objective. The strength of this signal is governed by the residual mode uncertainty, defined as the expected posterior variance of the mode indicator. We prove that the excess risk is bounded above and below by the KL divergence between mode weights scaled by the schedule-averaged residual mode uncertainty. These two-sided bounds hold for arbitrary mask schedules and uniformly over a fixed compact interval of mode weights containing the truth, with constants depending only on that interval.

Under mode pinning, sufficiently large visible contexts leave exponentially small residual mode uncertainty. Mask schedules that always retain such large contexts admit nonvanishing mode-weight errors at an excess risk that decays exponentially with the minimum visible-context size (Fig. 1). This forces the associated approximate-tensorization constants to grow at least exponentially. Conversely, when low-visibility masks—those leaving few coordinates visible—retain nontrivial mode uncertainty, assigning them positive schedule mass yields explicit mode-weight recovery bounds within the mode-reweighting family. Further, full masking directly scores the joint distribution. Positive full-mask probability characterizes uniform control of joint KL divergence by excess risk. Mask schedules thus govern both prediction difficulty and the stability of joint-distribution recovery.

Contributions. **(i)** We introduce an ε-identifiability modulus measuring the worst-case joint-distribution error at an excess risk at most ε, distinguishing exact identification from stable recovery (Defn. 2.2). **(ii)** We establish exponential mode blindness under mode pinning for schedules revealing only large contexts (Thm. 4.1), yielding exponential lower bounds on block approximate-tensorization constants in the bimodal setting (Cor. 4.1). **(iii)** For bimodal reweightings over a fixed weight interval, we prove two-sided sensitivity bounds in terms of schedule-averaged residual mode uncertainty (Thm. 4.2), yielding recovery bounds from low-visibility mass within this family. Under joint masked-block log loss, positive full-mask probability further characterizes distribution-free control of joint KL divergence (Prop. 4.1). **(iv)** We validate mode blindness, its recovery under low-visibility masking, and predictions for recovery speed through exact calculations and controlled gradient descent within the mode-reweighting family (Sec. 5).

2 Problem Setup

This section formulates the masked-prediction objective on a finite product space and introduces our theoretical framework. Appx. A outlines the notation and mathematical conventions used through-

out, while Appx. B reviews related literature. Auxiliary lemmas and complete proofs of all results in the main text are deferred to Appx. C and Appx. D, respectively.

Data space and mask schedule. Let $\mathcal{A}$ be a finite alphabet with $|\mathcal{A}| \geq 2$, and let $\mathcal{X} := \mathcal{A}^N$ be the space of length-N sequences. Let $p \in \Delta(\mathcal{X})$ denote the true data distribution, and write $X = (X_1, \ldots, X_N) \sim p$. A *mask schedule* is a probability distribution μ over subsets $K \subseteq [N]$, where K indexes the masked coordinates and $K^{\complement} := [N] \setminus K$ indexes the visible coordinates.[1] For a *visibility threshold* $s \in \{0, \ldots, N\}$, define

$$v_{\min}(\mu) := \min_{K \in \mathrm{supp}(\mu)} |K^{\complement}|, \qquad \pi_s(\mu) := \mu\big(|K^{\complement}| \leq s\big). \tag{1}$$

Here, $v_{\min}(\mu)$ is the minimum visible-context size under schedule μ, and $\pi_s(\mu)$ is the probability of observing at most s visible coordinates. We refer to $\pi_s(\mu)$ as the *low-visibility mass*; in particular, $\pi_0(\mu) = \mu(K = [N])$ is the *full-mask probability* (where the visible context is empty, $|K^{\complement}| = 0$).

Admissible model distributions. A model is represented by a joint distribution $q \in \Delta(\mathcal{X})$ via its masked-block conditionals $q_K(\cdot \mid x_{K^{\complement}})$. We call support-covering models *admissible* and define

$$\mathcal{Q}(p) := \{q \in \Delta(\mathcal{X}) : \mathrm{supp}(p) \subseteq \mathrm{supp}(q)\}. \tag{2}$$

For every $q \in \mathcal{Q}(p)$, every mask K, and each context $x_{K^{\complement}}$ with $p_{K^{\complement}}(x_{K^{\complement}}) > 0$, the model conditional $q_K(\cdot \mid x_{K^{\complement}})$ is well defined and satisfies $p_K(\cdot \mid x_{K^{\complement}}) \ll q_K(\cdot \mid x_{K^{\complement}})$. Restricting model conditionals to derive from a joint distribution isolates schedule-induced information loss from the issue of incompatible conditionals. This strengthens our negative results since our mode-blind constructions are themselves valid joint distributions, while properly scoping our positive guarantees.

Definition 2.1 (Masked log discrepancy)**.** For $q \in \mathcal{Q}(p)$ and a fixed mask $K \subseteq [N]$, define the fixed-mask log discrepancy

$$\mathfrak{D}_K(p\|q) := \mathbb{E}_{X_{K^{\complement}} \sim p_{K^{\complement}}} \left[D_{\mathrm{KL}}(p_K(\cdot \mid X_{K^{\complement}}) \,\|\, q_K(\cdot \mid X_{K^{\complement}})) \right]. \tag{3}$$

For a mask schedule μ, define the schedule-averaged masked log discrepancy

$$\mathfrak{D}_\mu(p\|q) := \mathbb{E}_{K \sim \mu} \left[\mathfrak{D}_K(p\|q) \right]. \tag{4}$$

Remark 2.1 (Excess-risk interpretation). We evaluate masked prediction via the joint masked-block logarithmic loss $-\log q_K(X_K \mid X_{K^{\complement}})$. For every $q \in \mathcal{Q}(p)$,

$$\mathfrak{D}_\mu(p\|q) = \mathbb{E}_{K \sim \mu} \mathbb{E}_{X \sim p} \left[-\log q_K(X_K \mid X_{K^{\complement}}) + \log p_K(X_K \mid X_{K^{\complement}}) \right]. \tag{5}$$

Hence, $\mathfrak{D}_\mu(p\|q)$ corresponds to the *population masked-prediction excess risk* of model distribution q relative to data distribution p: the mask K is sampled from μ, data $X \sim p$ provides both context and target, and the model q is scored on the joint conditional. Throughout, our primary setting scores the entire masked block jointly (see Appx. C.7 for extensions to token-factorized heads).

ε-Identifiability modulus. While the discrepancy $\mathfrak{D}_\mu(p\|q)$ measures conditional agreement on visible contexts drawn from p, near-zero conditional excess risk does not inherently imply that the joint distributions are close. To analyze this stability of identification, we introduce a modulus that quantifies the worst-case joint error compatible with a given excess-risk budget ε.

Definition 2.2 (ε-identifiability modulus)**.** Fix a schedule μ and a metric d continuous on $\Delta(\mathcal{X}) \times \Delta(\mathcal{X})$. For a data distribution $p \in \Delta(\mathcal{X})$ and risk budget $\varepsilon \geq 0$, the ε-identifiability modulus is

$$\mathfrak{A}_d^\mu(p, \varepsilon) := \sup \{d(p, q) : q \in \mathcal{Q}(p),\ \mathfrak{D}_\mu(p\|q) \leq \varepsilon\}. \tag{6}$$

Because $\Delta(\mathcal{X})$ is compact and d is continuous, $\mathfrak{A}_d^\mu(p, \varepsilon)$ is always finite and nondecreasing in ε (see Appx. C.1 for the properties). The modulus distinguishes two types of guarantees:

- A *lower bound* on $\mathfrak{A}_d^\mu(p, \varepsilon)$ requires constructing just one candidate distribution $q \in \mathcal{Q}(p)$ that stays within the risk budget ε while remaining distant from p under d.
- An *upper bound* must uniformly control *every* admissible model inside that risk budget.

At risk budget $\varepsilon = 0$, Defn. 2.2 reduces to exact objective-level identifiability: $\mathfrak{A}_d^\mu(p, 0) = 0$ if and only if p is uniquely identified by zero masked discrepancy. For $\varepsilon > 0$, it characterizes the stability of joint-distribution recovery under near-optimal prediction. Although defined here for masked log loss and joint-distribution recovery, the modulus extends naturally to other objectives and targets.

[1] Notation: while $K \subseteq [N]$ and $K^{\complement}$ strictly denote a mask drawn from μ and its visible coordinates, we use V and $V^{\complement} = [N] \setminus V$ to denote general, schedule-independent conditioning sets.

3 From Approximate Tensorization to the Bimodal Regime

In this section, we first examine the identifiability modulus when block approximate tensorization converts masked discrepancy into joint-distribution error. We then introduce a bimodal data geometry where this local-to-global control weakens, formalizing residual mode uncertainty and a mode-reweighting family whose masked discrepancy admits an exact information decomposition.

3.1 Quantitative Identifiability via Block Approximate Tensorization

Block heat-bath dynamics. Given a distribution $r \in \Delta(\mathcal{X})$, a configuration $x \in \mathrm{supp}(r)$, and a mask K, a block heat-bath update holds the visible coordinates fixed and redraws the masked block according to

$$\widetilde{X} = (\widetilde{X}_{K^\complement}, \widetilde{X}_K), \qquad \widetilde{X}_{K^\complement} = x_{K^\complement}, \qquad \widetilde{X}_K \sim r_K(\cdot \mid x_{K^\complement}). \tag{7}$$

Sampling $K \sim \mu$ at each step defines the block dynamics induced by (r, μ). Because each fixed-mask update is reversible with respect to r, their mixture under μ remains reversible, rendering r stationary for the resulting dynamics. Taking $r = p$ or $r = q$ yields the data and model block dynamics, respectively.

Approximate tensorization. We say a distribution $r \in \Delta(\mathcal{X})$ satisfies block *approximate tensorization (AT)* with constant $\bar{C}_{\mathrm{AT}}(r, \mu) \in [0, \infty]$, defined as the smallest constant such that

$$D_{\mathrm{KL}}(p' \,\|\, r) \;\le\; \bar{C}_{\mathrm{AT}}(r, \mu)\, \mathfrak{D}_\mu(p' \,\|\, r) \qquad \text{for every } p' \in \Delta(\mathcal{X}) \text{ with } \mathrm{supp}(p') \subseteq \mathrm{supp}(r). \tag{8}$$

This formulation represents the density-ratio form of the block factorization of relative entropy. The size of $\bar{C}_{\mathrm{AT}}(r, \mu)$ determines how strongly conditional discrepancies control joint KL divergence (Caputo & Parisi, 2021; Li et al., 2024). Applying Eq. (8) with reference distribution $r = q$ for an admissible model and $p' = p$ converts the masked discrepancy into a joint-distribution bound, which via Pinsker's inequality yields[2]

$$\mathfrak{A}^{\mu}_{\mathrm{TV}}(p, \varepsilon) \le \sqrt{\tfrac{1}{2}\, \bar{C}^{\varepsilon}_{\mathrm{AT}}(p, \mu)\, \varepsilon}, \quad \bar{C}^{\varepsilon}_{\mathrm{AT}}(p, \mu) := \sup\big\{\bar{C}_{\mathrm{AT}}(q, \mu) : q \in \mathcal{Q}(p),\ \mathfrak{D}_\mu(p \| q) \le \varepsilon\big\}. \tag{9}$$

The AT constant is evaluated at the model distribution q because Eq. (8) uses q as the reference measure. See Appx. C.2 for the formal transfer and comparisons to classical functional inequalities.

For any sequence (p_N, μ_N) and excess-risk budgets ε_N, Eq. (9) guarantees vanishing total-variation error whenever $\bar{C}^{\varepsilon_N}_{\mathrm{AT}}(p_N, \mu_N)\varepsilon_N \to 0$. We next introduce a bimodal data geometry for which the AT constants grow exponentially with the minimum visible-context size.

3.2 Bimodal Geometry and the Mode-Reweighting Direction

Suppose the data distribution exhibits two well-separated global regimes, partitioning the sample space into disjoint mode regions $\mathcal{X} = \mathcal{X}_+ \sqcup \mathcal{X}_-$ with $p(\mathcal{X}_\tau) > 0$ for each $\tau \in \{+, -\}$. This geometric partition is a property of p and independent of any mask schedule. Let $w := p(\mathcal{X}_+) \in (0, 1)$ denote the true global mode weight. For each mode $\tau \in \{+, -\}$, we write $-\tau$ for the complementary mode, define the within-mode conditional distribution $p^\tau(\cdot) := p(\cdot \mid \mathcal{X}_\tau)$, and let p^τ_V denote its marginal over coordinates $V \subseteq [N]$.

Residual mode uncertainty. To quantify the residual ambiguity regarding the active mode after observing context on $V \subseteq [N]$, define the binary mode indicator $Z := \mathbb{1}\{X \in \mathcal{X}_+\}$ and its posterior probability

$$\beta_V(x_V) := p(Z = 1 \mid X_V = x_V). \tag{10}$$

The *residual mode uncertainty* is captured by the conditional minimum mean-square error (MMSE):

$$\mathrm{mmse}_p(Z \mid X_V) := \mathbb{E}_p\big[(Z - \mathbb{E}_p[Z \mid X_V])^2\big] = \mathbb{E}_{X_V \sim p_V}\big[\beta_V(X_V)(1 - \beta_V(X_V))\big]. \tag{11}$$

This variance vanishes if and only if the visible context determines the mode almost surely. For a fixed mode weight w, it is maximized at $w(1 - w)$ when X_V carries no information about Z.

[2]Hereafter, we specialize to total variation distance $d = D_{\mathrm{TV}}$ and write $\mathfrak{A}^{\mu}_{\mathrm{TV}} := \mathfrak{A}^{\mu}_{D_{\mathrm{TV}}}$.

Mode-weight perturbations. For any $\lambda \in (0,1)$, define the mode-reweighted distribution

$$q_\lambda := \lambda\, p^+ + (1-\lambda)\, p^-. \tag{12}$$

Because the mode supports are disjoint, q_λ preserves the within-mode distributions of p exactly and changes only the global mixture weight to $q_\lambda(\mathcal{X}_+) = \lambda$. Furthermore, $\lambda \in (0,1)$ guarantees $\mathrm{supp}(q_\lambda) = \mathrm{supp}(p)$, ensuring $q_\lambda \in \mathcal{Q}(p)$. As in Eq. (10), let $\beta_{\lambda,V}(x_V) := q_\lambda(Z = 1 \mid X_V = x_V)$ denote the posterior mode probability under q_λ.

The following lemma shows that mode reweighting acts as an exact logit tilt for each visible context, reducing the conditional KL divergence between masked blocks to the binary KL divergence between their mode posteriors.

Lemma 3.1 (Posterior shift under mode reweighting)**.** *Let q_λ be defined by Eq. (12) for $\lambda \in (0,1)$. Fix $V \subseteq [N]$, let $V^{\complement} = [N] \setminus V$, and take $x_V \in \mathrm{supp}(p_V)$. The following properties hold:*

(i) *The posterior mode probability under q_λ satisfies*

$$\beta_{\lambda,V}(x_V) = \frac{\frac{\lambda}{w}\,\beta_V(x_V)}{\frac{\lambda}{w}\,\beta_V(x_V) + \frac{1-\lambda}{1-w}\,\big(1-\beta_V(x_V)\big)}. \tag{13}$$

(ii) *If $\beta_V(x_V) \in (0,1)$, reweighting induces a logit tilt:*

$$\mathrm{logit}\,\big(\beta_{\lambda,V}(x_V)\big) = \mathrm{logit}\,\big(\beta_V(x_V)\big) + \mathrm{logit}(\lambda) - \mathrm{logit}(w). \tag{14}$$

(iii) *The conditional masked-block distributions satisfy*

$$D_{\mathrm{KL}}\big(p_{V^{\complement}}(\cdot \mid x_V) \,\|\, q_{\lambda,V^{\complement}}(\cdot \mid x_V)\big) = \mathrm{kl}\big(\beta_V(x_V) \,\|\, \beta_{\lambda,V}(x_V)\big), \tag{15}$$

where $\mathrm{kl}(\cdot\|\cdot)$ is the binary KL divergence (Eq. (36)). In particular, whenever $\beta_V(x_V) \in \{0,1\}$, we have $\beta_{\lambda,V}(x_V) = \beta_V(x_V)$ and the divergence in Eq. (15) vanishes identically.

Proposition 3.1 (Mode-weight information decomposition)**.** *Let q_λ be defined as in Eq. (12). For every mask $K \subseteq [N]$ with visible context coordinates $K^{\complement} = [N] \setminus K$,*

$$\mathfrak{D}_K(p\|q_\lambda) \;=\; \mathbb{E}_{X_{K^{\complement}} \sim p_{K^{\complement}}}\big[\mathrm{kl}\big(\beta_{K^{\complement}}(X_{K^{\complement}}) \,\|\, \beta_{\lambda,K^{\complement}}(X_{K^{\complement}})\big)\big], \tag{16}$$

$$\mathrm{kl}(w\|\lambda) \;=\; D_{\mathrm{KL}}(p_{K^{\complement}}\|q_{\lambda,K^{\complement}}) + \mathfrak{D}_K(p\|q_\lambda). \tag{17}$$

Remark 3.1 (KL decomposition under conditioning). Eq. (17) specializes the KL chain rule to mode reweighting: the visible-marginal term accounts for the mismatch in the context distribution, while the masked objective scores only the conditional remainder. When visible context nearly resolves mode uncertainty, almost all global distinguishability is absorbed by the visible marginal, leaving a correspondingly negligible supervisory signal. Conversely, at the full-mask endpoint ($K = [N]$), the visible context is empty and absorbs no information, yielding $\mathfrak{D}_{[N]}(p\|q_\lambda) = \mathrm{kl}(w\|\lambda)$. This decomposition governs both sides of our analysis: mode pinning diminishes the residual training signal, while preserving mode uncertainty restores objective-level sensitivity to the global structure.

4 Mode Blindness and Schedule-Controlled Recovery

We now characterize how masking controls the visibility of global mode weights. Large visible contexts make mode reweighting exponentially hard to detect, low-visibility masks restore this sensitivity, and the full-mask probability yields universal control across all admissible models.

4.1 Mode Blindness of Masked Prediction

Mode blindness arises when large visible contexts leave exponentially little uncertainty about the active global mode. We formalize this geometry through the following mode-pinning condition.

Assumption 4.1 (Large-context mode pinning)**.** *There exist constants $c_0 \in (0, 1/2]$, $c_1, \kappa > 0$, and a visibility threshold $s_{\mathrm{pin}} \in [N]$ such that two conditions hold. First, the global mode weights are strictly bounded away from zero: $p(\mathcal{X}_\tau) \geq c_0$ for both modes $\tau \in \{+,-\}$. Second, for every visible set $V \subseteq [N]$ of size $|V| \geq s_{\mathrm{pin}}$ and each mode τ, the typical set of visible contexts,*

$$\mathcal{E}_V^\tau := \big\{x_V \in \mathrm{supp}(p_V^\tau) : p(\mathcal{X}_{-\tau} \mid X_V = x_V) \leq e^{-\kappa|V|}\big\}, \tag{18}$$

has probability at least $1 - e^{-c_1|V|}$ under p_V^τ.

For asymptotic statements under mode pinning, the constants c_0, c_1, κ are independent of N. Throughout, we denote the combined decay rate by $c := \min\{c_1, \kappa\}$, and note that Asm. 4.1 implies $w \in [c_0, 1 - c_0]$. Uniformity over all sufficiently large visible sets makes the condition schedule-independent, though it can be restrictive when some large visible sets remain weakly informative about the mode. See Appx. G.1 for further discussion of its interpretation.

Remark 4.1 (Mode-pinning bottleneck). A visible context typical of mode $\mathcal{X}_\tau$ assigns exponentially small posterior probability to the opposite mode. Consequently, a block heat-bath update that conditions on such a context remains trapped in $\mathcal{X}_\tau$ with overwhelming probability. Cross-mode transitions are exponentially unlikely on typical large contexts; low-visibility masks and atypical contexts contribute the remaining terms in the transition bound. Concretely, letting $\widetilde{X}$ denote one step of the data block dynamics (Eq. (7)), Lemma C.4 yields

$$\mathbb{P}(\widetilde{X} \in \mathcal{X}_{-\tau} | X \in \mathcal{X}_\tau) \le \pi_s(\mu) + 2e^{-cs}, \qquad \text{for every} \quad s \ge s_{\mathrm{pin}}. \tag{19}$$

When the schedule assigns little probability to low-visibility contexts, the partition $\mathcal{X}_+ \sqcup \mathcal{X}_-$ forms a low-flow bottleneck. This metastable reading motivates the slow-mixing interpretation. Our objective-level arguments below use only the static geometry.

Mode reweighting as a blind direction. By Prop. 3.1, the discrepancy of a reweighted distribution corresponds to the expected binary KL divergence between the posterior mode probabilities under p and q_λ. The family $\{q_\lambda\}$ therefore isolates the global mode-weight direction: every member preserves the within-mode distributions of p and perturbs only their relative proportion. Under Asm. 4.1, large visible contexts leave exponentially small uncertainty regarding the mode, rendering the posterior mode probability nearly invariant to perturbations in the prior weight. Averaging over contexts and masks yields the following result (see also Appx. E for the extension to $M \ge 2$ modes):

Theorem 4.1 (Mode blindness of masked prediction)**.** *Suppose Asm. 4.1 holds with combined rate $c = \min\{c_1, \kappa\}$. Consider the mode-reweighting family $\{q_\lambda \in \mathcal{Q}(p) : \lambda \in I_{\mathrm{bl}}\}$, where q_λ is defined in Eq. (12) and $I_{\mathrm{bl}} := [c_0/2,\, 1 - c_0/2]$. There exists a constant $C \le 6c_0^{-2}$, depending only on c_0, such that every mask schedule μ with $v_{\min}(\mu) \ge s_{\mathrm{pin}}$ satisfies*

$$\sup_{\lambda \in I_{\mathrm{bl}}} \mathfrak{D}_\mu(p \| q_\lambda) \;\le\; C\, \mathbb{E}_{K \sim \mu}\big[e^{-c|K^{\mathfrak{c}}|}\big] \;\le\; C e^{-c\, v_{\min}(\mu)}. \tag{20}$$

Hence, the modulus (Defn. 2.2) remains constant-order even at this exponentially small excess risk:

$$\mathfrak{A}^{\mu}_{\mathrm{TV}}(p, \varepsilon) \;\ge\; \frac{1 - c_0}{2}, \qquad \textit{for every} \quad \varepsilon \ge C\, \mathbb{E}_{K \sim \mu}\big[e^{-c|K^{\mathfrak{c}}|}\big]. \tag{21}$$

Together, Eqs. (20) and (21) establish the exponential instability of population identification: a constant mode-weight error (and therefore constant total-variation distance) persists at a masked excess risk that decays exponentially with the minimum visible context size. Because the reweighting family preserves all within-mode conditionals, this failure isolates a global blind direction rather than an optimization failure. This construction also demonstrates why the AT recovery guarantee in Sec. 3.1 must deteriorate under this well-separated data regime:

Corollary 4.1 (Exponential lower bounds on block AT constants)**.** *Suppose Asm. 4.1 holds with combined rate $c = \min\{c_1, \kappa\}$. Let μ be a mask schedule with $v_{\min}(\mu) \ge s_{\mathrm{pin}}$. Then there exists $\lambda \in I_{\mathrm{bl}}$ such that*

$$\bar{C}_{\mathrm{AT}}(q_\lambda, \mu) \;\ge\; \frac{c_0^2(1 - c_0)^2}{12}\, e^{c\, v_{\min}(\mu)} \quad \textit{and} \quad \bar{C}_{\mathrm{AT}}(p, \mu) \;\ge\; \frac{c_0^4(1 - c_0)^2}{32}\, e^{c\, v_{\min}(\mu)}. \tag{22}$$

Remark 4.2 (Interpretation of the AT lower bounds). The two bounds in Cor. 4.1 play distinct roles. Set $\varepsilon_\mu := Ce^{-c\, v_{\min}(\mu)}$, with C from Thm. 4.1. Since the reweighted distribution q_λ lies in the ε_μ-discrepancy ball, the first bound yields $\bar{C}^{\varepsilon_\mu}_{\mathrm{AT}}(p, \mu)\, \varepsilon_\mu \ge \bar{C}_{\mathrm{AT}}(q_\lambda, \mu)\, \varepsilon_\mu \ge Cc_0^2(1 - c_0)^2/12$. Thus, Eq. (9) cannot certify vanishing joint-distribution error at the mode-blindness scale. The second bound uses Lemma C.5 to show that this blowup also occurs for the data distribution itself. Along any sequence with $v_{\min}(\mu_N) = \Omega(N)$, the data-side AT constants grow at least as $e^{\Omega(N)}$. Hence AT-based guarantees requiring constants uniformly bounded in N cannot cover this family.

4.2 Residual Mode Uncertainty Governs Mode-Weight Sensitivity

Mode blindness demonstrates that large visible contexts suppress objective-level sensitivity to global mode weights. The information decomposition in Prop. 3.1 isolates the quantity governing what remains: the residual uncertainty about the mode after conditioning on the visible context.

Lemma 4.1 (Reweighting sensitivity bounds)**.** *Fix a compact interval $I = [a, b] \subset (0, 1)$ and define*

$$\delta_I := \operatorname{logit}(b) - \operatorname{logit}(a), \qquad m_I := \min_{u \in I} u(1-u), \qquad c_I := 4e^{-\delta_I}, \qquad C_I := \frac{e^{\delta_I}}{m_I}.$$

Whenever $w, \lambda \in I$, every mask $K \subseteq [N]$ with visible context $K^{\complement}$ satisfies

$$c_I \operatorname{mmse}_p(Z \mid X_{K^{\complement}}) \operatorname{kl}(w\|\lambda) \;\le\; \mathfrak{D}_K(p\|q_\lambda) \;\le\; C_I \operatorname{mmse}_p(Z \mid X_{K^{\complement}}) \operatorname{kl}(w\|\lambda). \tag{23}$$

Remark 4.3 (Exact local sensitivity coefficient). The sensitivity bounds in Eq. (23) require no mixing or visibility assumptions; they follow entirely from the bimodal mixture structure. Furthermore, for every mask K with $\operatorname{mmse}_p(Z \mid X_{K^{\complement}}) > 0$, a second-order expansion as $\lambda \to w$ yields

$$\lim_{\lambda \to w} \frac{\mathfrak{D}_K(p\|q_\lambda)}{\operatorname{mmse}_p(Z \mid X_{K^{\complement}}) \operatorname{kl}(w\|\lambda)} = \frac{1}{w(1-w)}, \tag{24}$$

independently of the mask K. The quantity $\mathcal{I}(w) := (w(1-w))^{-1}$ is the Fisher information of the Bernoulli mode weight. Thus, at fixed w, visible contexts with smaller residual mode MMSE have weaker local sensitivity to changes in the global mode weight. See Remark D.1 for the interval dependence of Eq. (23) and the underlying reduction.

High- and low-visibility regimes. We next characterize residual mode uncertainty across the two visibility regimes. Mode pinning governs the large-context regime:

Corollary 4.2 (MMSE upper bound under mode pinning)**.** *Under Asm. 4.1, every visible set $V \subseteq [N]$ satisfying $|V| \ge s_{\text{pin}}$ obeys $\operatorname{mmse}_p(Z \mid X_V) \le \frac{5}{4} e^{-c|V|}$.*

At low visibility, recovery requires the complementary behavior—small contexts must retain non-trivial mode uncertainty:

Assumption 4.2 (Low-visibility residual mode uncertainty)**.** *There exist constants $u_0, L > 0$ and a visibility scale $s_{\text{ov}} \ge s_{\text{pin}}$ such that every $V \subseteq [N]$ with $|V| \le s_{\text{ov}}$ satisfies*

$$\operatorname{mmse}_p(Z \mid X_V) \ge u_0 \, e^{-L|V|}. \tag{25}$$

Asm. 4.2 requires residual mode uncertainty not to decay too rapidly in the low-visibility regime. This condition is compatible with large-context mode pinning: Appx. G.3 constructs an explicit distribution satisfying both assumptions with constants independent of N. Appxs. C.8 and G.2 give further interpretations and sufficient conditions.

Averaging the fixed-mask sensitivity bounds in Eq. (23) over the mask schedule yields our mode-weight sensitivity theorem:

Theorem 4.2 (Schedule-controlled mode-weight sensitivity)**.** *Fix a compact interval $I \subset (0, 1)$ containing w, and let c_I, C_I be the constants from Lemma 4.1. For every mask schedule μ and every parameter $\lambda \in I$,*

$$c_I \operatorname{kl}(w\|\lambda) \, \mathbb{E}_{K\sim\mu}\big[\operatorname{mmse}_p(Z \mid X_{K^{\complement}})\big] \;\le\; \mathfrak{D}_\mu(p\|q_\lambda) \;\le\; C_I \operatorname{kl}(w\|\lambda) \, \mathbb{E}_{K\sim\mu}\big[\operatorname{mmse}_p(Z \mid X_{K^{\complement}})\big]. \tag{26}$$

Under Asm. 4.2, every threshold $0 \le s \le s_{\text{ov}}$ satisfies

$$\mathfrak{D}_\mu(p\|q_\lambda) \;\ge\; c_I \, u_0 \, e^{-Ls} \, \pi_s(\mu) \operatorname{kl}(w\|\lambda). \tag{27}$$

Under Asm. 4.1, every threshold $s \ge s_{\text{pin}}$ satisfies

$$\mathfrak{D}_\mu(p\|q_\lambda) \;\le\; \frac{5C_I}{4}\Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|} \, \mathbb{1}\{|K^{\complement}| > s\}\big]\Big) \operatorname{kl}(w\|\lambda). \tag{28}$$

Eq. (26) gives two-sided sensitivity control for two-mode reweightings under arbitrary schedules, *without* mode-pinning or visibility assumptions. Under the respective assumptions, Eqs. (27) and (28) further express this scale through low-visibility mass $\pi_s(\mu)$ and a high-visibility remainder. The multimode distinction is discussed in Appx. E.4. For fixed s, common mask schedules can assign polynomially small, exponentially small, or zero low-visibility mass (see Appx. F for schedule-specific calculations).

Corollary 4.3 (Mode-weight recovery from low-visibility mass)**.** *Adopt the setting of Thm. 4.2 and the constants c_I, C_I, and m_I of Lemma 4.1. Fix a mask schedule μ, a weight $\lambda \in I$, and an excess-risk tolerance $\varepsilon \geq 0$. Under Asm. 4.2, for every $0 \leq s \leq s_{\mathrm{ov}}$ with $\pi_s(\mu) > 0$, the condition $\mathfrak{D}_\mu(p\|q_\lambda) \leq \varepsilon$ guarantees*

$$|w - \lambda| \;\leq\; \frac{e^{Ls/2}}{\sqrt{2c_I u_0}} \sqrt{\frac{\varepsilon}{\pi_s(\mu)}}. \tag{29}$$

Conversely, under Asm. 4.1, for every $s \geq s_{\mathrm{pin}}$, the bound $\mathfrak{D}_\mu(p\|q_\lambda) \leq \varepsilon$ holds whenever

$$\frac{5C_I}{4m_I}\Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\,\mathbb{1}\{|K^{\complement}| > s\}\big]\Big)|w-\lambda|^2 \leq \varepsilon. \tag{30}$$

Interpretation of mode-weight recovery. For fixed s, I, u_0, and L, Eq. (29) bounds the mode-weight error by $|w - \lambda| = O\big(\sqrt{\varepsilon/\pi_s(\mu)}\big)$. Thus, mode-weight errors within $\{q_\lambda : \lambda \in I\}$ vanish whenever $\varepsilon/\pi_s(\mu) \to 0$. Conversely, Eq. (30) gives a sufficient condition for a mode-weight perturbation to remain within the ε-discrepancy ball. When both bounds apply at the same s and

$$\mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\,\mathbb{1}\{|K^{\complement}| > s\}\big] = O\big(\pi_s(\mu)\big),$$

they have matching square-root dependence on $\varepsilon/\pi_s(\mu)$. These results characterize sensitivity along the one-dimensional reweighting family $\{q_\lambda\}$, with the within-mode distributions held fixed. We next ask when the masked objective controls joint-distribution error uniformly over all $q \in \mathcal{Q}(p)$.

4.3 Full-Mask Boundary: Distribution-Free KL Control

At the full-mask endpoint $K = [N]$, the visible context is empty and $\mathfrak{D}_{[N]}(p\|q) = D_{\mathrm{KL}}(p\|q)$. Hence, positive full-mask probability suffices to control joint KL divergence by the masked discrepancy. The following result shows that it is also necessary for distribution-free control over all admissible models.

Proposition 4.1 (Full-mask characterization)**.** *A schedule μ admits a constant $C_\mu < \infty$ such that*

$$D_{\mathrm{KL}}(p\|q) \;\leq\; C_\mu\, \mathfrak{D}_\mu(p\|q) \qquad \text{for all } p \in \Delta(\mathcal{X}) \text{ and all } q \in \mathcal{Q}(p) \tag{31}$$

if and only if $\pi_0(\mu) > 0$. When $\pi_0(\mu) > 0$, the constant $C_\mu = 1/\pi_0(\mu)$ is valid, and consequently

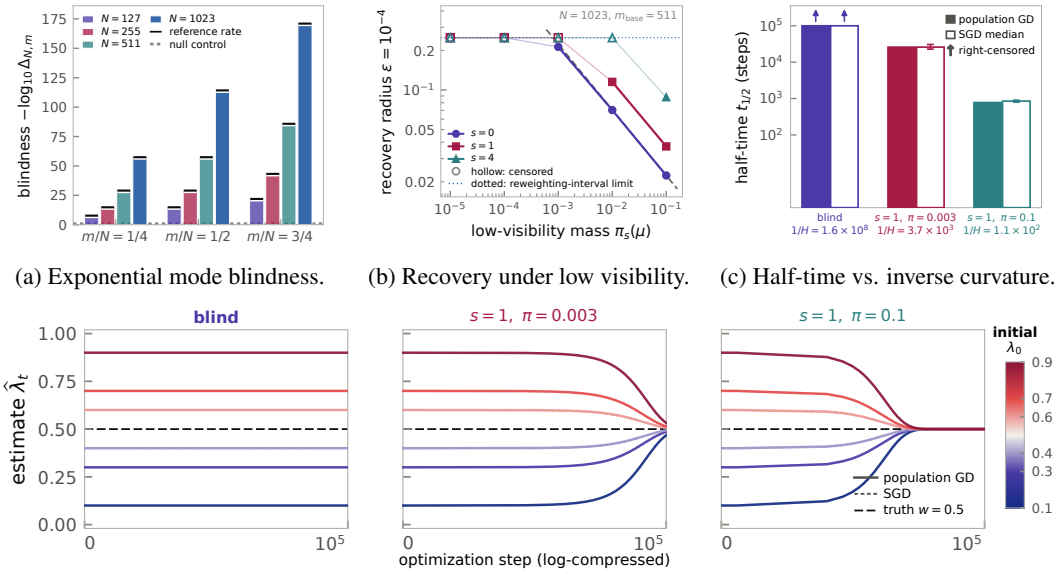
$$\mathfrak{A}^{\mu}_{\mathrm{TV}}(p, \varepsilon) \;\leq\; \sqrt{\varepsilon/(2\,\pi_0(\mu))}. \tag{32}$$

Unlike Cor. 4.3, Prop. 4.1 provides KL control uniformly over both data distributions and admissible models. While particular distributions may admit control without full masks, necessity concerns the distribution-free bound above. This characterization does not extend to per-token log loss, whose full-mask term scores only coordinate marginals rather than the joint distribution (Appx. C.7). See Remark D.2 for further discussion of the sharpness of this guarantee.

5 Experiments

We empirically test two predictions in the conditioned-product mixture of Prop. G.1: visibility controls the mode-weight signal, and residual uncertainty predicts gradient-descent recovery speed. Holding within-mode distributions fixed isolates the effect of the schedule, with all conditionals induced by the shared joint distribution q_λ. Both exact evaluation and optimization are conducted within this model family. Appx. H provides full experimental protocols, robustness checks, as well as interacting and unimodal controls.

Mask visibility controls mode-weight sensitivity. At joint TV error $1/4$, masked discrepancy decreases exponentially with visible-context size $m := |K^{\complement}|$ (Fig. 2a; Thm. 4.1). Across the evaluated grid with $w = 1/2$, discrepancy normalized by the product of mode-weight KL and residual uncertainty remains in $[4.000, 4.408]$ (Appx. H.2), supporting the sensitivity characterization of Thm. 4.2. Assigning mass $\pi_s(\mu)$ to masks leaving $s \in \{0, 1, 4\}$ coordinates visible shrinks the largest admissible mode-weight error at $\varepsilon = 10^{-4}$ (Fig. 2b). Away from interval saturation, the full-mask curve has log–log slope -0.49 against $\pi_0(\mu)$, consistent with $\pi_0(\mu)^{-1/2}$ scaling. These non-full masks restore mode-weight sensitivity within the reweighting family (Cor. 4.3).



(a) Exponential mode blindness. (b) Recovery under low visibility. (c) Half-time vs. inverse curvature.

(d) Mode-weight optimization trajectories across mask schedules.

Figure 2: Population geometry and optimization within the reweighting family q_λ. (a) Mode blindness at fixed joint TV error $1/4$; dotted curves show the unimodal control. (b) Largest admissible mode-weight error vs. low-visibility mass at $\varepsilon = 10^{-4}$; hollow markers are interval-limited. (c) Population and median stochastic half-times compared with the inverse-curvature prediction; the blind case is right-censored. (d) Solid curves show population gradient descent toward $w = 1/2$. Uncensored SGD traces and bands end after sustained half-error crossings.

Residual uncertainty predicts optimization speed. We parameterize $\lambda = (1 + e^{-a})^{-1}$ and optimize conditional mode-label cross-entropy, the exact λ-dependent part of the joint masked-block negative log-likelihood. Writing $R_\mu(a)$ for the population risk, its curvature is $H_\mu := R''_\mu(\mathrm{logit}(w)) = \mathbb{E}_{K\sim\mu}[\mathrm{mmse}_p(Z \mid X_{K^c})]$. At the fixed learning rate $\eta = 0.1$, local gradient descent predicts $t_{1/2} \simeq \log 2/(\eta H_\mu)$ for halving the mode-weight error $|\lambda - w|$. For $N = 63$, $m_{\text{base}} = 31$, and $\lambda_0 = 0.4$, the blind schedule has a predicted local half-time of 1.11×10^9 steps; neither population gradient descent nor any of the five stochastic runs reaches the half-error threshold within 10^5 steps. Assigning mass $\pi_1(\mu) = 0.1$ to masks leaving one coordinate visible reduces the prediction to 770 steps. The observed half-times are 776 steps for population gradient descent and a median of 900 steps for stochastic gradient descent. Population and median stochastic half-times preserve the predicted ordering across uncensored schedules (Fig. 2c), while trajectories show recovery under non-full-mask interventions (Fig. 2d).

6 Conclusion

Mask schedules should be judged not only by the conditional prediction tasks they create, but also by the joint-distribution errors they expose. When visible context nearly determines the active mode, masked prediction can remain near-optimal despite substantial errors in global mode weights. Small masked-prediction excess risk must therefore be interpreted relative to both the schedule and the underlying data geometry. When global frequencies matter, schedule design should retain supervision that constrains them while supporting effective learning within modes: concretely, low-visibility masks that preserve mode uncertainty yield recovery bounds within the mode-reweighting family, while positive full-mask probability characterizes distribution-free joint-KL control under joint masked-block log loss. For representation learning more broadly, these findings motivate asking how strongly an objective constrains the target of interest, and what level of excess risk would support its recovery. Understanding how these population-level constraints interact with finite-sample generalization, optimization, and architectural inductive biases remains an important next step. See Appx. I for broader implications, limitations, and open directions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with providing feedback on interpreting experimental results. We did not use generative AI tools to generate synthetic datasets, and other required-disclosure tasks not relevant to this work are not applicable. Additionally, we used generative AI tools for language editing, drafting and revising portions of the manuscript, literature search and related-work discovery, and suggestions on presentation. All AI-assisted experimental interpretations were checked against our experimental results, literature references were verified against primary sources, and AI-assisted manuscript text was reviewed and revised by the authors for accuracy and consistency. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies population-level properties of masked-prediction objectives and does not involve human participants or the collection of new data. The natural-text experiments use existing publicly available corpora in accordance with their documented licenses and report only aggregate measurements of residual mode uncertainty. We do not release derived text data or attempt to identify individuals. We are not aware of additional material ethical risks arising from this work.

REPRODUCIBILITY STATEMENT

Assumptions and formal statements are provided in the main text, with complete proofs in Appx. D. The appendices provide auxiliary results, the extension to multiple modes, mask-schedule calculations, and further interpretations. Appx. H reports the full experimental protocols, hyperparameters, numerical audits, uncertainty-estimation procedures, and additional results. All sources of randomness are documented for the stochastic experiments.

REFERENCES

Marianne Arriola, Aaron Gokaslan, Justin Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 50726–50753, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7ede97c3e082c6df10a8d6103a2eebd2-Paper-Conference.pdf.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.pdf.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 229–237, 2010. URL https://aclanthology.org/N10-1027/.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974. URL https://doi.org/10.1111/j.2517-6161.1974.tb00999.x.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003. URL https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Yichao Cai, Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi. CLAP: Isolating content from style through contrastive learning with augmented prompts. In *European Conference on Computer Vision*, pp. 130–147. Springer, 2024. URL https://link.springer.com/chapter/10.1007/978-3-031-72664-4_8.

Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3KtPujOw5z.

Yichao Cai, Zhen Zhang, Yuhang Liu, and Javen Qinfeng Shi. The geometric mechanics of contrastive representation learning: Alignment potentials, entropic dispersion, and cross-modal divergence. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=X9VtjdBlQw.

Pietro Caputo and Daniel Parisi. Block factorization of the relative entropy via spatial mixing. *Communications in Mathematical Physics*, 388(2):793–818, 2021. URL https://doi.org/10.1007/s00220-021-04237-1.

CodeParrot. GitHub Code Clean. Hugging Face Hub, https://huggingface.co/datasets/codeparrot/github-code-clean, 2022. Derived from codeparrot/github-code; accessed 2026-07-20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL https://aclanthology.org/N19-1423/.

Jian Ding, Eyal Lubetzky, and Yuval Peres. The mixing time evolution of Glauber dynamics for the mean-field Ising model. *Communications in Mathematical Physics*, 289(2):725–764, 2009. URL https://doi.org/10.1007/s00220-009-0781-9.

Richard S Ellis and Charles M Newman. The statistics of Curie-Weiss models. *Journal of Statistical Physics*, 19(2):149–161, 1978. URL https://doi.org/10.1007/BF01012508.

Reza Gheissari and Alistair Sinclair. Low-temperature Ising dynamics with random initializations. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1445–1458, 2022. URL https://doi.org/10.1145/3519935.3519964.

Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. Exposing the implicit energy networks behind masked language models via Metropolis–Hastings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=6PvWo1kEvlT.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021. URL https://proceedings.neurips.cc/paper/2021/file/27debb435021eb68b3965290b5e24c49-Paper.pdf.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Lucas Torroba Hennigen and Yoon Kim. Deriving language models from masked language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1149–1159, 2023. URL https://aclanthology.org/2023.acl-short.99/.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999. URL https://sigir.org/wp-content/uploads/2017/06/p211.pdf.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Seonwoo Kim and Jungkyoung Lee. Sharp mixing time asymptotics of Glauber dynamics for the Curie-Weiss-Potts model at low temperatures. *arXiv preprint arXiv:2602.19545*, 2026. URL https://arxiv.org/abs/2602.19545.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TD7AnQjNzR6.

Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7918–7928, June 2023. URL https://doi.org/10.1109/CVPR52729.2023.00765.

David A Levin, Malwina J Luczak, and Yuval Peres. Glauber dynamics for the mean-field Ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146 (1):223, 2010. URL https://doi.org/10.1007/s00440-008-0189-z.

Yuchen Li, Alexandre Kirchmeyer, Aashay Mehta, Yilong Qin, Boris Dadachev, Kishore Papineni, Sanjiv Kumar, and Andrej Risteski. Promises and pitfalls of generative masked language modeling: Theoretical framework and practical guidelines. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BTkaKA74mS.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=S7Evzt9uit3.

Bingbin Liu, Daniel J Hsu, Pradeep Ravikumar, and Andrej Risteski. Masked prediction: A parameter identifiability view. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21241–21254. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/85dd09d356ca561169b2c03e43cf305e-Paper-Conference.pdf.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Yuhang Liu, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton van den Hengel, and Javen Qinfeng Shi. I predict therefore I am: Is next token prediction enough to learn human-interpretable concepts from data? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vVYD74U5KE.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

Emanuele Marconato, Sebastien Lachapelle, Sebastian Weichwald, and Luigi Gresele. All or none: Identifiable linear properties of next-token predictors in language modeling. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=XCmIlemQP5.

Yu Meng, Jitin Krishnan, Sinong Wang, Qifan Wang, Yuning Mao, Han Fang, Marjan Ghazvininejad, Jiawei Han, and Luke Zettlemoyer. Representation deficiency in masked language modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=b3l0piOrGU.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VMAHQDOtjp.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Wa3cfE3Iay.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sMyXP8Tanm.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web*, pp. 91–100, 2008. URL https://doi.org/10.1145/1367497.1367510.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. CANDI: Hybrid discrete-continuous diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=03KQGUvw1e.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/volume21/20-074/20-074.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184. Curran Associates, Inc., 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/eb0b13cc515724ab8015bc978fdde0ad-Paper-Conference.pdf.

Ramkrishna Jyoti Samanta, Somabha Mukherjee, and Jiang Zhang. Mixing phases and metastability for the Glauber dynamics on the p-spin Curie-Weiss model. *arXiv preprint arXiv:2412.16952*, 2024. URL https://arxiv.org/abs/2412.16952.

Suvadip Sana, Sami Wolf, Neer Mehta, Alina Shah, Aitzaz Shaikh, Janna Goodman, and Lionel Levine. Mixing times of Glauber dynamics on masked language models. *arXiv preprint arXiv:2605.16378*, 2026. URL https://arxiv.org/abs/2605.16378.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri

and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, pp. 5–42, 2011. URL https://www.jstor.org/stable/24309261.

Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36, 2019. URL https://aclanthology.org/W19-2304.pdf.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k/wang20k.pdf.

Tom Young, Yunan Chen, and Yang You. Inconsistencies in masked language models. *arXiv preprint arXiv:2301.00068*, 2022. URL https://arxiv.org/abs/2301.00068.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148.

Qi Zhang, Runyu Zhou, Yifei Wang, and Yisen Wang. A theoretical framework for masked pretraining (MPT). *Journal of Machine Learning Research*, 27(199):1–46, 2026. URL https://www.jmlr.org/papers/volume27/25-0477/25-0477.pdf.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CTC7CmirNr.

Contents of Appendix

A Notation and Conventions **16**

B Related Work **18**

C Lemmas and Auxiliary Results **19**

C.1 Basic Properties of the Identifiability Modulus 19

C.2 Approximate Tensorization and the Modulus 19

C.3 Shared-Component Mixtures 21

C.4 Pooled Posterior Confidence 22

C.5 Cross-Mode Transition Bound 23

C.6 Reverse Mode Blindness 24

C.7 Factorized Per-Token Losses 26

C.8 A Sufficient Condition for Asm. 4.2 27

D Proofs of Statements **27**

D.1 Proof of Lemma 3.1 27

D.2 Proof of Prop. 3.1 28

D.3 Proof of Thm. 4.1 29

D.4 Proof of Cor. 4.1 30

D.5 Proof of Lemma 4.1 31

D.6 Proof of Cor. 4.2 33

D.7 Proof of Thm. 4.2 33

D.8 Proof of Cor. 4.3 34

D.9 Proof of Prop. 4.1 35

E Mode Blindness with Multiple Modes ($M \geq 2$) **36**

E.1 Mode Structure and Pinning 36

E.2 Mixture Identities and Posterior Tilting 37

E.3 Mode Blindness and the Modulus Lower Bound 39

E.4 Directional Sensitivity Beyond Two Modes 40

F Low-Visibility Mass for Common Mask Schedules **41**

G Interpretation and Instantiation of Assumptions **43**

G.1 Large-Context Mode Pinning 43

G.2 Low-Visibility Uncertainty 43

G.3 An Explicit Distribution Satisfying Both Assumptions 43

H Experimental Protocols and Additional Results **47**

H.1 Controlled Distributions, Estimands, and Computation 47

H.2 Exact Computation of Mode-Blindness Geometry 48

H.3 Low-Visibility Schedule Intervention 50

H.4 Direct Optimization within q_λ 51

H.5 Sampling Cost of Mode-Weight Errors 53

H.6 Exploratory Natural-Text Data Geometry 54

I Discussion **55**

I.1 Implications for Mask Schedule Design 55

I.2 Limitations and Open Directions 56

A NOTATION AND CONVENTIONS

This section introduces the notation and conventions used throughout this work, with Tab. 1 providing a quick reference guide.

General conventions. For an integer $N \geq 1$, we write $[N] := \{1, \ldots, N\}$. Unless stated otherwise, asymptotic statements are taken as $N \to \infty$. For real-valued sequences $f = f_N$ and $g = g_N$, we write $f = O(g)$ if $|f_N| \leq C|g_N|$ for some constant $C > 0$ and all sufficiently large N. We write $f = \Omega(g)$ if $g = O(f)$, and $f = \Theta(g)$ (or equivalently $f \asymp g$) if both relations hold. For sequences with $g_N \neq 0$ for all sufficiently large N, we write $f \sim g$ if $f_N/g_N \to 1$ and $f = o(g)$ if $f_N/g_N \to 0$. We use $\simeq$ for approximate equality. A sequence is called *sublinear* if it is $o(N)$. Calligraphic letters such as $\mathcal{A}$, $\mathcal{X}$, and $\mathcal{Q}(p)$ denote spaces, classes, or distinguished subsets. Uppercase letters generally denote random variables, and lowercase letters denote their realizations. We also use uppercase letters for coordinate sets such as $K, V, S \subseteq [N]$. For a subset $S \subseteq [N]$, we define the projected vectors $X_S := (X_i)_{i \in S}$ and $x_S := (x_i)_{i \in S}$.

Probability distributions. For a finite space $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the simplex of probability distributions on $\mathcal{X}$. Given a distribution $r \in \Delta(\mathcal{X})$ on the configuration space $\mathcal{X} = \mathcal{A}^N$ and a subset $S \subseteq [N]$, r_S denotes the marginal distribution of X_S under r. For complementary sets $V \subseteq [N]$ and $K = [N] \setminus V$, the conditional distribution of X_K given $X_V = x_V$ is written as $r_K(\cdot \mid x_V)$ whenever $r_V(x_V) > 0$.

We write $r(A)$ and $r(A \mid X_V = x_V)$ for probabilities under a specified distribution r. We also use $\mathbb{P}$ when the underlying distribution or random experiment is specified by context or a subscript. Expectations follow the same convention. Subscripts such as $\mathbb{E}_r$, $\mathbb{E}_{X \sim p}$, and $\mathbb{E}_{K \sim \mu}$ indicate the source of randomness when needed.

Information-theoretic quantities. All logarithms are natural. For probability distributions $p, q \in \Delta(\mathcal{X})$, the total-variation distance is

$$D_{\mathrm{TV}}(p, q) := \sup_{A \subseteq \mathcal{X}} |p(A) - q(A)| = \tfrac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)|. \tag{33}$$

The KL divergence is

$$D_{\mathrm{KL}}(p \| q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}, \tag{34}$$

with the conventions $0 \log \frac{0}{b} := 0$ for $b \geq 0$ and $a \log \frac{a}{0} := +\infty$ for $a > 0$.

For $a \in (0, 1)$, define the log-odds

$$\mathrm{logit}(a) := \log \frac{a}{1 - a}, \tag{35}$$

with $\mathrm{logit}(0) := -\infty$ and $\mathrm{logit}(1) := +\infty$.

For $a, b \in [0, 1]$, define the binary KL divergence

$$\mathrm{kl}(a \| b) := a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b}, \tag{36}$$

using the same extended-value conventions. In particular, $\mathrm{kl}(a \| b) \geq 0$, with equality if and only if $a = b$.

Minimum mean-square error. For jointly distributed random variables Y and X under a probability distribution r, with $Y \in \mathbb{R}^d$ and $\mathbb{E}_r \|Y\|_2^2 < \infty$, define the minimum mean-square error

$$\mathrm{mmse}_r(Y \mid X) := \mathbb{E}_r \left[\|Y - \mathbb{E}_r[Y \mid X]\|_2^2 \right]. \tag{37}$$

For scalar Y, this reduces to $\mathbb{E}_r[(Y - \mathbb{E}_r[Y \mid X])^2]$.

Table 1: Canonical notation used throughout the paper.

Symbol	Meaning
Spaces and probability distributions	
$[N]$	Coordinate index set $\{1, \dots, N\}$.
$\mathcal{A}$	Finite alphabet.
$\mathcal{X} = \mathcal{A}^N,\ \mathcal{F} = 2^{\mathcal{X}}$	Configuration space and its full σ-field.
$\Delta(S)$	Probability simplex over a finite set S.
p	Data distribution on $\mathcal{X}$; unless stated otherwise, $X \sim p$.
q, r	Generic model distribution and generic probability distribution on $\mathcal{X}$.
$\mathrm{supp}(r)$	Support of a distribution r.
$\mathcal{Q}(p)$	Support-covering model class $\{q \in \Delta(\mathcal{X}) : \mathrm{supp}(p) \subseteq \mathrm{supp}(q)\}$.
Coordinates, marginals, and conditionals	
X, x	Random configuration and a realization.
$S \subseteq [N]$	Generic coordinate set.
$K,\ K^{\complement}$	Masked coordinate set and visible set $K^{\complement} := [N] \setminus K$.
V	Generic visible or conditioning coordinate set.
X_S, x_S	Subvector indexed by S and its realization.
r_S	Marginal distribution of X_S under r.
$r_K(\cdot \mid x_V)$	Conditional distribution of X_K given $X_V = x_V$.
Mask schedules and block updates	
$\mu,\ \mathrm{supp}(\mu)$	Mask schedule on $2^{[N]}$ and its support.
$v_{\min}(\mu)$	Minimum visible-set size $\min_{K \in \mathrm{supp}(\mu)} \lvert K^{\complement}\rvert$.
$\pi_s(\mu)$	Low-visibility mass $\mu(\{K : \lvert K^{\complement}\rvert \le s\})$.
Masked objective and identifiability	
$\mathfrak{D}_K(p\Vert q)$	Fixed-mask log discrepancy (Defn. 2.1).
$\mathfrak{D}_\mu(p\Vert q)$	Schedule-averaged masked log discrepancy $\mathbb{E}_{K\sim\mu}\mathfrak{D}_K(p\Vert q)$.
$\mathfrak{A}_d^\mu(p, \varepsilon)$	ε-identifiability modulus under distance d (Defn. 2.2).
$\mathfrak{A}_{\mathrm{TV}}^\mu(p, \varepsilon)$	Total-variation identifiability modulus.
Mode structure and constructions	
$\mathcal{X}_+, \mathcal{X}_-$	Mode regions forming the partition $\mathcal{X} = \mathcal{X}_+ \sqcup \mathcal{X}_-$.
$\tau, -\tau$	Mode label in $\{+, -\}$ and the opposite label.
Z	Binary mode indicator $Z := \mathbb{1}\{X \in \mathcal{X}_+\}$.
p^τ	Mode-conditioned distribution $p(\cdot \mid \mathcal{X}_\tau)$.
p_V^τ	Marginal of p^τ on coordinates V.
$\mathcal{E}_V^\tau$	Typical visible contexts from mode region $\mathcal{X}_\tau$.
$\beta_V(x_V)$	Posterior mode probability $p(Z = 1 \mid X_V = x_V)$.
$\mathrm{mmse}_p(Z \mid X_V)$	Residual mode MMSE $\mathbb{E}_p[(Z - \mathbb{E}_p[Z \mid X_V])^2] = \mathbb{E}[\beta_V(1 - \beta_V)]$.
q_λ	Reweighted bimodal distribution $\lambda\, p(\cdot \mid \mathcal{X}_+) + (1 - \lambda)\, p(\cdot \mid \mathcal{X}_-)$.
c_0	Uniform lower bound on the probability mass of each mode region.
c_1	Exponential rate controlling atypical visible contexts.
κ	Exponential rate controlling the opposite-mode posterior on $\mathcal{E}_V^\tau$.
s_{pin}	Visibility threshold in the mode-pinning assumption (Asm. 4.1).
s_{ov}	Visibility scale in the low-visibility uncertainty assumption (Asm. 4.2).
c	Combined rate $c := \min\{c_1, \kappa\}$.
Divergences and constants	
$D_{\mathrm{KL}}(\cdot\Vert\cdot)$	KL divergence.
$\mathrm{kl}(\cdot\Vert\cdot)$	Binary KL divergence.
$D_{\mathrm{TV}}(\cdot, \cdot)$	Total-variation distance.
$\mathrm{logit}(\cdot)$	Log-odds.
$\mathrm{diam}_d(\Delta(\mathcal{X}))$	Diameter of the probability simplex under d.
$\bar{C}_{\mathrm{AT}}(r, \mu)$	Block approximate-tensorization constant of a distribution r (Eq. (8)).
$\bar{C}_{\mathrm{AT}}^{\varepsilon}(p, \mu)$	Budget-uniform AT constant over $\{q \in \mathcal{Q}(p) : \mathfrak{D}_\mu(p\Vert q) \le \varepsilon\}$ (Eq. (9)).

B Related Work

Masked-pretraining objectives and theory. Estimating a joint distribution through conditional likelihoods is the basis of pseudolikelihood and conditional composite likelihood (Besag, 1974; Varin et al., 2011). Prediction from partial observations also underlies masked language modeling (Devlin et al., 2019; Liu et al., 2019), masked autoencoding (He et al., 2022), masked latent prediction (Assran et al., 2023; Bardes et al., 2024; Nam et al., 2026), and absorbing-state discrete diffusion (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024; Ou et al., 2025). For masked diffusion, ELBO formulations can be expressed as weighted masked-prediction losses, including discrete mixtures over mask counts (Shi et al., 2024; Zheng et al., 2025).

Complementary theory studies how masking shapes learned representations and latent recovery. Kong et al. (2023) analyze hierarchical latent-variable identifiability in masked autoencoders, characterizing how masking ratio and patch size govern which latent variables are recovered. Meng et al. (2024) identify representation deficiencies caused by dimensions devoted to `[MASK]` tokens, while Zhang et al. (2026) connect masked reconstruction to implicit alignment, dimensional collapse, and downstream guarantees across various masking strategies. Pynadath et al. (2026) study a distinct notion of token identifiability, examining how Gaussian corruption degrades discrete token identity. While these works characterize representation geometry or corruption mechanics, our focus is fundamentally on how the mask schedule controls the recovery of the underlying joint distribution from near-optimal conditional prediction.

Joint-distribution recovery. While Liu et al. (2022) establish exact identification for latent-variable models, analyzing distributional recovery through functional inequalities has recently gained traction: Koehler et al. (2023) connect score matching and pseudolikelihood to functional inequalities, including mixture-weight failures in score matching, and Li et al. (2024) establish finite-sample joint recovery for masked prediction via block AT (Caputo & Parisi, 2021). Complementing these guarantees, we characterize the population excess-risk geometry of mode reweightings over a fixed interval: masked discrepancy is bounded above and below by the mode-weight KL divergence scaled by schedule-averaged residual mode uncertainty. This yields the exponential mode-blindness scale and corresponding blowup of AT constants under mode pinning, as well as explicit low-visibility recovery guarantees within the reweighting family.

Metastability and conditional compatibility. Our formulation draws intuition from phase coexistence and slow mixing in Curie–Weiss models (Ellis & Newman, 1978; Levin et al., 2010; Ding et al., 2009; Gheissari & Sinclair, 2022; Kim & Lee, 2026; Samanta et al., 2024), as well as mixture descriptions of text (Hofmann, 1999; Blei et al., 2003). A separate line of work shows that conditionals learned by masked models need not be compatible with a single joint distribution (Wang & Cho, 2019; Young et al., 2022; Goyal et al., 2022; Hennigen & Kim, 2023). Recent work also studies iterative masked-token sampling, proving exponentially slow escape from semantic basins under a local margin condition at low temperatures (Sana et al., 2026). Mode blindness is orthogonal to conditional incompatibility: it occurs even when all model conditionals are induced by a single coherent joint distribution.

Population objectives and identifiability. Theoretical analyses of contrastive learning relate its objectives to latent-class structure (Saunshi et al., 2019), spectral properties (HaoChen et al., 2021), alignment and uniformity (Wang & Isola, 2020), dimensional collapse (Jing et al., 2022), distributional geometry (Cai et al., 2026), and cross-modal geometry and feature recovery (Cai et al., 2024; Liang et al., 2022; Cai et al., 2025). For next-token prediction, related work characterizes representations of distributionally equivalent predictors (Marconato et al., 2025) and conditions for recovering human-interpretable latent structure (Liu et al., 2026). Representation-level near-identifiability further bounds, up to transformations, discrepancies between representations at exact population-risk minimizers (Nelson et al., 2026). In contrast to representation-space invariants, our ε-identifiability modulus measures the largest joint-distribution error among model distributions with excess risk at most ε. It thus characterizes the stability of distributional recovery from the population objective without positing internal representations or transformation symmetries.

C LEMMAS AND AUXILIARY RESULTS

This section gives lemmas and auxiliary results that support the main proofs or clarify theoretical interpretations of the main results.

C.1 BASIC PROPERTIES OF THE IDENTIFIABILITY MODULUS

This section records basic well-posedness and monotonicity properties of the ε-identifiability modulus (Defn. 2.2).

Proposition C.1 (Basic properties of the modulus)**.** *For every $p \in \Delta(\mathcal{X})$, mask schedule μ, continuous distance d, and discrepancy budget $\varepsilon \geq 0$,*

$$0 \leq \mathfrak{A}_d^\mu(p, \varepsilon) \leq \operatorname{diam}_d(\Delta(\mathcal{X})) < \infty, \qquad \textit{where} \quad \operatorname{diam}_d(\Delta(\mathcal{X})) := \sup_{q,q' \in \Delta(\mathcal{X})} d(q, q').$$

Moreover, $\mathfrak{A}_d^\mu(p, \varepsilon)$ is nondecreasing in ε.

Proof. Write $\mathcal{Q}_\varepsilon^\mu(p) := \{q \in \mathcal{Q}(p) : \mathfrak{D}_\mu(p\|q) \leq \varepsilon\}$ for the admissible set in Eq. (6). We first show it is nonempty. The data distribution itself belongs to $\mathcal{Q}(p)$. Moreover, for every mask K, the fixed-mask discrepancy

$$\mathfrak{D}_K(p\|p) = \mathbb{E}_{X_{K^\complement} \sim p_{K^\complement}}[D_{\mathrm{KL}}(p_K(\cdot \mid X_{K^\complement}) \| p_K(\cdot \mid X_{K^\complement}))] = 0,$$

since each conditional KL divergence between identical distributions vanishes. Averaging over $K \sim \mu$ gives

$$\mathfrak{D}_\mu(p\|p) = \mathbb{E}_{K \sim \mu} \mathfrak{D}_K(p\|p) = 0 \leq \varepsilon,$$

so $p \in \mathcal{Q}_\varepsilon^\mu(p)$. Hence the supremum in Eq. (6) is over a nonempty set, and since $d(p, p) = 0$,

$$\mathfrak{A}_d^\mu(p, \varepsilon) \geq 0.$$

Next, for every $q \in \mathcal{Q}_\varepsilon^\mu(p)$,

$$d(p, q) \leq \operatorname{diam}_d(\Delta(\mathcal{X})).$$

Because $\mathcal{A}$ is a finite alphabet, $\mathcal{X} = \mathcal{A}^N$ is finite and the simplex $\Delta(\mathcal{X})$ is compact. Since d is continuous on $\Delta(\mathcal{X}) \times \Delta(\mathcal{X})$, the diameter $\operatorname{diam}_d(\Delta(\mathcal{X}))$ is finite. Taking the supremum over $\mathcal{Q}_\varepsilon^\mu(p)$ therefore gives

$$0 \leq \mathfrak{A}_d^\mu(p, \varepsilon) \leq \operatorname{diam}_d(\Delta(\mathcal{X})) < \infty.$$

Finally, let $0 \leq \varepsilon \leq \varepsilon'$. Then $\mathcal{Q}_\varepsilon^\mu(p) \subseteq \mathcal{Q}_{\varepsilon'}^\mu(p)$, because every model distribution with discrepancy at most ε also has discrepancy at most ε'. Taking the supremum of $d(p, q)$ over these nested sets yields

$$\mathfrak{A}_d^\mu(p, \varepsilon) \leq \mathfrak{A}_d^\mu(p, \varepsilon').$$

Thus the modulus is nondecreasing in ε. □

C.2 APPROXIMATE TENSORIZATION AND THE MODULUS

This section relates the entropy-functional form of block AT (Caputo & Parisi, 2021) to its formulation in terms of conditional KL divergences (Li et al., 2024), and derives the modulus bound of Eq. (9).

Throughout, we fix a mask schedule μ. For a distribution $r \in \Delta(\mathcal{X})$ and a function $\phi \colon \mathcal{X} \to [0, \infty)$, the entropy functional is

$$\operatorname{Ent}_r(\phi) := \mathbb{E}_r[\phi \log \phi] - \mathbb{E}_r[\phi] \log \mathbb{E}_r[\phi], \tag{38}$$

with the convention $0 \log 0 := 0$. If $\mathbb{E}_r[\phi] = 0$, nonnegativity implies $\phi = 0$ r-almost surely and hence $\operatorname{Ent}_r(\phi) = 0$. Jensen's inequality applied to the convex function $t \mapsto t \log t$ gives $\operatorname{Ent}_r(\phi) \geq 0$. Because every expectation below integrates ϕ against r or its conditionals, the values of ϕ outside $\operatorname{supp}(r)$ have no effect on the integrals. We therefore assume without loss of generality that ϕ vanishes outside this support.

The subtracted term in Eq. (38) extends the relative entropy to unnormalized arguments: When $\mathbb{E}_r[\phi] > 0$, normalizing ϕ defines the distribution $p_\phi := (\phi/\mathbb{E}_r[\phi])\, r$. Expanding its relative entropy, $D_{\mathrm{KL}}(p_\phi \| r) = \mathbb{E}_r[(\phi/\mathbb{E}_r[\phi]) \log(\phi/\mathbb{E}_r[\phi])]$, gives

$$\mathrm{Ent}_r(\phi) = \mathbb{E}_r[\phi]\, D_{\mathrm{KL}}\big(p_\phi \,\big\|\, r\big). \tag{39}$$

Thus, Ent_r is the homogeneous extension of relative entropy from normalized densities to nonnegative functions. In particular, for a distribution $p' \in \Delta(\mathcal{X})$ with $\mathrm{supp}(p') \subseteq \mathrm{supp}(r)$, its density with respect to r,

$$\phi_{p'} := \frac{\mathrm{d}p'}{\mathrm{d}r}, \qquad \phi_{p'}(x) = \begin{cases} p'(x)/r(x), & x \in \mathrm{supp}(r), \\ 0, & \text{otherwise}, \end{cases} \tag{40}$$

satisfies $\mathbb{E}_r[\phi_{p'}] = p'(\mathrm{supp}(r)) = 1$, and thus recovers the relative entropy $\mathrm{Ent}_r(\phi_{p'}) = D_{\mathrm{KL}}(p' \,\|\, r)$.

For a mask K and a context $x_{K^\complement}$ with $r_{K^\complement}(x_{K^\complement}) > 0$, we write $\mathrm{Ent}_{r_K(\cdot|x_{K^\complement})}(\phi)$ for the entropy of the fiber restriction $x_K \mapsto \phi(x_{K^\complement}, x_K)$ under the conditional distribution. Finally, for a distribution p' with $\mathrm{supp}(p') \subseteq \mathrm{supp}(r)$ we have $r \in \mathcal{Q}(p')$ by Eq. (2), so $\mathfrak{D}_\mu(p' \,\|\, r)$ is well defined by Defn. 2.1.

Lemma C.1 (Equivalent forms of block AT and a modulus bound)**.** *Fix a mask schedule μ.*

(i) For every distribution $r \in \Delta(\mathcal{X})$ and every constant $\bar{C} \in [0, \infty)$, the entropy-functional block-factorization inequality

$$\mathrm{Ent}_r(\phi) \;\le\; \bar{C}\, \mathbb{E}_{K\sim\mu}\, \mathbb{E}_{X_{K^\complement}\sim r_{K^\complement}} \Big[\mathrm{Ent}_{r_K(\cdot|X_{K^\complement})}(\phi) \Big] \quad \textit{for every } \phi\colon \mathcal{X} \to [0,\infty) \tag{41}$$

holds if and only if the density-ratio inequality of Eq. (8) holds at the same constant:

$$D_{\mathrm{KL}}(p' \,\|\, r) \;\le\; \bar{C}\, \mathfrak{D}_\mu(p' \,\|\, r) \qquad \textit{for every } p' \in \Delta(\mathcal{X}) \textit{ with } \mathrm{supp}(p') \subseteq \mathrm{supp}(r).$$

Thus, the optimal constants of the two inequalities coincide with $\bar{C}_{\mathrm{AT}}(r, \mu)$ in Eq. (8).

(ii) For every $p \in \Delta(\mathcal{X})$ and every $\varepsilon \ge 0$, writing $\mathcal{Q}^\mu_\varepsilon(p) := \{q \in \mathcal{Q}(p) : \mathfrak{D}_\mu(p\|q) \le \varepsilon\}$ for the ε-discrepancy ball, the modulus bound of Eq. (9) holds:

$$\mathfrak{A}^\mu_{\mathrm{TV}}(p, \varepsilon) \;\le\; \sqrt{\tfrac{1}{2}\, \bar{C}^\varepsilon_{\mathrm{AT}}(p, \mu)\, \varepsilon}, \qquad \textit{with } \bar{C}^\varepsilon_{\mathrm{AT}}(p, \mu) = \sup_{q \in \mathcal{Q}^\mu_\varepsilon(p)} \bar{C}_{\mathrm{AT}}(q, \mu),$$

where the right-hand side is understood as $+\infty$ when $\bar{C}^\varepsilon_{\mathrm{AT}}(p, \mu) = \infty$.

Proof intuition. For a normalized density $\phi = p'/r$, the global entropy $\mathrm{Ent}_r(\phi)$ equals $D_{\mathrm{KL}}(p'\|r)$. The key point is that conditional entropy supplies the context weights required by the masked discrepancy: on a fiber with visible context x_V, the conditional mean of ϕ is $p'_V(x_V)/r_V(x_V)$. The fiber entropy is therefore the conditional KL divergence multiplied by this ratio. Integrating against r_V converts the weights to p'_V, recovering the fixed-mask discrepancy. Homogeneity extends the identity to unnormalized functions. Applying block AT with reference distribution q, followed by Pinsker's inequality, yields the modulus bound.

Proof. Part (i). Step 1: normalization. Both sides of Eq. (41) are homogeneous of degree one in ϕ: for $\lambda > 0$, expanding $(\lambda t)\log(\lambda t) = \lambda\, t \log t + \lambda t \log \lambda$ shows that the $\log \lambda$ contributions cancel between the two terms of Eq. (38), so $\mathrm{Ent}_r(\lambda\phi) = \lambda\, \mathrm{Ent}_r(\phi)$, and the same computation applies to every conditional entropy on the right-hand side. Moreover, if $\mathbb{E}_r[\phi] = 0$, then ϕ vanishes on $\mathrm{supp}(r)$, both sides are 0 by convention, and the inequality holds trivially.

Together, these observations reduce Eq. (41) to functions with $\mathbb{E}_r[\phi] = 1$. Such functions are the densities $\phi_{p'}$ of Eq. (40): given such a ϕ, the distribution $p' := \phi\, r$ satisfies $\mathrm{supp}(p') \subseteq \mathrm{supp}(r)$ and $\phi = \phi_{p'}$, and the correspondence $\phi_{p'} \leftrightarrow p'$ is a bijection.

Step 2: the global and fiber identities. Eq. (39) applies with $\mathbb{E}_r[\phi_{p'}] = 1$ and $p_{\phi_{p'}} = p'$, giving $\mathrm{Ent}_r(\phi_{p'}) = D_{\mathrm{KL}}(p' \,\|\, r)$, finite because $\mathcal{X}$ is finite and $\mathrm{supp}(p') \subseteq \mathrm{supp}(r)$.

Fix a mask K and a context $x_{K^\complement} \in \mathrm{supp}(r_{K^\complement})$, and set the ratio $g := p'_{K^\complement}(x_{K^\complement})/r_{K^\complement}(x_{K^\complement})$. If $p'_{K^\complement}(x_{K^\complement}) = 0$, then $\phi_{p'}$ vanishes on this fiber, and the conditional entropy is zero. Otherwise, for every completion x_K with $r_K(x_K \mid x_{K^\complement}) > 0$, the density factorizes as

$$\phi_{p'}(x_{K^\complement}, x_K) = g \cdot \frac{p'_K(x_K \mid x_{K^\complement})}{r_K(x_K \mid x_{K^\complement})},$$

where both sides vanish when $p'_K(x_K \mid x_{K^\complement}) = 0$. The fiber restriction of $\phi_{p'}$ thus has conditional mean g, and its normalization $\phi_{p'}/g$ is the density of $p'_K(\cdot \mid x_{K^\complement})$ with respect to $r_K(\cdot \mid x_{K^\complement})$. Applying Eq. (39) under the conditional distribution gives

$$\mathrm{Ent}_{r_K(\cdot|x_{K^\complement})}(\phi_{p'}) = g\, D_{\mathrm{KL}}\big(p'_K(\cdot \mid x_{K^\complement}) \,\big\|\, r_K(\cdot \mid x_{K^\complement})\big),$$

which is finite by the fiberwise domination inherited from $\mathrm{supp}(p') \subseteq \mathrm{supp}(r)$.

Step 3: averaging. Averaging the fiber identity over $X_{K^\complement} \sim r_{K^\complement}$, the ratio g converts the averaging measure into $p'_{K^\complement}$:

$$\mathbb{E}_{X_{K^\complement}\sim r_{K^\complement}}\Big[\mathrm{Ent}_{r_K(\cdot|X_{K^\complement})}(\phi_{p'})\Big] = \mathbb{E}_{X_{K^\complement}\sim p'_{K^\complement}}\Big[D_{\mathrm{KL}}\big(p'_K(\cdot \mid X_{K^\complement}) \,\big\|\, r_K(\cdot \mid X_{K^\complement})\big)\Big] = \mathfrak{D}_K(p' \,\|\, r),$$

by Defn. 2.1, with contexts outside $\mathrm{supp}(p'_{K^\complement})$ contributing zero to both sides. Averaging over $K \sim \mu$ yields $\mathfrak{D}_\mu(p' \,\|\, r)$.

By Steps 2–3, under the bijection of Step 1 every instance of Eq. (41) has the same left-hand side and the same right-hand side as the corresponding instance of Eq. (8). The two families of inequalities are therefore valid for the same constants $\bar{C}$, and their optimal constants, the infima of the respective sets of valid constants with $\inf \emptyset := \infty$, coincide.

Part (ii). If $\bar{C}^{\varepsilon}_{\mathrm{AT}}(p, \mu) = \infty$, the bound holds vacuously with its right-hand side read as $+\infty$; assume therefore that it is finite. Let $q \in \mathcal{Q}^{\mu}_{\varepsilon}(p)$. Because $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$, Eq. (8) applies with $r = q$ and $p' = p$, yielding

$$D_{\mathrm{KL}}(p \,\|\, q) \;\le\; \bar{C}_{\mathrm{AT}}(q, \mu)\, \mathfrak{D}_\mu(p \,\|\, q) \;\le\; \bar{C}_{\mathrm{AT}}(q, \mu)\, \varepsilon \;\le\; \bar{C}^{\varepsilon}_{\mathrm{AT}}(p, \mu)\, \varepsilon.$$

Pinsker's inequality then gives $D_{\mathrm{TV}}(p, q) \le \sqrt{\frac{1}{2}\, \bar{C}^{\varepsilon}_{\mathrm{AT}}(p, \mu)\, \varepsilon}$. This bound is independent of q, so taking the supremum over $q \in \mathcal{Q}^{\mu}_{\varepsilon}(p)$, which is the supremum defining $\mathfrak{A}^{\mu}_{\mathrm{TV}}(p, \varepsilon)$ in Defn. 2.2, completes the proof. □

Remark C.1 (Relation to the literature). Caputo & Parisi (2021) establish entropy-functional block factorization under strong spatial-mixing conditions. Their normalization includes the coordinate-coverage factor $\gamma(\mu) := \min_{i\in[N]} \sum_{K \ni i} \mu(K)$. With block weights $\mu(K)$, their constant C yields $\bar{C}_{\mathrm{AT}}(r, \mu) \le C/\gamma(\mu)$ whenever $\gamma(\mu) > 0$. Li et al. (2024) use the conditional-KL formulation of block AT in their finite-sample recovery analysis, with the constant evaluated at the learned distribution. Their framework allows configuration-dependent masking and conditions on $(X_{K^\complement}, K)$. For masks drawn independently of X, conditioning additionally on K does not change the conditional distribution of the masked coordinates, so their conditional discrepancy reduces to $\mathfrak{D}_\mu(p' \| r)$. Lemma C.1(ii) applies the same AT-to-TV argument uniformly over all distributions within the discrepancy budget. In the continuous setting, Koehler et al. (2023) use the log-Sobolev constant of the fitted distribution to convert excess score-matching risk into a bound on joint KL divergence.

C.3 Shared-Component Mixtures

This section provides two elementary identities for mixtures that share the same disjoint components and differ only in their mixing weights.

Lemma C.2 (Shared-component mixtures)**.** *Let $\mathcal{Z}$ be a finite set, and let $\nu_0, \nu_1 \in \Delta(\mathcal{Z})$ have disjoint supports. For $a \in [0, 1]$, define $m_a := a\nu_1 + (1 - a)\nu_0$. Then, for every $a, b \in [0, 1]$,*

$$D_{\mathrm{TV}}(m_a, m_b) = |a - b| \tag{42}$$

and

$$D_{\mathrm{KL}}(m_a \| m_b) = \mathrm{kl}(a \| b), \tag{43}$$

where both sides of Eq. (43) are interpreted using the usual extended-value conventions.

Proof intuition. Disjoint supports make the component label directly observable from the sample. Since both mixtures share identical component distributions, this label captures all distinguishing information. The total variation distance and KL divergence between the two mixtures therefore reduce to those between Bernoulli distributions with parameters a and b.

Proof. Let $S_j := \mathrm{supp}(\nu_j)$ for $j \in \{0,1\}$. Since $S_0 \cap S_1 = \varnothing$,

$$m_a - m_b = (a-b)\nu_1 - (a-b)\nu_0.$$

The two signed terms have disjoint supports, so

$$\|m_a - m_b\|_1 = |a-b|\big(\|\nu_1\|_1 + \|\nu_0\|_1\big) = 2|a-b|,$$

which proves Eq. (42) by substituting it into the definition of the total-variation distance (Eq. (33)).

For the relative entropy, first suppose $a, b \in (0,1)$. The complement of $S_0 \cup S_1$ contributes zero. Splitting the sum over the two supports gives

$$\begin{aligned} D_{\mathrm{KL}}(m_a \| m_b) &= \sum\nolimits_{z \in S_1} a\nu_1(z) \log \frac{a\nu_1(z)}{b\nu_1(z)} + \sum\nolimits_{z \in S_0} (1-a)\nu_0(z) \log \frac{(1-a)\nu_0(z)}{(1-b)\nu_0(z)} \\ &= a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b} \\ &= \mathrm{kl}(a \| b). \end{aligned} \tag{44}$$

At the boundary points $a, b \in [0,1]$, the same identity follows directly from the extended-value conventions. □

C.4 Pooled Posterior Confidence

This lemma converts the mode-wise guarantees in Asm. 4.1 into a single high-probability statement under the pooled visible marginal. It shows that a large visible context is rarely ambiguous about the global mode.

Lemma C.3 (Pooled posterior confidence)**.** *Suppose Asm. 4.1 holds, and let c_1 and κ be as there. Recall $\beta_V(x_V) = p(Z = 1 \mid X_V = x_V)$ from Eq. (10). For every $V \subseteq [N]$ with $|V| \geq s_{\mathrm{pin}}$, define*

$$G_V := \Big\{ x_V \in \mathrm{supp}(p_V) : \min\{\beta_V(x_V), 1 - \beta_V(x_V)\} \leq e^{-\kappa|V|} \Big\}. \tag{45}$$

Then

$$p_V(G_V^{\complement}) \leq e^{-c_1|V|},$$

where the complement is taken relative to $\mathrm{supp}(p_V)$*, i.e.,* $G_V^{\complement} := \mathrm{supp}(p_V) \setminus G_V$.

Proof intuition. Under either mode, posterior ambiguity can occur only outside that mode's typical context set. Each mode assigns probability at most $e^{-c_1|V|}$ to this exceptional event. The overall mixture probability is a weighted average of these two probabilities, and thus inherits the same bound.

Proof. Let $w_\tau := p(\mathcal{X}_\tau)$ for $\tau \in \{+,-\}$. Since $\mathcal{X} = \mathcal{X}_+ \sqcup \mathcal{X}_-$, the visible marginal decomposes as

$$p_V = w_+ p_V^+ + w_- p_V^-. \tag{46}$$

For $x_V \in \mathcal{E}_V^+$, Asm. 4.1 gives

$$1 - \beta_V(x_V) = p(\mathcal{X}_- \mid X_V = x_V) \leq e^{-\kappa|V|},$$

and hence $x_V \in G_V$. Thus,

$$\mathcal{E}_V^+ \subseteq G_V.$$

Similarly, for $x_V \in \mathcal{E}_V^-$,

$$\beta_V(x_V) = p(\mathcal{X}_+ \mid X_V = x_V) \leq e^{-\kappa|V|},$$

so

$$\mathcal{E}_V^- \subseteq G_V.$$

Therefore, for each $\tau \in \{+,-\}$,

$$p_V^\tau(G_V^{\complement}) \le p_V^\tau\big((\mathcal{E}_V^\tau)^{\complement}\big) \le e^{-c_1|V|},$$

where the final inequality is exactly the high-probability condition in Asm. 4.1. Averaging these two bounds using Eq. (46) yields

$$p_V(G_V^{\complement}) \;=\; w_+ p_V^+(G_V^{\complement}) + w_- p_V^-(G_V^{\complement}) \;\le\; (w_+ + w_-)e^{-c_1|V|} \;=\; e^{-c_1|V|}.$$

□

C.5 Cross-Mode Transition Bound

This section shows that mode pinning suppresses one-step transitions between the two mode regions when a block update retains a large visible context.

Lemma C.4 (One-step cross-mode transition bound)**.** *Suppose Asm. 4.1 holds, and recall the combined rate $c = \min\{c_1, \kappa\}$. Fix a mode $\tau \in \{+,-\}$ and let $X \sim p^\tau$. Unless indicated otherwise, $\mathbb{P}$ includes randomness from the initial configuration, the heat-bath resampling, and the mask draw when the mask is random.*

(i) *Fix a mask $K \subseteq [N]$ with $|K^{\complement}| \ge s_{\mathrm{pin}}$. Let X' be a single fixed-mask heat-bath update of X: set $X'_{K^{\complement}} = X_{K^{\complement}}$ and draw $X'_K \sim p_K(\cdot \mid X_{K^{\complement}})$. Then*

$$\mathbb{P}(X' \in \mathcal{X}_{-\tau}) \le 2e^{-c|K^{\complement}|}. \tag{47}$$

(ii) *Let $\widetilde{X}$ be a single step of the block dynamics induced by (p,μ): draw $K \sim \mu$ independently of X and set $\widetilde{X}_{K^{\complement}} = X_{K^{\complement}}$, $\widetilde{X}_K \sim p_K(\cdot \mid X_{K^{\complement}})$. Then*

$$\mathbb{P}\big(\widetilde{X} \in \mathcal{X}_{-\tau}\big) \le \pi_s(\mu) + 2e^{-cs}, \qquad \text{for every } s \in \{s_{\mathrm{pin}}, \dots, N\}. \tag{48}$$

(iii) *In particular, if $v_{\min}(\mu) \ge s_{\mathrm{pin}}$, then*

$$\mathbb{P}\big(\widetilde{X} \in \mathcal{X}_{-\tau}\big) \le 2e^{-c\, v_{\min}(\mu)}. \tag{49}$$

Proof intuition. Given the retained context on $V = K^{\complement}$, a heat-bath update crosses modes with probability equal to the opposite-mode posterior. Under p^τ, this posterior is at most $e^{-\kappa|V|}$ on typical contexts, while atypical contexts have probability at most $e^{-c_1|V|}$. Summing these two contributions yields the fixed-mask bound. Under a random schedule, masks retaining at most s coordinates contribute at most $\pi_s(\mu)$, while the remaining masks contribute at most $2e^{-cs}$.

Proof. Step 1: reduction to the opposite-mode posterior. Fix $\tau \in \{+,-\}$ and a mask $K \subseteq [N]$ with $|K^{\complement}| \ge s_{\mathrm{pin}}$. Let $X \sim p^\tau$, and let X' be the fixed-mask update of part (i): $X'_{K^{\complement}} = X_{K^{\complement}}$ and $X'_K \sim p_K(\cdot \mid X_{K^{\complement}})$. Since the update keeps the visible coordinates fixed and redraws the masked block from $p_K(\cdot \mid X_{K^{\complement}})$, the updated configuration has conditional distribution

$$X' \mid X_{K^{\complement}} \sim p(\cdot \mid X_{K^{\complement}}),$$

and therefore

$$\mathbb{P}(X' \in \mathcal{X}_{-\tau} \,|\, X_{K^{\complement}}) = p(\mathcal{X}_{-\tau} \mid X_{K^{\complement}}). \tag{50}$$

Since $X \sim p^\tau$, its visible marginal satisfies $X_{K^{\complement}} \sim p^\tau_{K^{\complement}}$. Taking the expectation in Eq. (50) over $X_{K^{\complement}} \sim p^\tau_{K^{\complement}}$ gives

$$\mathbb{P}(X' \in \mathcal{X}_{-\tau}) = \mathbb{E}_{X_{K^{\complement}} \sim p^\tau_{K^{\complement}}}\left[p(\mathcal{X}_{-\tau} \mid X_{K^{\complement}})\right].$$

Step 2: typical–atypical split. Because $|K^{\complement}| \ge s_{\mathrm{pin}}$, Asm. 4.1 applies with $V = K^{\complement}$. We split the expectation according to whether the visible context belongs to the typical set $\mathcal{E}^\tau_{K^{\complement}}$:

$$\mathbb{P}\left(X' \in \mathcal{X}_{-\tau}\right) = \mathbb{E}_{X_{K^{\complement}} \sim p^\tau_{K^{\complement}}}\left[p(\mathcal{X}_{-\tau} \mid X_{K^{\complement}}) \mathbb{1}\{X_{K^{\complement}} \in \mathcal{E}^\tau_{K^{\complement}}\}\right] + \mathbb{E}_{X_{K^{\complement}} \sim p^\tau_{K^{\complement}}}\left[p(\mathcal{X}_{-\tau} \mid X_{K^{\complement}}) \mathbb{1}\{X_{K^{\complement}} \notin \mathcal{E}^\tau_{K^{\complement}}\}\right]. \tag{51}$$

For the first term in Eq. (51), the definition of $\mathcal{E}^\tau_{K^\complement}$ in Asm. 4.1 guarantees $p(\mathcal{X}_{-\tau} \mid X_{K^\complement} = x_{K^\complement}) \le e^{-\kappa|K^\complement|}$ for every $x_{K^\complement} \in \mathcal{E}^\tau_{K^\complement}$, hence

$$\mathbb{E}_{X_{K^\complement}\sim p^\tau_{K^\complement}}\left[p(\mathcal{X}_{-\tau} \mid X_{K^\complement})\mathbb{1}\{X_{K^\complement} \in \mathcal{E}^\tau_{K^\complement}\}\right] \le e^{-\kappa|K^\complement|}. \tag{52}$$

For the second term, we use the trivial bound $p(\mathcal{X}_{-\tau} \mid X_{K^\complement}) \le 1$ alongside the typical-set mass bound $p^\tau_{K^\complement}(\mathcal{E}^\tau_{K^\complement}) \ge 1 - e^{-c_1|K^\complement|}$ from Asm. 4.1:

$$\mathbb{E}_{X_{K^\complement}\sim p^\tau_{K^\complement}}\left[p(\mathcal{X}_{-\tau} \mid X_{K^\complement})\mathbb{1}\{X_{K^\complement} \notin \mathcal{E}^\tau_{K^\complement}\}\right] \;\le\; p^\tau_{K^\complement}\big((\mathcal{E}^\tau_{K^\complement})^\complement\big) \;\le\; e^{-c_1|K^\complement|}. \tag{53}$$

Combining Eqs. (51) to (53) yields

$$\mathbb{P}(X' \in \mathcal{X}_{-\tau}) \le e^{-\kappa|K^\complement|} + e^{-c_1|K^\complement|}.$$

Since $c = \min\{c_1, \kappa\}$, both exponents are bounded by $-c|K^\complement|$. Therefore,

$$\mathbb{P}(X' \in \mathcal{X}_{-\tau}) \le 2e^{-c|K^\complement|},$$

which proves Eq. (47).

Step 3: averaging over the schedule. We now consider the mixed block dynamics. Draw $K \sim \mu$ independently of X and let $\widetilde{X}$ be the induced update of part (ii). Conditioning on the drawn mask, the law of total probability gives

$$\mathbb{P}\Big(\widetilde{X} \in \mathcal{X}_{-\tau}\Big) = \mathbb{E}_{K\sim\mu}\left[\mathbb{P}\Big(\widetilde{X} \in \mathcal{X}_{-\tau}\,\Big|\, K\Big)\right].$$

Given K, the configuration $\widetilde{X}$ is the fixed-mask update of part (i) for that mask, so the conditional crossing probability is the quantity bounded above whenever $|K^\complement| \ge s_{\mathrm{pin}}$. Fix $s \ge s_{\mathrm{pin}}$ and split the mask space into $\{|K^\complement| \le s\}$ and $\{|K^\complement| > s\}$. On the first event we use the trivial upper bound one; on the second, $|K^\complement| > s \ge s_{\mathrm{pin}}$, so Eq. (47) applies. Hence,

$$\mathbb{P}\Big(\widetilde{X} \in \mathcal{X}_{-\tau}\Big) \le \mu\big(|K^\complement| \le s\big) + \mathbb{E}_{K\sim\mu}\left[2e^{-c|K^\complement|}\mathbb{1}\{|K^\complement| > s\}\right]. \tag{54}$$

The first term is $\pi_s(\mu)$ by Eq. (1). For the second term, on $\{|K^\complement| > s\}$ we have $e^{-c|K^\complement|} \le e^{-cs}$ since $c > 0$, so

$$\mathbb{E}_{K\sim\mu}\left[2e^{-c|K^\complement|}\mathbb{1}\{|K^\complement| > s\}\right] \;\le\; 2e^{-cs}\,\mathbb{P}_{K\sim\mu}\big(|K^\complement| > s\big) \;\le\; 2e^{-cs}.$$

Substituting both into Eq. (54) proves Eq. (48).

Step 4: minimum visible size. Finally, suppose $v_{\min}(\mu) \ge s_{\mathrm{pin}}$. By the definition of $v_{\min}(\mu)$, $|K^\complement| \ge v_{\min}(\mu) \ge s_{\mathrm{pin}}$ for every $K \in \mathrm{supp}(\mu)$, so Eq. (47) applies to every mask in the support of μ. Averaging over $K \sim \mu$,

$$\mathbb{P}\Big(\widetilde{X} \in \mathcal{X}_{-\tau}\Big) \;\le\; \mathbb{E}_{K\sim\mu}\left[2e^{-c|K^\complement|}\right] \;\le\; 2e^{-c\,v_{\min}(\mu)},$$

where the last inequality uses $|K^\complement| \ge v_{\min}(\mu)$. This proves Eq. (49). □

C.6 Reverse Mode Blindness

This section shows that the mode-blindness bound persists when the masked discrepancy is taken in the reverse KL direction.

Lemma C.5 (Mode blindness for reverse KL)**.** *Suppose Asm. 4.1 holds with combined rate $c = \min\{c_1, \kappa\}$, and let $B_0 = \log\frac{(2-c_0)(1-c_0)}{c_0^2}$ and $A_0 = e^{B_0} - 1 + B_0$ be the constants of the proof of Thm. 4.1. Set $C_2 := 2e^{B_0}A_0 + B_0 \le 16\,c_0^{-4}$. Then every mask schedule μ with $v_{\min}(\mu) \ge s_{\mathrm{pin}}$ satisfies*

$$\sup_{\lambda\in I_{\mathrm{bl}}} \mathfrak{D}_\mu(q_\lambda \,\|\, p) \;\le\; C_2\,\mathbb{E}_{K\sim\mu}\big[e^{-c|K^\complement|}\big] \;\le\; C_2\,e^{-c\,v_{\min}(\mu)}. \tag{55}$$

Proof intuition. Reversing the KL divergence changes both the order of the posterior comparison and the distribution used to average contexts. A bounded logit shift preserves exponential posterior confidence, so the reversed conditional KL remains exponentially small on the confidence set. Its complement has exponentially small probability under each component separately, and therefore remains negligible after changing the mixture weights. Together, these two bounds preserve the overall exponential decay rate.

Proof. Fix $\lambda \in I_{\mathrm{bl}}$ and $K \in \mathrm{supp}(\mu)$. Since $v_{\min}(\mu) \geq s_{\mathrm{pin}}$, every visible set below satisfies $|K^{\complement}| \geq s_{\mathrm{pin}}$.

Step 1: binary reduction. Since $\lambda \in (0,1)$, we have $\mathrm{supp}(q_\lambda) = \mathrm{supp}(p)$, meaning that all marginals and conditionals of the two distributions share the same support. For any $x_{K^{\complement}} \in \mathrm{supp}(p_{K^{\complement}})$ such that $\beta_{K^{\complement}} \in (0,1)$, both masked-block conditionals act as mixtures of the same disjointly supported components $\nu_{\pm}$. As established in the proof of Lemma 3.1(iii), these mixtures have weights $(\beta_{K^{\complement}}, 1-\beta_{K^{\complement}})$ and $(\beta_{\lambda,K^{\complement}}, 1-\beta_{\lambda,K^{\complement}})$. Applying Lemma C.2 in either direction yields

$$D_{\mathrm{KL}}\big(q_{\lambda,K}(\cdot \mid x_{K^{\complement}}) \,\big\|\, p_K(\cdot \mid x_{K^{\complement}})\big) = \mathrm{kl}\big(\beta_{\lambda,K^{\complement}} \,\big\|\, \beta_{K^{\complement}}\big).$$

Furthermore, the boundary fibers where $\beta_{K^{\complement}} \in \{0,1\}$ contribute zero to the divergence since $\beta_{\lambda,K^{\complement}} = \beta_{K^{\complement}}$ in those cases. It then follows that

$$\mathfrak{D}_K(q_\lambda \| p) = \mathbb{E}_{X_{K^{\complement}} \sim q_{\lambda,K^{\complement}}}[\mathrm{kl}(\beta_{\lambda,K^{\complement}}(X_{K^{\complement}}) \| \beta_{K^{\complement}}(X_{K^{\complement}}))].$$

Step 2: fiberwise bound. On the interior fibers, Lemma 3.1(ii) gives that $\mathrm{logit}(\beta_{K^{\complement}}) = \mathrm{logit}(\beta_{\lambda,K^{\complement}}) - \Delta_\lambda$. Recall $\Delta_\lambda := \mathrm{logit}(\lambda) - \mathrm{logit}(w)$. Since $w \in [c_0, 1-c_0]$ and $\lambda \in I_{\mathrm{bl}} = [c_0/2, 1-c_0/2]$, we have $|\Delta_\lambda| \leq B_0$.

By applying the fiberwise bound from Step 3 of the proof of Thm. 4.1 with a base point of $\beta_{\lambda,K^{\complement}}$ and a shift of $-\Delta_\lambda$, we obtain

$$\mathrm{kl}(\beta_{\lambda,K^{\complement}} \| \beta_{K^{\complement}}) \leq \min\big\{B_0,\ A_0 \min\{\beta_{\lambda,K^{\complement}}, 1-\beta_{\lambda,K^{\complement}}\}\big\}.$$

A logit shift of size at most B_0 increases the smaller posterior probability by at most a factor $2e^{B_0}$. To see this, write $o := \beta_{K^{\complement}}/(1-\beta_{K^{\complement}})$ and $o_\lambda := \beta_{\lambda,K^{\complement}}/(1-\beta_{\lambda,K^{\complement}})$ for the two posterior odds, and note the elementary sandwich, valid for every $\beta \in (0,1)$ with odds o,

$$\min\{\beta,\, 1-\beta\} \ \leq\ \min\{o,\, 1/o\} \ \leq\ 2\min\{\beta,\, 1-\beta\},$$

which follows, for $\beta \leq \frac{1}{2}$, from $\beta \leq \beta/(1-\beta) \leq 2\beta$, and by the symmetry $(\beta, o) \mapsto (1-\beta, 1/o)$ otherwise. Since $\mathrm{logit}(\beta_{\lambda,K^{\complement}}) = \mathrm{logit}(\beta_{K^{\complement}}) + \Delta_\lambda$, the odds transform as $o_\lambda = e^{\Delta_\lambda} o$, whence

$$\min\{o_\lambda, 1/o_\lambda\} \leq e^{|\Delta_\lambda|} \min\{o, 1/o\} \leq e^{B_0} \min\{o, 1/o\}.$$

Chaining the three bounds,

$$\min\{\beta_{\lambda,K^{\complement}},\, 1-\beta_{\lambda,K^{\complement}}\} \ \leq\ \min\{o_\lambda,\, 1/o_\lambda\} \ \leq\ e^{B_0} \min\{o,\, 1/o\} \ \leq\ 2e^{B_0} \min\{\beta_{K^{\complement}},\, 1-\beta_{K^{\complement}}\}.$$

Step 3: averaging. Let $G_{K^{\complement}}$ denote the pooled confidence set defined in Lemma C.3. Within $G_{K^{\complement}}$, the fiberwise bound evaluates to at most $2e^{B_0} A_0\, e^{-\kappa|K^{\complement}|}$, while outside $G_{K^{\complement}}$ we rely on the uniform bound B_0. The proof of Lemma C.3 gives

$$p^{\tau}_{K^{\complement}}(G^{\complement}_{K^{\complement}}) \leq e^{-c_1|K^{\complement}|} \qquad \text{for each } \tau \in \{+,-\}.$$

Since $q_{\lambda,K^{\complement}} = \lambda p^{+}_{K^{\complement}} + (1-\lambda) p^{-}_{K^{\complement}}$,

$$\begin{aligned} q_{\lambda,K^{\complement}}(G^{\complement}_{K^{\complement}}) &= \lambda p^{+}_{K^{\complement}}(G^{\complement}_{K^{\complement}}) + (1-\lambda) p^{-}_{K^{\complement}}(G^{\complement}_{K^{\complement}}) \\ &\leq e^{-c_1|K^{\complement}|}. \end{aligned}$$

Combining these observations gives

$$\mathfrak{D}_K(q_\lambda \| p) \leq 2e^{B_0} A_0\, e^{-\kappa|K^{\complement}|} + B_0\, e^{-c_1|K^{\complement}|} \leq C_2\, e^{-c|K^{\complement}|}.$$

Taking the average over $K \sim \mu$ establishes the main claim. Finally, to derive the numerical bound, we let $x := \frac{(2-c_0)(1-c_0)}{c_0^2} \leq 2c_0^{-2}$. By applying the inequality $\log x \leq x - 1$, we conclude that

$$C_2 = 2x(x - 1 + \log x) + \log x \leq (x-1)(4x+1) \leq 4x^2 \leq 16\, c_0^{-4}.$$

Since this bound is uniform in $\lambda \in I_{\mathrm{bl}}$, averaging over $K \sim \mu$ and taking the supremum over λ gives the claimed schedule bound. □

C.7 Factorized Per-Token Losses

Defn. 2.1 scores the *joint* masked-block conditional, whereas BERT-style pretraining sums *per-token* losses over the masked block (Devlin et al., 2019; Liu et al., 2019), the masked analog of pseudolikelihood (Besag, 1974). For $q \in \mathcal{Q}(p)$, a mask K, and $i \in K$, write $q_i(\cdot \mid x_{K^\complement})$ for the coordinate-i conditional marginal, and define

$$\mathfrak{D}_K^{\mathrm{fac}}(p\|q) := \mathbb{E}_{X_{K^\complement}\sim p_{K^\complement}} \sum_{i\in K} D_{\mathrm{KL}}\big(p_i(\cdot \mid X_{K^\complement})\|q_i(\cdot \mid X_{K^\complement})\big), \quad \mathfrak{D}_\mu^{\mathrm{fac}}(p\|q) := \mathbb{E}_{K\sim\mu}\, \mathfrak{D}_K^{\mathrm{fac}}(p\|q). \tag{56}$$

Corollary C.1 (Mode blindness under per-token log loss)**.** *Suppose Asm. 4.1 holds with combined rate c, and let $C \le 6c_0^{-2}$ be the constant of Thm. 4.1. Then every mask schedule μ with $v_{\min}(\mu) \ge s_{\mathrm{pin}}$ satisfies*

$$\sup_{\lambda\in I_{\mathrm{bl}}} \mathfrak{D}_\mu^{\mathrm{fac}}(p\|q_\lambda) \;\le\; C\,N\,\mathbb{E}_{K\sim\mu}\big[e^{-c|K^\complement|}\big] \;\le\; C\,e^{-c\,v_{\min}(\mu)+\log N}. \tag{57}$$

Proof intuition. Projecting a masked block onto one coordinate cannot increase KL divergence. The joint masked-block discrepancy therefore bounds each per-token discrepancy, and summing over the masked coordinates introduces a factor of at most N. When the visible count grows linearly with N, this exponential decay easily absorbs the factor N, keeping the overall bound exponentially small.

Proof. Fix $\lambda \in I_{\mathrm{bl}}$, $K \in \mathrm{supp}(\mu)$, and $x_{K^\complement} \in \mathrm{supp}(p_{K^\complement})$. The coordinate projection $x_K \mapsto x_i$ is a stochastic map, so the data-processing inequality and Lemma 3.1(iii) give, for every $i \in K$,

$$\begin{aligned} D_{\mathrm{KL}}\big(p_i(\cdot \mid x_{K^\complement})\|q_{\lambda,i}(\cdot \mid x_{K^\complement})\big) &\le D_{\mathrm{KL}}\big(p_K(\cdot \mid x_{K^\complement})\|q_{\lambda,K}(\cdot \mid x_{K^\complement})\big) \\ &= \mathrm{kl}\big(\beta_{K^\complement}(x_{K^\complement})\|\beta_{\lambda,K^\complement}(x_{K^\complement})\big). \end{aligned}$$

Summing over $i \in K$ and using $|K| \le N$, then averaging over contexts as in Eq. (69) of the proof of Thm. 4.1 and over $K \sim \mu$, proves Eq. (57). □

Mode information under factorization. If $v_{\min}(\mu) \ge \max\{s_{\mathrm{pin}}, 2c^{-1}\log N\}$, then Eq. (57) gives

$$\sup_{\lambda\in I_{\mathrm{bl}}} \mathfrak{D}_\mu^{\mathrm{fac}}(p\|q_\lambda) \le Ce^{-c\,v_{\min}(\mu)/2}.$$

Since $D_{\mathrm{TV}}(p, q_\lambda) = |w - \lambda|$, the same reweighted alternatives give a nonvanishing lower bound on the modulus defined through $\mathfrak{D}_\mu^{\mathrm{fac}}$, as in Eq. (21). In particular, when $v_{\min}(\mu) = \Omega(N)$, this occurs at an excess-risk tolerance exponentially small in N. The mode-weight recovery bound does not follow from the same comparison. Although the fiber components $\nu_\pm$ have disjoint supports, their coordinate marginals $\nu_{\pm,i}$ may overlap. Coordinate projection can therefore discard mode information, and data processing provides only an upper bound on the per-token KL divergence. At the full mask, for example,

$$D_{\mathrm{TV}}(p_i, q_{\lambda,i}) = |w - \lambda|\, D_{\mathrm{TV}}(p_i^+, p_i^-).$$

Universal joint-distribution control from the full-mask term also fails under factorization. That term scores only the unconditional coordinate marginals. In particular, the product distribution $q^{\mathrm{prod}} := \bigotimes_{i=1}^N p_i$ belongs to $\mathcal{Q}(p)$ and satisfies

$$\mathfrak{D}_{[N]}^{\mathrm{fac}}(p\|q^{\mathrm{prod}}) = 0.$$

Indeed, $p(x) > 0$ implies $p_i(x_i) > 0$ for every coordinate, so q^{prod} covers the support of p. For the construction in Prop. G.1, with $\theta \in (0,1)$ and $w \in (0,1)$ fixed as $N \to \infty$,

$$D_{\mathrm{TV}}(p, q^{\mathrm{prod}}) = 1 - o(1).$$

To see this, p concentrates near empirical magnetizations $\pm\theta$, whereas q^{prod} concentrates near $(2w-1)\theta$, which lies strictly between them. When $w = 1/2$, q^{prod} is the uniform distribution. The full-mask joint-KL control of Prop. 4.1 therefore relies on scoring the joint masked-block distribution.

C.8 A Sufficient Condition for Asm. 4.2

Lemma C.6 (Bounded visible evidence implies residual uncertainty). *Fix $L_0 > 0$ and $V \subseteq [N]$, and suppose the mode-conditional visible marginals satisfy*

$$\left|\log \frac{p_V^+(x_V)}{p_V^-(x_V)}\right| \le L_0|V| \qquad \text{for all } x_V \in \mathrm{supp}(p_V). \tag{58}$$

Let $w := p(\mathcal{X}_+) \in (0,1)$. Then

$$\mathrm{mmse}_p(Z \mid X_V) \ge w(1-w)\, e^{-L_0|V|}. \tag{59}$$

In particular, Asm. 4.2 holds on any scale $s_{\mathrm{ov}} \ge s_{\mathrm{pin}}$ for which Eq. (58) holds for every $|V| \le s_{\mathrm{ov}}$ with the same constant L_0, taking $u_0 = w(1-w)$ and $L = L_0$.

> **Proof intuition.** Bayes' rule reweights the two modes by their visible-context likelihoods. After dividing by one likelihood, the normalizing constant becomes a convex combination of 1 and the likelihood ratio, and is thus bounded between them. This controls how much reweighting can diminish the posterior mode variance for any given context, a property preserved after averaging over all contexts.

Proof. Fix $x_V \in \mathrm{supp}(p_V)$ and write $\ell := \log\big(p_V^+(x_V)/p_V^-(x_V)\big)$, so $|\ell| \le L_0|V|$ by Eq. (58). By Bayes' rule with prior $p(\mathcal{X}_+) = w$,

$$\beta_V(x_V) = \frac{w\, p_V^+(x_V)}{w\, p_V^+(x_V) + (1-w)\, p_V^-(x_V)} = \frac{w\, e^{\ell}}{w\, e^{\ell} + (1-w)},$$

dividing through by $p_V^-(x_V) > 0$, as implied by Eq. (58). Hence

$$\beta_V(x_V)\big(1-\beta_V(x_V)\big) = \frac{w(1-w)\, e^{\ell}}{\big(w\, e^{\ell} + (1-w)\big)^2}.$$

We upper-bound the denominator by splitting on the sign of ℓ. If $\ell \ge 0$, then $we^{\ell} + (1-w) \le we^{\ell} + (1-w)e^{\ell} = e^{\ell}$; if $\ell < 0$, then $we^{\ell} + (1-w) \le w + (1-w) = 1$. Writing $\ell^+ := \max\{\ell, 0\}$, both cases are subsumed by

$$we^{\ell} + (1-w) \;\le\; e^{\ell^+}.$$

Therefore

$$\begin{aligned}\beta_V(x_V)\big(1-\beta_V(x_V)\big) &\ge w(1-w)e^{\ell - 2\ell^+} \\ &= w(1-w)e^{-|\ell|} \\ &\ge w(1-w)e^{-L_0|V|},\end{aligned}$$

where the last step uses $|\ell| \le L_0|V|$ from Eq. (58).

Taking $\mathbb{E}_{X_V \sim p_V}$ of both sides and recalling $\mathrm{mmse}_p(Z \mid X_V) = \mathbb{E}[\beta_V(1-\beta_V)]$ proves Eq. (59). □

D Proofs of Statements

This section provides full proofs for all formal statements deferred from the main text.

D.1 Proof of Lemma 3.1

Proof. Fix $x_V \in \mathrm{supp}(p_V)$ and abbreviate $\beta_V = \beta_V(x_V)$ and $\beta_{\lambda,V} = \beta_{\lambda,V}(x_V)$.

Step 1: density ratio and the reweighted posterior. The reweighted distribution satisfies

$$q_\lambda(x) = \begin{cases} \dfrac{\lambda}{w}\, p(x), & x \in \mathcal{X}_+, \\ \dfrac{1-\lambda}{1-w}\, p(x), & x \in \mathcal{X}_-. \end{cases} \tag{60}$$

Summing Eq. (60) over completions in the two mode regions gives

$$q_\lambda(\mathcal{X}_+, X_V = x_V) = \frac{\lambda}{w}\, p(\mathcal{X}_+, X_V = x_V), \tag{61}$$

$$q_\lambda(\mathcal{X}_-, X_V = x_V) = \frac{1-\lambda}{1-w}\, p(\mathcal{X}_-, X_V = x_V). \tag{62}$$

By Bayes' rule, dividing the numerator and denominator by $p_V(x_V) > 0$ yields Eq. (13):

$$\beta_{\lambda,V} = \frac{q_\lambda(\mathcal{X}_+, X_V = x_V)}{q_\lambda(\mathcal{X}_+, X_V = x_V) + q_\lambda(\mathcal{X}_-, X_V = x_V)} = \frac{\frac{\lambda}{w}\,\beta_V}{\frac{\lambda}{w}\,\beta_V + \frac{1-\lambda}{1-w}\,(1-\beta_V)}.$$

Step 2: the logit shift. If $\beta_V \in (0,1)$, both terms in the denominator of Eq. (13) are positive. Taking posterior odds gives

$$\frac{\beta_{\lambda,V}}{1-\beta_{\lambda,V}} = \frac{\lambda}{1-\lambda}\,\frac{1-w}{w}\,\frac{\beta_V}{1-\beta_V}.$$

Taking logarithms proves Eq. (14):

$$\operatorname{logit}\big(\beta_{\lambda,V}(x_V)\big) = \operatorname{logit}\big(\beta_V(x_V)\big) + \operatorname{logit}(\lambda) - \operatorname{logit}(w).$$

Step 3: fiber mixtures and the binary reduction. Continue with the case $\beta_V \in (0,1)$. Both conditional component distributions

$$\nu_\tau := p(X_{V^\complement} \in \cdot \,|\, X_V = x_V,\ X \in \mathcal{X}_\tau), \qquad \tau \in \{+,-\},$$

are well defined. Reweighting changes only the prior masses of the two mode regions, so the same conditional component distributions appear under q_λ. Hence

$$p_{V^\complement}(\cdot \mid x_V) = \beta_V\,\nu_+ + (1-\beta_V)\,\nu_-, \tag{63}$$

$$q_{\lambda,V^\complement}(\cdot \mid x_V) = \beta_{\lambda,V}\,\nu_+ + (1-\beta_{\lambda,V})\,\nu_-. \tag{64}$$

The supports of ν_+ and ν_- are disjoint. Indeed, for fixed x_V, any completion $x_{V^\complement}$ determines a unique full configuration $(x_V, x_{V^\complement})$, which belongs to at most one of the disjoint regions $\mathcal{X}_+$ and $\mathcal{X}_-$. Applying the KL identity of Lemma C.2 to the mixtures Eqs. (63) and (64) proves Eq. (15) when $\beta_V \in (0,1)$:

$$D_{\mathrm{KL}}\big(p_{V^\complement}(\cdot \mid x_V) \,\|\, q_{\lambda,V^\complement}(\cdot \mid x_V)\big) = \mathrm{kl}(\beta_V \,\|\, \beta_{\lambda,V}).$$

Step 4: boundary posteriors. Finally, if $\beta_V = 1$, then $p(\mathcal{X}_-, X_V = x_V) = 0$. By Eq. (62), the same is true under q_λ, so $\beta_{\lambda,V} = 1$. Moreover, both conditional block distributions equal the within-$\mathcal{X}_+$ conditional distribution. The case $\beta_V = 0$ is symmetric. Thus the conditional divergence in Eq. (15) vanishes, completing the proof. □

D.2 Proof of Prop. 3.1

> **Proof intuition.** Because the two distributions share their within-mode components, their joint KL divergence depends only on the mode weights. The KL chain rule decomposes this divergence into a visible-marginal term and a masked-conditional term. The masked objective therefore measures only the portion of the mode-weight discrepancy that remains unresolved by the visible coordinates.

Proof. Fix a mask $K \subseteq [N]$. Because $\lambda \in (0,1)$, the model distribution satisfies $\mathrm{supp}(q_\lambda) = \mathrm{supp}(p)$, ensuring that the conditional $q_{\lambda,K}(\cdot \mid x_{K^\complement})$ is well defined for all $x_{K^\complement} \in \mathrm{supp}(p_{K^\complement})$.

By Lemma 3.1(iii), the pointwise divergence between the masked conditionals satisfies

$$D_{\mathrm{KL}}(p_K(\cdot \mid x_{K^\complement}) \,\|\, q_{\lambda,K}(\cdot \mid x_{K^\complement})) = \mathrm{kl}\big(\beta_{K^\complement}(x_{K^\complement}) \,\|\, \beta_{\lambda,K^\complement}(x_{K^\complement})\big).$$

Taking the expectation over $X_{K^\complement} \sim p_{K^\complement}$ directly yields Eq. (16) by the definition of the masked log discrepancy $\mathfrak{D}_K(p\|q_\lambda)$ (Defn. 2.1).

To establish Eq. (17), decompose the joint divergence along $(X_{K^\complement}, X_K)$ by the KL chain rule:

$$D_{\mathrm{KL}}(p\|q_\lambda) = D_{\mathrm{KL}}(p_{K^\complement}\|q_{\lambda,K^\complement}) + \mathbb{E}_{X_{K^\complement}\sim p_{K^\complement}}[D_{\mathrm{KL}}(p_K(\cdot \mid X_{K^\complement}) \,\|\, q_{\lambda,K}(\cdot \mid X_{K^\complement}))], \tag{65}$$

where the conditional term is $\mathfrak{D}_K(p\|q_\lambda)$ by Defn. 2.1.

The mode distributions p^+ and p^- have disjoint supports, so p and q_λ are shared-component mixtures in the sense of Lemma C.2, with weights w and λ:

$$D_{\mathrm{KL}}(p\|q_\lambda) = \mathrm{kl}(w\|\lambda). \tag{66}$$

Substituting Eq. (66) into Eq. (65) and rearranging gives

$$\mathfrak{D}_K(p\|q_\lambda) = \mathrm{kl}(w\|\lambda) - D_{\mathrm{KL}}(p_{K^\complement}\|q_{\lambda,K^\complement}),$$

which proves Eq. (17). □

D.3 Proof of Thm. 4.1

Proof intuition. Reweighting shifts the posterior log-odds by a bounded amount. The resulting conditional KL divergence is controlled by the smaller posterior mode probability. Mode pinning makes this probability exponentially small on typical visible contexts, while atypical contexts carry exponentially small probability. Averaging over contexts yields the masked-discrepancy upper bound. Meanwhile, the mixing weights can still maintain a separation bounded away from zero, establishing the total-variation lower bound.

Proof. Step 1: uniform bound on the logit shift. Recall $I_{\mathrm{bl}} = [c_0/2, 1 - c_0/2]$ from the theorem statement. Under Asm. 4.1, the true weight satisfies $w \in [c_0, 1 - c_0]$. Since logit is increasing and $\mathrm{logit}(1 - x) = -\mathrm{logit}(x)$, the posterior log-odds shift $\Delta_\lambda := \mathrm{logit}(\lambda) - \mathrm{logit}(w)$ satisfies

$$|\Delta_\lambda| \le \mathrm{logit}(1 - \tfrac{c_0}{2}) - \mathrm{logit}(c_0) = \log\frac{(2 - c_0)(1 - c_0)}{c_0^2} =: B_0$$

for every $\lambda \in I_{\mathrm{bl}}$. This bound depends only on c_0. Set $A_0 := e^{B_0} - 1 + B_0$.

Step 2: reduction to the binary divergence. Fix $\lambda \in I_{\mathrm{bl}}$ and a mask $K \in \mathrm{supp}(\mu)$. Since $v_{\min}(\mu) \ge s_{\mathrm{pin}}$, we have $|K^\complement| \ge s_{\mathrm{pin}}$. For every $x_{K^\complement} \in \mathrm{supp}(p_{K^\complement})$, abbreviating $\beta_{K^\complement} = \beta_{K^\complement}(x_{K^\complement})$ and $\beta_{\lambda,K^\complement} = \beta_{\lambda,K^\complement}(x_{K^\complement})$, Lemma 3.1(iii) provides

$$D_{\mathrm{KL}}(p_K(\cdot \mid x_{K^\complement})\|q_{\lambda,K}(\cdot \mid x_{K^\complement})) = \mathrm{kl}(\beta_{K^\complement}\|\beta_{\lambda,K^\complement}). \tag{67}$$

Step 3: fiberwise divergence bounds. We bound this binary divergence in two regimes. For $\beta_{K^\complement} \in (0, 1)$, Lemma 3.1(ii) gives $\mathrm{logit}(\beta_{\lambda,K^\complement}) = \mathrm{logit}(\beta_{K^\complement}) + \Delta_\lambda$, which algebraically rearranges to

$$\mathrm{kl}(\beta_{K^\complement}\|\beta_{\lambda,K^\complement}) = \log\big(1 - \beta_{K^\complement} + \beta_{K^\complement}e^{\Delta_\lambda}\big) - \beta_{K^\complement}\Delta_\lambda.$$

This expression admits two immediate upper bounds. First, since $1 - \beta_{K^\complement} + \beta_{K^\complement}e^{\Delta_\lambda} \le e^{\max\{0,\Delta_\lambda\}}$, we obtain the uniform bound

$$\mathrm{kl}(\beta_{K^\complement}\|\beta_{\lambda,K^\complement}) \le \max\{0, \Delta_\lambda\} - \beta_{K^\complement}\Delta_\lambda \le |\Delta_\lambda| \le B_0.$$

Second, applying $\log(1 + t) \le t$ yields

$$\mathrm{kl}(\beta_{K^\complement}\|\beta_{\lambda,K^\complement}) \le \beta_{K^\complement}(e^{|\Delta_\lambda|} - 1 + |\Delta_\lambda|) \le A_0\beta_{K^\complement}.$$

By symmetry, replacing $(\beta_{K^\complement}, \Delta_\lambda)$ with $(1 - \beta_{K^\complement}, -\Delta_\lambda)$ in the same argument yields the complementary bound $A_0(1 - \beta_{K^\complement})$. Combining these properties gives

$$\mathrm{kl}(\beta_{K^\complement}\|\beta_{\lambda,K^\complement}) \le \min\big\{B_0,\ A_0\min\{\beta_{K^\complement}, 1 - \beta_{K^\complement}\}\big\}. \tag{68}$$

The same bound holds when $\beta_{K^\complement} \in \{0, 1\}$, since $\beta_{\lambda,K^\complement} = \beta_{K^\complement}$ and the divergence is zero.

Step 4: averaging over contexts and masks. We now integrate over the visible context. Let $G_{K^{\complement}}$ denote the pooled confidence set from Lemma C.3, which applies because $|K^{\complement}| \geq s_{\text{pin}}$. On $G_{K^{\complement}}$, the residual uncertainty is exponentially small:

$$\min\{\beta_{K^{\complement}}(X_{K^{\complement}}), 1-\beta_{K^{\complement}}(X_{K^{\complement}})\} \leq e^{-\kappa|K^{\complement}|}.$$

On the complement $G_{K^{\complement}}^{\complement}$, we use the uniform bound B_0, noting that $p_{K^{\complement}}(G_{K^{\complement}}^{\complement}) \leq e^{-c_1|K^{\complement}|}$. Splitting the expectation of Eq. (68) over $X_{K^{\complement}} \sim p_{K^{\complement}}$ accordingly yields

$$\begin{aligned}\mathbb{E}_{X_{K^{\complement}}\sim p_{K^{\complement}}}\left[\mathrm{kl}\big(\beta_{K^{\complement}}(X_{K^{\complement}})\|\beta_{\lambda,K^{\complement}}(X_{K^{\complement}})\big)\right] &\leq A_0 e^{-\kappa|K^{\complement}|} + B_0\, p_{K^{\complement}}(G_{K^{\complement}}^{\complement}) \\ &\leq (A_0+B_0)e^{-c|K^{\complement}|}, \qquad \text{where } c=\min\{c_1,\kappa\}.\end{aligned} \tag{69}$$

Define $C := A_0 + B_0$. We can write $C = x - 1 + 2\log x$ where $x := \frac{(2-c_0)(1-c_0)}{c_0^2}$. Since $\log x \leq x-1$ and $x \leq 2c_0^{-2}$, it follows immediately that $C \leq 3(x-1) \leq 6c_0^{-2}$.

By Defn. 2.1 and Eq. (67), the masked discrepancy $\mathfrak{D}_\mu(p\|q_\lambda)$ is the average of the left-hand side of Eq. (69) over $K \sim \mu$. Therefore, taking the supremum over $\lambda \in I_{\text{bl}}$ gives

$$\sup_{\lambda\in I_{\text{bl}}} \mathfrak{D}_\mu(p\|q_\lambda) \leq C\,\mathbb{E}_{K\sim\mu}\left[e^{-c|K^{\complement}|}\right] \leq C\, e^{-c\, v_{\min}(\mu)},$$

which proves Eq. (20).

Step 5: identifiability modulus lower bound. Since the reweighted models satisfy $q_\lambda \in \mathcal{Q}(p)$, any tolerance $\varepsilon \geq C\,\mathbb{E}_{K\sim\mu}[e^{-c|K^{\complement}|}]$ ensures that the entire family $\{q_\lambda\}_{\lambda\in I_{\text{bl}}}$ achieves discrepancy $\mathfrak{D}_\mu(p\|q_\lambda) \leq \varepsilon$ and is thus admissible in the supremum defining $\mathfrak{A}_{\text{TV}}^{\mu}(p,\varepsilon)$. Because p and q_λ share the same disjointly supported components, Lemma C.2 yields $D_{\text{TV}}(p,q_\lambda) = |w-\lambda|$. Using the structural bound $w \in [c_0, 1-c_0] \subseteq I_{\text{bl}}$, we explicitly evaluate the supremum over the interval:

$$\begin{aligned}\mathfrak{A}_{\text{TV}}^{\mu}(p,\varepsilon) \;\geq\; \sup_{\lambda\in I_{\text{bl}}}|w-\lambda| &= \max\{w-c_0/2,\, 1-c_0/2-w\} \\ &\geq \frac{(w-c_0/2)+(1-c_0/2-w)}{2} = \frac{1-c_0}{2}.\end{aligned}$$

This proves Eq. (21) and completes the proof. □

D.4 Proof of Cor. 4.1

> **Proof intuition.** An AT inequality controls the joint KL divergence by a constant times the masked discrepancy. The chosen reweighting keeps the joint divergence bounded away from zero while making the masked discrepancy exponentially small, forcing the AT constant to grow exponentially. The forward comparison lower-bounds the constant for q_λ, while the reverse comparison lower-bounds the constant for p.

Proof. Step 1: reweighting and joint separation. Pick $\lambda \in \{c_0/2,\, 1-c_0/2\}$ maximizing $|w-\lambda|$. Since Asm. 4.1 gives $w \in [c_0, 1-c_0]$, as in Step 5 of the proof of Thm. 4.1,

$$|w-\lambda| = \max\{w-c_0/2,\; 1-c_0/2-w\} \geq \frac{(w-c_0/2)+(1-c_0/2-w)}{2} = \frac{1-c_0}{2}.$$

Because p and q_λ share the same disjointly supported components, Lemma C.2 gives the binary KL identities. Pinsker's inequality then yields

$$D_{\text{KL}}(p\|q_\lambda) = \mathrm{kl}(w\|\lambda) \geq 2(w-\lambda)^2 \geq \tfrac{(1-c_0)^2}{2}, \qquad D_{\text{KL}}(q_\lambda\|p) = \mathrm{kl}(\lambda\|w) \geq \tfrac{(1-c_0)^2}{2}.$$

The equality $\mathrm{supp}(p) = \mathrm{supp}(q_\lambda)$ makes both applications of Eq. (8) below admissible. In each case, the AT constant is bounded below by the ratio of the joint divergence to the masked discrepancy, with a positive numerator divided by zero interpreted as $+\infty$.

Step 2: model-side bound. Thm. 4.1 gives $\mathfrak{D}_\mu(p\|q_\lambda) \le C\, e^{-c\, v_{\min}(\mu)}$ with $C \le 6c_0^{-2}$. Taking $r = q_\lambda$, $p' = p$ in Eq. (8),

$$\bar{C}_{\mathrm{AT}}(q_\lambda, \mu) \;\ge\; \frac{D_{\mathrm{KL}}(p\|q_\lambda)}{\mathfrak{D}_\mu(p\|q_\lambda)} \;\ge\; \frac{(1-c_0)^2/2}{6c_0^{-2}}\, e^{c\, v_{\min}(\mu)} \;=\; \frac{c_0^2(1-c_0)^2}{12}\, e^{c\, v_{\min}(\mu)}.$$

Step 3: data-side bound. Lemma C.5 gives $\mathfrak{D}_\mu(q_\lambda\|p) \le C_2\, e^{-c\, v_{\min}(\mu)}$ with $C_2 \le 16\, c_0^{-4}$. Taking $r = p$, $p' = q_\lambda$ in Eq. (8),

$$\bar{C}_{\mathrm{AT}}(p, \mu) \;\ge\; \frac{D_{\mathrm{KL}}(q_\lambda\|p)}{\mathfrak{D}_\mu(q_\lambda\|p)} \;\ge\; \frac{(1-c_0)^2/2}{16c_0^{-4}}\, e^{c\, v_{\min}(\mu)} \;=\; \frac{c_0^4(1-c_0)^2}{32}\, e^{c\, v_{\min}(\mu)},$$

completing the proof. □

D.5 Proof of Lemma 4.1

Proof intuition. The prior and each context-conditional posterior undergo the same log-odds shift. Along this shift, the second derivative of the binary KL divergence is the Bernoulli variance. Comparing these variances relates the conditional divergence to the prior divergence by a factor proportional to the posterior variance. Averaging this posterior variance over visible contexts yields the residual mode MMSE.

Proof. Recall $\delta_I = \mathrm{logit}(b) - \mathrm{logit}(a)$ with $I = [a, b]$ from the statement of Lemma 4.1. Fix a mask $K \subseteq [N]$ and weights $w, \lambda \in I$. Set

$$\delta := \mathrm{logit}(\lambda) - \mathrm{logit}(w), \qquad |\delta| \le \delta_I.$$

For $\beta \in (0,1)$ and $t \in \mathbb{R}$, define the logit path

$$\beta_t := \mathrm{logit}^{-1}\big(\mathrm{logit}(\beta) + t\big),$$

so that $\beta_0 = \beta$ and, by Lemma 3.1(ii), the reweighted posterior on every interior fiber is the endpoint of this path: if $\beta = \beta_{K^{\complement}}(x_{K^{\complement}}) \in (0,1)$, then $\beta_{\lambda,K^{\complement}}(x_{K^{\complement}}) = \beta_\delta$.

For the prior weight, define

$$w_t := \mathrm{logit}^{-1}\big(\mathrm{logit}(w) + t\big), \qquad t \in \mathbb{R}.$$

Then $w_0 = w$ and $w_\delta = \lambda$ by the definition of δ.

Step 1: integral representation of the binary divergence. Fix $\beta \in (0,1)$ and let $f(t) := \mathrm{kl}(\beta\|\beta_t)$. Since $t \mapsto \beta_t$ is smooth with values in $(0,1)$, f is smooth. Differentiating and using the logistic identity $\frac{\mathrm{d}}{\mathrm{d}t}\beta_t = \beta_t(1-\beta_t)$,

$$f'(t) = \beta_t(1-\beta_t)\left(\frac{1-\beta}{1-\beta_t} - \frac{\beta}{\beta_t}\right) = \beta_t - \beta, \qquad f''(t) = \beta_t(1-\beta_t).$$

Hence $f(0) = 0$ and $f'(0) = 0$, and Taylor's theorem with integral remainder gives, for either sign of δ,

$$\mathrm{kl}(\beta\|\beta_\delta) = \int_0^\delta (\delta - t)\, \beta_t(1-\beta_t)\, \mathrm{d}t = \delta^2 \int_0^1 (1-s)\, \beta_{\delta s}(1-\beta_{\delta s})\, \mathrm{d}s,$$

where the second form follows from the substitution $t = \delta s$ and is manifestly nonnegative.

Step 2: variance comparison along the path. For every $\beta \in (0,1)$ and $t \in \mathbb{R}$, writing $z := e^{\mathrm{logit}(\beta)}$,

$$\frac{\beta_t(1-\beta_t)}{\beta(1-\beta)} = e^t\, \frac{(1+z)^2}{(1+ze^t)^2}.$$

Since $\min\{1, e^t\}(1+z) \le 1 + ze^t \le \max\{1, e^t\}(1+z)$, both sign cases give

$$e^{-|t|}\, \beta(1-\beta) \;\le\; \beta_t(1-\beta_t) \;\le\; e^{|t|}\, \beta(1-\beta). \tag{70}$$

Along the integral of Step 1 we have $|t| \le |\delta| \le \delta_I$. Substituting Eq. (70) and $\int_0^1 (1-s)\,\mathrm{d}s = \frac{1}{2}$ yields the two-sided fiber bound

$$\frac{e^{-\delta_I}}{2}\,\beta(1-\beta)\,\delta^2 \;\le\; \mathrm{kl}(\beta\|\beta_\delta) \;\le\; \frac{e^{\delta_I}}{2}\,\beta(1-\beta)\,\delta^2. \tag{71}$$

Step 3: comparison with the prior divergence. Recall $m_I := \min_{u\in I} u(1-u) > 0$. For every $s \in [0,1]$,

$$\mathrm{logit}(w_{\delta s}) = (1-s)\,\mathrm{logit}(w) + s\,\mathrm{logit}(\lambda).$$

Since logit^{-1} is increasing, $w_{\delta s}$ lies between w and λ and therefore belongs to I. Hence

$$m_I \le w_{\delta s}(1-w_{\delta s}) \le \tfrac{1}{4}.$$

Applying the integral representation from Step 1 gives

$$\mathrm{kl}(w\|\lambda) = \delta^2 \int_0^1 (1-s)\,w_{\delta s}(1-w_{\delta s})\;\mathrm{d}s.$$

Using $\int_0^1 (1-s)\;\mathrm{d}s = \frac{1}{2}$, we obtain

$$\frac{m_I}{2}\,\delta^2 \le \mathrm{kl}(w\|\lambda) \le \frac{1}{8}\,\delta^2, \qquad \text{equivalently} \quad 8\,\mathrm{kl}(w\|\lambda) \le \delta^2 \le \frac{2}{m_I}\,\mathrm{kl}(w\|\lambda). \tag{72}$$

Step 4: two-sided fiberwise bound. Substituting Eq. (72) into Eq. (71) eliminates δ^2: for every $\beta \in (0,1)$,

$$4e^{-\delta_I}\,\beta(1-\beta)\,\mathrm{kl}(w\|\lambda) \;\le\; \mathrm{kl}(\beta\|\beta_\delta) \;\le\; \frac{e^{\delta_I}}{m_I}\,\beta(1-\beta)\,\mathrm{kl}(w\|\lambda). \tag{73}$$

For $\beta \in \{0,1\}$, extend the path by setting $\beta_t := \beta$ for every $t \in \mathbb{R}$. Both $\mathrm{kl}(\beta\|\beta_\delta)$ and $\beta(1-\beta)$ then vanish, so Eq. (73) holds for every $\beta \in [0,1]$.

By Lemma 3.1, for every $x_{K^c} \in \mathrm{supp}(p_{K^c})$, setting $\beta = \beta_{K^c}(x_{K^c})$ gives $\beta_\delta = \beta_{\lambda,K^c}(x_{K^c})$, including when the posterior is zero or one. Thus Eq. (73) applies to every supported visible context.

Step 5: average over contexts. Taking the expectation of Eq. (73) over $X_{K^c} \sim p_{K^c}$ and using Eq. (16) of Prop. 3.1 for the middle term and Eq. (11) for the outer terms,

$$4e^{-\delta_I}\;\mathrm{mmse}_p(Z \mid X_{K^c})\,\mathrm{kl}(w\|\lambda) \;\le\; \mathfrak{D}_K(p\|q_\lambda) \;\le\; \frac{e^{\delta_I}}{m_I}\;\mathrm{mmse}_p(Z \mid X_{K^c})\,\mathrm{kl}(w\|\lambda).$$

This proves Eq. (23) with

$$c_I := 4e^{-\delta_I}, \qquad C_I := \frac{e^{\delta_I}}{m_I},$$

both depending only on I. □

Remark D.1 (Exact binary reduction). The tractability of the analysis comes from the mode-reweighting family itself. Because $\{q_\lambda\}$ preserves the two within-mode distributions and varies only their mixture weight, the N-dimensional comparison reduces on every visible fiber to a one-parameter binary problem (Lemma 3.1). This yields both the exact information decomposition of Prop. 3.1 and the Fisher limit in Eq. (24). The local limit in Eq. (24) has coefficient $1/(w(1-w))$. Consequently, any upper constant valid uniformly over $\lambda, w \in I$ for a nondegenerate interval I must be at least $\sup_{u\in I} 1/(u(1-u))$. Thus no finite upper constant works uniformly over intervals approaching the boundary of $(0,1)$. This reduction also enables the enumerations in Sec. 5 and Appx. H to compare the theoretical bounds with directly computed masked discrepancies.

D.6 Proof of Cor. 4.2

Proof. Step 1: elementary bounds. By Eq. (11), the residual mode uncertainty is given by

$$\mathrm{mmse}_p(Z \mid X_V) = \mathbb{E}_{X_V \sim p_V}\big[\beta_V(X_V)(1-\beta_V(X_V))\big].$$

We abbreviate $\beta_V(X_V)$ as β_V when the argument is clear. For every $\beta_V \in [0,1]$, the following two bounds hold. First, because $\max\{\beta_V, 1-\beta_V\} \le 1$, the product satisfies

$$\beta_V(1-\beta_V) = \min\{\beta_V, 1-\beta_V\} \cdot \max\{\beta_V, 1-\beta_V\} \le \min\{\beta_V, 1-\beta_V\}. \tag{74}$$

Second,

$$\beta_V(1-\beta_V) = \tfrac{1}{4} - (\beta_V - \tfrac{1}{2})^2 \le \tfrac{1}{4}. \tag{75}$$

Step 2: the pooled confidence set. Recall from Lemma C.3 the pooled confidence set:

$$G_V = \big\{x_V \in \mathrm{supp}(p_V) : \min\{\beta_V(x_V), 1-\beta_V(x_V)\} \le e^{-\kappa|V|}\big\}.$$

Since $|V| \ge s_{\mathrm{pin}}$ by hypothesis, Lemma C.3 ensures that the probability of encountering an atypical context is exponentially small:

$$p_V(G_V^{\complement}) \le e^{-c_1|V|}, \qquad \text{where} \quad G_V^{\complement} := \mathrm{supp}(p_V) \setminus G_V.$$

Step 3: splitting the expectation. We bound the total expectation by splitting it over these two sets,

$$\mathrm{mmse}_p(Z \mid X_V) = \mathbb{E}_{X_V \sim p_V}\big[\beta_V(1-\beta_V)\mathbb{1}\{X_V \in G_V\}\big] + \mathbb{E}_{X_V \sim p_V}\big[\beta_V(1-\beta_V)\mathbb{1}\{X_V \in G_V^{\complement}\}\big].$$

On the typical set G_V, Eq. (74) and the definition of G_V give $\beta_V(1-\beta_V) \le e^{-\kappa|V|}$, so the first term is at most $e^{-\kappa|V|}\, p_V(G_V) \le e^{-\kappa|V|}$. On the atypical set $G_V^{\complement}$, Eq. (75) bounds the integrand by $\frac{1}{4}$, so the second term is at most $\frac{1}{4}\, p_V(G_V^{\complement}) \le \frac{1}{4} e^{-c_1|V|}$. Summing these two contributions yields:

$$\mathrm{mmse}_p(Z \mid X_V) \le e^{-\kappa|V|} + \frac{1}{4} e^{-c_1|V|} \le \frac{5}{4} e^{-c|V|},$$

where the final inequality uses $c = \min\{c_1, \kappa\}$, so that $e^{-\kappa|V|} \le e^{-c|V|}$ and $e^{-c_1|V|} \le e^{-c|V|}$. □

D.7 Proof of Thm. 4.2

Proof intuition. The fixed-mask bound reduces schedule dependence to the average residual mode uncertainty. Under the low-visibility assumption, each mask below the visibility threshold retains a guaranteed minimum level of uncertainty, yielding a lower bound proportional to $\pi_s(\mu)$. For the upper bound, the Bernoulli variance is at most $1/4$, so low-visibility masks contribute at most $\pi_s(\mu)/4$, while mode pinning exponentially suppresses the contribution from masks above the threshold.

Proof. Step 1: averaging the fixed-mask sensitivity. By Lemma 4.1, for every mask $K \subseteq [N]$ and every $\lambda \in I$,

$$c_I\ \mathrm{mmse}_p(Z \mid X_{K^{\complement}})\, \mathrm{kl}(w\|\lambda) \ \le\ \mathfrak{D}_K(p\|q_\lambda) \ \le\ C_I\ \mathrm{mmse}_p(Z \mid X_{K^{\complement}})\, \mathrm{kl}(w\|\lambda).$$

Since $\mathrm{kl}(w\|\lambda)$ is independent of K, averaging over $K \sim \mu$ and using Defn. 2.1 gives Eq. (26).

For any visibility threshold s, split the schedule average of the residual MMSE as

$$\begin{aligned}\mathbb{E}_{K\sim\mu}\big[\,\mathrm{mmse}_p(Z \mid X_{K^{\complement}})\big] = &\underbrace{\mathbb{E}_{K\sim\mu}\big[\,\mathrm{mmse}_p(Z \mid X_{K^{\complement}})\,\mathbb{1}\{|K^{\complement}| \le s\}\big]}_{=:\mathsf{T}_{\mathrm{low}}} \\ &+ \underbrace{\mathbb{E}_{K\sim\mu}\big[\,\mathrm{mmse}_p(Z \mid X_{K^{\complement}})\,\mathbb{1}\{|K^{\complement}| > s\}\big]}_{=:\mathsf{T}_{\mathrm{high}}}.\end{aligned} \tag{76}$$

Step 2: the low-visibility lower bound. Suppose Asm. 4.2 holds and fix $0 \le s \le s_{\mathrm{ov}}$. On the event $\{|K^{\complement}| \le s\}$, we have $|K^{\complement}| \le s \le s_{\mathrm{ov}}$, so Asm. 4.2 applies to the visible set $K^{\complement}$ and gives

$$\mathrm{mmse}_p(Z \mid X_{K^{\complement}}) \ \ge\ u_0\, e^{-L|K^{\complement}|} \ \ge\ u_0\, e^{-Ls},$$

where the second inequality uses $|K^{\complement}| \le s$ and $L > 0$. Therefore

$$\mathsf{T}_{\text{low}} \;\ge\; u_0\, e^{-Ls}\, \mathbb{P}_{K\sim\mu}\big(|K^{\complement}| \le s\big) = u_0\, e^{-Ls}\, \pi_s(\mu), \tag{77}$$

by the definition of the low-visibility mass. Since $\mathsf{T}_{\text{high}} \ge 0$, combining Eqs. (76) and (77) with the lower half of Eq. (26) gives

$$\mathfrak{D}_\mu(p\|q_\lambda) \;\ge\; c_I\, \mathrm{kl}(w\|\lambda)\, \mathsf{T}_{\text{low}} \;\ge\; c_I\, u_0\, e^{-Ls}\, \pi_s(\mu)\, \mathrm{kl}(w\|\lambda),$$

which proves Eq. (27).

Step 3: the high-visibility upper bound. Suppose Asm. 4.1 holds and fix $s \ge s_{\text{pin}}$. On the event $\{|K^{\complement}| \le s\}$, the elementary bound $\beta(1-\beta) \le \frac{1}{4}$ for $\beta \in [0,1]$ gives $\mathrm{mmse}_p(Z \mid X_{K^{\complement}}) \le \frac{1}{4}$, hence

$$\mathsf{T}_{\text{low}} \;\le\; \tfrac{1}{4}\, \pi_s(\mu). \tag{78}$$

On the event $\{|K^{\complement}| > s\}$, we have $|K^{\complement}| > s \ge s_{\text{pin}}$, so Cor. 4.2 applies and gives $\mathrm{mmse}_p(Z \mid X_{K^{\complement}}) \le \frac{5}{4} e^{-c|K^{\complement}|}$, hence

$$\mathsf{T}_{\text{high}} \;\le\; \tfrac{5}{4}\, \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\, \mathbb{1}\{|K^{\complement}| > s\}\big]. \tag{79}$$

Combining Eqs. (76), (78) and (79) with the upper half of Eq. (26),

$$\begin{aligned}\mathfrak{D}_\mu(p\|q_\lambda) \;&\le\; C_I \left(\tfrac{1}{4}\, \pi_s(\mu) + \tfrac{5}{4}\, \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\, \mathbb{1}\{|K^{\complement}| > s\}\big]\right) \mathrm{kl}(w\|\lambda) \\ &\le\; \tfrac{5}{4} C_I \Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\, \mathbb{1}\{|K^{\complement}| > s\}\big]\Big) \mathrm{kl}(w\|\lambda).\end{aligned}$$

This proves Eq. (28) and completes the proof. □

D.8 Proof of Cor. 4.3

Proof. Throughout, fix $\lambda \in I$ and recall $m_I = \min_{u\in I} u(1-u)$ from Lemma 4.1.

Step 1: recovery via the low-visibility lower bound. Suppose Asm. 4.2 holds. Fix $0 \le s \le s_{\text{ov}}$ with $\pi_s(\mu) > 0$, and assume $\mathfrak{D}_\mu(p\|q_\lambda) \le \varepsilon$. Since $\pi_s(\mu) > 0$ and $c_I, u_0 > 0$, the lower bound Eq. (27) of Thm. 4.2 can be solved for the binary divergence:

$$\mathrm{kl}(w\|\lambda) \;\le\; \frac{\mathfrak{D}_\mu(p\|q_\lambda)}{c_I\, u_0\, e^{-Ls}\, \pi_s(\mu)} \;\le\; \frac{e^{Ls}\, \varepsilon}{c_I\, u_0\, \pi_s(\mu)}. \tag{80}$$

By Pinsker's inequality applied to the pair of Bernoulli distributions with parameters w and λ, whose total-variation distance is $|w-\lambda|$,

$$\mathrm{kl}(w\|\lambda) \;\ge\; 2\,|w-\lambda|^2. \tag{81}$$

Combining Eqs. (80) and (81) and taking square roots,

$$|w-\lambda| \;\le\; \sqrt{\frac{e^{Ls}\, \varepsilon}{2\, c_I\, u_0\, \pi_s(\mu)}} \;=\; \frac{e^{Ls/2}}{\sqrt{2 c_I\, u_0}} \sqrt{\frac{\varepsilon}{\pi_s(\mu)}},$$

which proves Eq. (29).

Step 2: a quadratic upper bound on the binary divergence. For the converse, we first bound $\mathrm{kl}(w\|\lambda)$ by a quadratic. Using $\log x \le x - 1$ for $x > 0$ on both terms,

$$\mathrm{kl}(w\|\lambda) = w \log \frac{w}{\lambda} + (1-w) \log \frac{1-w}{1-\lambda} \;\le\; w\, \frac{w-\lambda}{\lambda} + (1-w)\, \frac{\lambda - w}{1-\lambda} = \frac{(w-\lambda)^2}{\lambda(1-\lambda)},$$

where the final equality follows from

$$\frac{w}{\lambda} - \frac{1-w}{1-\lambda} = \frac{w(1-\lambda) - \lambda(1-w)}{\lambda(1-\lambda)} = \frac{w-\lambda}{\lambda(1-\lambda)}.$$

Since $\lambda \in I$, we conclude

$$\mathrm{kl}(w\|\lambda) \;\le\; \frac{|w-\lambda|^2}{m_I}. \tag{82}$$

Step 3: the converse via the high-visibility upper bound. Suppose Asm. 4.1 holds. Fix $s \ge s_{\mathrm{pin}}$, and assume the hypothesis of Eq. (30):

$$\frac{5C_I}{4m_I}\Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\,\mathbb{1}\{|K^{\complement}|>s\}\big]\Big)\,|w-\lambda|^2 \;\le\; \varepsilon.$$

Then, combining the upper bound Eq. (28) of Thm. 4.2 with Eq. (82),

$$\begin{aligned}\mathfrak{D}_\mu(p\|q_\lambda) \;&\le\; \tfrac{5}{4}C_I\Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\,\mathbb{1}\{|K^{\complement}|>s\}\big]\Big)\,\mathrm{kl}(w\|\lambda)\\ &\le\; \frac{5C_I}{4m_I}\Big(\pi_s(\mu) + \mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\,\mathbb{1}\{|K^{\complement}|>s\}\big]\Big)\,|w-\lambda|^2\\ &\le\; \varepsilon,\end{aligned}$$

which proves Eq. (30) and completes the proof. □

D.9 Proof of Prop. 4.1

Proof intuition. The full mask directly measures the joint KL divergence. For necessity, consider two configurations that differ in every coordinate. Any nonempty visible context uniquely identifies which configuration occurred, so perturbing their mixture weights leaves all conditionals under non-full masks unchanged. Only the full mask detects the weight mismatch, which establishes the sharpness of the control constant.

Proof. Step 1: sufficiency and the modulus bound. Suppose $\pi_0(\mu) > 0$, and fix $p \in \Delta(\mathcal{X})$ and $q \in \mathcal{Q}(p)$. At the full mask, the visible context is empty, so

$$\mathfrak{D}_{[N]}(p\|q) = D_{\mathrm{KL}}(p\|q). \tag{83}$$

Since every fixed-mask discrepancy is nonnegative,

$$\begin{aligned}\mathfrak{D}_\mu(p\|q) &= \pi_0(\mu)\,D_{\mathrm{KL}}(p\|q) + \sum_{K\subsetneq[N]}\mu(K)\,\mathfrak{D}_K(p\|q)\\ &\ge \pi_0(\mu)\,D_{\mathrm{KL}}(p\|q).\end{aligned} \tag{84}$$

Thus $C_\mu = 1/\pi_0(\mu)$ satisfies Eq. (31) uniformly over p and q.

If $\mathfrak{D}_\mu(p\|q) \le \varepsilon$, then Eq. (84) and Pinsker's inequality give

$$D_{\mathrm{TV}}(p,q) \le \sqrt{\frac{1}{2}\,D_{\mathrm{KL}}(p\|q)} \le \sqrt{\frac{\varepsilon}{2\,\pi_0(\mu)}}.$$

Taking the supremum over all such $q \in \mathcal{Q}(p)$ proves Eq. (32).

Step 2: necessity and optimality. Choose configurations $x^-, x^+ \in \mathcal{X}$ that differ in every coordinate, and define

$$p = \tfrac{1}{2}\delta_{x^-} + \tfrac{1}{2}\delta_{x^+}, \qquad q_\lambda = \lambda\delta_{x^+} + (1-\lambda)\delta_{x^-},$$

where $\lambda \in (0,1)\setminus\{\frac{1}{2}\}$. These distributions have the same support, so $p \ll q_\lambda$ and $q_\lambda \in \mathcal{Q}(p)$.

Fix $K \subsetneq [N]$ and write $V = K^{\complement}$. Because $V \ne \varnothing$, the two visible configurations x_V^- and x_V^+ are distinct. They constitute the support of p_V and both have positive probability under $(q_\lambda)_V$. Conditioning on either context therefore identifies the configuration under both distributions:

$$p_K(\cdot \mid x_V^\sigma) = (q_\lambda)_K(\cdot \mid x_V^\sigma) = \delta_{x_K^\sigma}, \qquad \sigma \in \{-,+\}.$$

Hence the conditional distributions coincide p_V-almost surely, and $\mathfrak{D}_K(p\|q_\lambda) = 0$.

Since p and q_λ are distinct and have the same finite support, the divergence $\kappa_\lambda := D_{\mathrm{KL}}(p\|q_\lambda)$ satisfies $0 < \kappa_\lambda < \infty$. Combining the vanishing non-full-mask discrepancies with Eq. (83) yields

$$\mathfrak{D}_\mu(p\|q_\lambda) = \pi_0(\mu)\,\kappa_\lambda = \pi_0(\mu)\,D_{\mathrm{KL}}(p\|q_\lambda). \tag{85}$$

Any finite constant satisfying Eq. (31) uniformly over p, q must therefore satisfy

$$\kappa_\lambda \le C_\mu\,\pi_0(\mu)\,\kappa_\lambda.$$

Dividing by $\kappa_\lambda > 0$ gives $1 \le C_\mu \pi_0(\mu)$. This inequality is impossible when $\pi_0(\mu) = 0$ and forces $C_\mu \ge 1/\pi_0(\mu)$ otherwise. Together with Step 1, this proves necessity and optimality. □

Remark D.2 (Sharpness of the full-mask TV scaling under pinning). The constant $1/\pi_0(\mu)$ in Eq. (31) is optimal uniformly over data distributions, as witnessed by the two-point construction in the proof. We now address sharpness of the TV bound Eq. (32) for a fixed distribution under Asm. 4.1. Write $\pi_0 := \pi_0(\mu) > 0$, and suppose all non-full masks in the support of μ satisfy $|K^{\complement}| \ge s_{\mathrm{pin}}$. For $\lambda \in I_{\mathrm{bl}}$, the full-mask case of Eq. (17) and Eq. (69) give

$$\mathfrak{D}_\mu(p\|q_\lambda) \;\le\; \pi_0\,\mathrm{kl}(w\|\lambda) + \underbrace{C\sum\nolimits_{K\subsetneq[N]} \mu(K) e^{-c|K^{\complement}|}}_{=:r_\mu}.$$

Define $m := \min_{u\in I_{\mathrm{bl}}} u(1-u) > 0$ and $\eta_I := \max_{\lambda\in I_{\mathrm{bl}}} \mathrm{kl}(w\|\lambda) > 0$. Suppose $r_\mu < \pi_0\eta_I$. For every tolerance satisfying

$$2r_\mu \le \varepsilon \le 2\pi_0\eta_I,$$

continuity of $\lambda \mapsto \mathrm{kl}(w\|\lambda)$ on I_{bl}, together with $\mathrm{kl}(w\|w) = 0$, ensures that there exists $\lambda \in I_{\mathrm{bl}}$ such that

$$\mathrm{kl}(w\|\lambda) = \frac{\varepsilon}{2\pi_0}.$$

For this choice, $\mathfrak{D}_\mu(p\|q_\lambda) \le \varepsilon$. Moreover, $\mathrm{kl}(w\|\lambda) \le |w-\lambda|^2/m$ and $D_{\mathrm{TV}}(p, q_\lambda) = |w-\lambda|$. Combining this with Eq. (32) yields

$$\sqrt{\frac{m}{2}}\sqrt{\frac{\varepsilon}{\pi_0}} \;\le\; \mathfrak{A}^{\mu}_{\mathrm{TV}}(p,\varepsilon) \;\le\; \frac{1}{\sqrt{2}}\sqrt{\frac{\varepsilon}{\pi_0}}.$$

Thus the $\sqrt{\varepsilon/\pi_0}$ scaling is sharp throughout this range, with constants controlled by the fixed weight interval I_{bl}. The lower tolerance threshold accounts for residual sensitivity from non-full masks, while the upper threshold keeps the witness within the admissible reweighting interval.

E Mode Blindness with Multiple Modes ($M \ge 2$)

This section extends Thm. 4.1 to partitions into a fixed number $M \ge 2$ of regions. Under M-mode pinning, large-context masked prediction can be exponentially insensitive to the global mixture weights while the within-mode distributions remain fixed.

E.1 Mode Structure and Pinning

Let $\mathcal{X} = \bigsqcup_{j\in[M]} \mathcal{X}_j$ be a partition where $w_j := p(\mathcal{X}_j) > 0$ for every j. We write $\boldsymbol{w} := (w_1, \dots, w_M)$ for the true mode-weight vector and $p^j := p(\cdot \mid \mathcal{X}_j)$ for the within-mode distributions. Let

$$\Delta_{M-1} := \Big\{\boldsymbol{\lambda} \in \mathbb{R}^M : \lambda_j \ge 0 \;\; \forall\, j \in [M],\; \sum\nolimits_{j\in[M]} \lambda_j = 1\Big\}.$$

The relative interior of this simplex is its interior within the affine hyperplane $\{\boldsymbol{\lambda} \in \mathbb{R}^M : \sum_{j=1}^M \lambda_j = 1\}$:

$$\mathrm{relint}(\Delta_{M-1}) := \{\boldsymbol{\lambda} \in \Delta_{M-1} : \lambda_j > 0 \text{ for every } j \in [M]\}.$$

For any $\boldsymbol{\lambda} \in \mathrm{relint}(\Delta_{M-1})$, we define the reweighted distribution

$$q_{\boldsymbol{\lambda}} := \sum\nolimits_{j\in[M]} \lambda_j\, p^j. \tag{86}$$

This construction preserves every within-mode conditional distribution of p while setting $q_{\boldsymbol{\lambda}}(\mathcal{X}_j) = \lambda_j$. As in the binary case, $\mathrm{supp}(q_{\boldsymbol{\lambda}}) = \mathrm{supp}(p)$, and thus $q_{\boldsymbol{\lambda}} \in \mathcal{Q}(p)$. The mode indicator becomes the one-hot vector $\mathbf{Z} := \big(\mathbb{1}\{X \in \mathcal{X}_j\}\big)_{j\in[M]}$, with posterior

$$\boldsymbol{\beta}_V(x_V) := \big(p(X \in \mathcal{X}_j \mid X_V = x_V)\big)_{j\in[M]} \in \Delta_{M-1}.$$

We denote the corresponding posterior under $q_{\boldsymbol{\lambda}}$ as $\boldsymbol{\beta}_{\boldsymbol{\lambda},V}$. We rely on two scalar summaries of $\boldsymbol{\beta}_V$. The first is the *vertex deficiency*

$$\gamma_V(x_V) := 1 - \max_{j\in[M]} \beta_{V,j}(x_V), \tag{87}$$

which vanishes when the visible context perfectly identifies the region. The second is the one-hot residual MMSE, equivalently the expected Gini impurity of the posterior:

$$\mathrm{mmse}_p(\mathbf{Z} \mid X_V) = \mathbb{E}\big[\|\mathbf{Z} - \mathbb{E}[\mathbf{Z} \mid X_V]\|^2\big] = \mathbb{E}_{X_V\sim p_V}\big[\,1 - \|\boldsymbol{\beta}_V(X_V)\|_2^2\,\big]. \tag{88}$$

For $M = 2$, this is twice the quantity in Eq. (11). Pointwise, the vertex deficiency and the posterior Gini impurity satisfy

$$\gamma_V(x_V) \le 1 - \|\boldsymbol{\beta}_V(x_V)\|_2^2 \le 2\gamma_V(x_V).$$

Indeed, writing $\gamma = \gamma_V(x_V)$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_V(x_V)$, we have $(1-\gamma)^2 \le \|\boldsymbol{\beta}\|_2^2 \le 1-\gamma$, which gives $\gamma \le 1 - \|\boldsymbol{\beta}\|_2^2 \le 2\gamma - \gamma^2 \le 2\gamma$. Averaging over visible contexts therefore yields

$$\mathbb{E}_{X_V\sim p_V}\big[\gamma_V(X_V)\big] \le \mathrm{mmse}_p(\mathbf{Z} \mid X_V) \le 2\,\mathbb{E}_{X_V\sim p_V}\big[\gamma_V(X_V)\big].$$

The pinning condition. With the M-mode geometry established, we assume that large visible contexts identify the mode with high confidence.

Assumption E.1 (M-mode pinning)**.** *There exist constants $c_0 \in (0, 1/M]$, $c_1, \kappa > 0$, and a visibility threshold $s_{\mathrm{pin}} \in [N]$ such that $w_j \ge c_0$ for every $j \in [M]$. Moreover, for every visible set $V \subseteq [N]$ with $|V| \ge s_{\mathrm{pin}}$ and every mode $j \in [M]$, the typical set*

$$\mathcal{E}_V^j := \big\{x_V \in \mathrm{supp}(p_V^j) : p(\mathcal{X} \setminus \mathcal{X}_j \mid X_V = x_V) \le e^{-\kappa|V|}\big\}$$

satisfies

$$p_V^j(\mathcal{E}_V^j) \ge 1 - e^{-c_1|V|}.$$

As before, we define the combined rate $c := \min\{c_1, \kappa\}$. For $M = 2$, Asm. E.1 reduces to Asm. 4.1.

E.2 Mixture Identities and Posterior Tilting

To analyze the discrepancy between the true distribution and its reweighted counterpart, we establish how total variation and KL divergence behave over disjoint mixture components.

Lemma E.1 (Shared-component mixtures with M components)**.** *Let $\mathcal{Z}$ be a finite set, and let $\nu_1, \dots, \nu_M \in \Delta(\mathcal{Z})$ have pairwise disjoint supports. For $\boldsymbol{a} \in \Delta_{M-1}$, define*

$$m_{\boldsymbol{a}} := \sum_{j=1}^M a_j \nu_j.$$

Then, for every $\boldsymbol{a}, \boldsymbol{b} \in \Delta_{M-1}$,

$$D_{\mathrm{TV}}(m_{\boldsymbol{a}}, m_{\boldsymbol{b}}) = \tfrac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|_1, \qquad D_{\mathrm{KL}}(m_{\boldsymbol{a}} \| m_{\boldsymbol{b}}) = D_{\mathrm{KL}}(\boldsymbol{a} \| \boldsymbol{b}),$$

with the usual extended-value conventions for zero coordinates.

Proof. Let $S_j := \mathrm{supp}(\nu_j)$ for $j \in [M]$. Since these supports are pairwise disjoint, for every $z \in S_j$,

$$m_{\boldsymbol{a}}(z) = a_j \nu_j(z), \qquad m_{\boldsymbol{b}}(z) = b_j \nu_j(z).$$

Hence

$$\|m_{\boldsymbol{a}} - m_{\boldsymbol{b}}\|_1 = \sum_{j=1}^M \sum_{z\in S_j} |a_j - b_j|\,\nu_j(z) = \sum_{j=1}^M |a_j - b_j| = \|\boldsymbol{a} - \boldsymbol{b}\|_1,$$

because $\sum_{z\in S_j}\nu_j(z)=1$. Therefore,

$$D_{\mathrm{TV}}(m_{\boldsymbol{a}},m_{\boldsymbol{b}})=\frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|_1.$$

For the KL divergence, splitting over the disjoint supports gives

$$D_{\mathrm{KL}}(m_{\boldsymbol{a}}\|m_{\boldsymbol{b}})=\sum_{j=1}^{M}\sum_{z\in S_j}a_j\nu_j(z)\log\frac{a_j\nu_j(z)}{b_j\nu_j(z)}=\sum_{j=1}^{M}a_j\log\frac{a_j}{b_j}=D_{\mathrm{KL}}(\boldsymbol{a}\|\boldsymbol{b}),$$

with the equality understood in the extended sense: if $a_j>0=b_j$ for some j, both sides are $+\infty$, while terms with $a_j=0$ vanish. □

Posterior tilting and information decomposition. Building on this mixture geometry, we next characterize how reweighting the global modes shifts the visible posterior. The tilt determines the fiberwise divergence, and hence the masked discrepancy.

Lemma E.2 (M-mode posterior tilt and information decomposition)**.** *Fix $V\subseteq[N]$, let $K=[N]\backslash V$, and take $x_V\in\mathrm{supp}(p_V)$ and $\boldsymbol{\lambda}\in\mathrm{relint}(\Delta_{M-1})$. Then*

$$\beta_{\boldsymbol{\lambda},V,j}(x_V)=\frac{(\lambda_j/w_j)\,\beta_{V,j}(x_V)}{\sum_{l\in[M]}(\lambda_l/w_l)\,\beta_{V,l}(x_V)},\tag{89}$$

so the posterior under $q_{\boldsymbol{\lambda}}$ is an exponential tilt of the posterior under p by the log-ratio vector $\boldsymbol{t}:=(\log(\lambda_j/w_j))_{j\in[M]}$. Moreover,

$$D_{\mathrm{KL}}\big(p_{V^{\complement}}(\cdot\mid x_V)\,\|\,q_{\boldsymbol{\lambda},V^{\complement}}(\cdot\mid x_V)\big)=D_{\mathrm{KL}}\big(\boldsymbol{\beta}_V(x_V)\,\|\,\boldsymbol{\beta}_{\boldsymbol{\lambda},V}(x_V)\big).\tag{90}$$

Consequently, for every mask K with visible set $K^{\complement}=[N]\setminus K$,

$$\mathfrak{D}_K(p\|q_{\boldsymbol{\lambda}})=\mathbb{E}_{X_{K^{\complement}}\sim p_{K^{\complement}}}\big[D_{\mathrm{KL}}(\boldsymbol{\beta}_{K^{\complement}}\|\boldsymbol{\beta}_{\boldsymbol{\lambda},K^{\complement}})\big]=D_{\mathrm{KL}}(\boldsymbol{w}\|\boldsymbol{\lambda})-D_{\mathrm{KL}}(p_{K^{\complement}}\|q_{\boldsymbol{\lambda},K^{\complement}}).\tag{91}$$

Proof. As in Lemma 3.1, $q_{\boldsymbol{\lambda}}$ has density λ_j/w_j relative to p on $\mathcal{X}_j$. Summing over completions on each region and normalizing yields Eq. (89).

For Eq. (90), fix $x_V\in\mathrm{supp}(p_V)$ and abbreviate $\beta_{V,j}=\beta_{V,j}(x_V)$ and $\beta_{\boldsymbol{\lambda},V,j}=\beta_{\boldsymbol{\lambda},V,j}(x_V)$. For each j with $\beta_{V,j}>0$, define

$$\nu_j:=p(X_K\in\cdot\,|\,X_V=x_V,\ X\in\mathcal{X}_j)\,.$$

By Eq. (89), $\beta_{\boldsymbol{\lambda},V,j}>0$ if and only if $\beta_{V,j}>0$. Reweighting preserves the conditional distribution within each mode, so

$$p_K(\cdot\mid x_V)=\sum_{j:\beta_{V,j}>0}\beta_{V,j}\nu_j,\qquad q_{\boldsymbol{\lambda},K}(\cdot\mid x_V)=\sum_{j:\beta_{V,j}>0}\beta_{\boldsymbol{\lambda},V,j}\nu_j.$$

For fixed x_V, each completion x_K determines a full configuration belonging to exactly one mode region. The components in these sums therefore have pairwise disjoint supports. The shared-component identity of Lemma E.1 gives

$$D_{\mathrm{KL}}\big(p_K(\cdot\mid x_V)\,\|\,q_{\boldsymbol{\lambda},K}(\cdot\mid x_V)\big)=\sum_{j:\beta_{V,j}>0}\beta_{V,j}\log\frac{\beta_{V,j}}{\beta_{\boldsymbol{\lambda},V,j}}=D_{\mathrm{KL}}\big(\boldsymbol{\beta}_V(x_V)\,\|\,\boldsymbol{\beta}_{\boldsymbol{\lambda},V}(x_V)\big),$$

where omitted indices have zero probability under both posteriors. This proves Eq. (90).

Finally, for every mask K, set $V=K^{\complement}$. The first equality in Eq. (91) is exactly Defn. 2.1. The second follows from the KL chain rule together with $D_{\mathrm{KL}}(p\|q_{\boldsymbol{\lambda}})=D_{\mathrm{KL}}(\boldsymbol{w}\|\boldsymbol{\lambda})$, which is Lemma E.1 applied to p and $q_{\boldsymbol{\lambda}}$. □

E.3 Mode Blindness and the Modulus Lower Bound

The decomposition in Lemma E.2 shows that the masked discrepancy depends on the KL divergence between the original and tilted posteriors. We bound this divergence using the vertex deficiency, capturing the behavior when the visible context is highly confident.

Lemma E.3 (KL bound near a simplex vertex)**.** *Let $B > 0$, and consider a vector $\boldsymbol{t} \in \mathbb{R}^M$ satisfying $\max_{j,l\in[M]} |t_j - t_l| \le 2B$. For any $\boldsymbol{\beta} \in \Delta_{M-1}$, let $\boldsymbol{\beta}_{\boldsymbol{t}}$ denote its tilt defined by $\beta_{\boldsymbol{t},j} \propto \beta_j e^{t_j}$. Setting $\gamma := 1 - \max_j \beta_j$, we have*

$$D_{\mathrm{KL}}(\boldsymbol{\beta} \| \boldsymbol{\beta}_{\boldsymbol{t}}) \;\le\; \min\{\, B_0,\; A_0\,\gamma \,\}, \qquad B_0 := 2B, \quad A_0 := e^{2B} - 1 + 2B. \tag{92}$$

Proof. Let $\psi(\boldsymbol{\beta}) := \log\big(\sum_l \beta_l e^{t_l}\big) - \sum_j \beta_j t_j$. A direct computation from the definition of the tilt yields $D_{\mathrm{KL}}(\boldsymbol{\beta} \| \boldsymbol{\beta}_{\boldsymbol{t}}) = \psi(\boldsymbol{\beta})$. Because replacing $\boldsymbol{t}$ with $\boldsymbol{t} + s\mathbf{1}$ adds s to both terms, ψ is invariant under scalar shifts of $\boldsymbol{t}$. The bounding hypothesis on $\boldsymbol{t}$ is likewise shift-invariant.

For the first bound, we have $\log\big(\sum_l \beta_l e^{t_l}\big) \le \max_l t_l$ and $-\sum_j \beta_j t_j \le -\min_l t_l$. Combining these inequalities yields $\psi \le \max_l t_l - \min_l t_l \le 2B = B_0$.

For the second bound, let $j^* \in \arg\max_j \beta_j$. This implies $\beta_{j^*} = 1 - \gamma$ and $\sum_{l \ne j^*} \beta_l = \gamma$. By shift invariance, we may assume $t_{j^*} = 0$. The hypothesis then guarantees $|t_l| \le 2B$ for every l. Using the inequality $\log(1 + u) \le u$, we obtain

$$\log\Big(\sum_l \beta_l e^{t_l}\Big) \le \log\big((1-\gamma) + \gamma e^{2B}\big) = \log\big(1 + \gamma(e^{2B} - 1)\big) \le \gamma\,(e^{2B} - 1).$$

Also, the second term is bounded as $-\sum_j \beta_j t_j \le \sum_{l \ne j^*} \beta_l |t_l| \le 2B\gamma$. Adding these two bounds together gives $\psi \le \gamma\,(e^{2B} - 1 + 2B) = A_0\gamma$. □

Main result: M-mode blindness. Combining Lemmas E.2 and E.3 yields a uniform discrepancy bound over the $(M-1)$-dimensional reweighting region defined below.

Theorem E.1 (Mode blindness with M modes)**.** *Suppose the M-mode pinning assumption above holds with combined rate $c = \min\{c_1, \kappa\}$, and let $\Lambda := \{\boldsymbol{\lambda} \in \Delta_{M-1} : \lambda_j \ge c_0/2 \;\; \forall j\}$. There is a constant $C \le 12 c_0^{-2}$, depending only on c_0, such that every mask schedule μ with $v_{\min}(\mu) \ge s_{\mathrm{pin}}$ satisfies*

$$\sup_{\boldsymbol{\lambda}\in\Lambda} \mathfrak{D}_\mu(p \| q_{\boldsymbol{\lambda}}) \;\le\; C\,\mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\big] \;\le\; C e^{-c\,v_{\min}(\mu)}. \tag{93}$$

Consequently, for every $\varepsilon \ge C\,\mathbb{E}_{K\sim\mu}\big[e^{-c|K^{\complement}|}\big]$,

$$\mathfrak{A}^{\mu}_{\mathrm{TV}}(p, \varepsilon) \ge (M-1)\left(\frac{1}{M} - \frac{c_0}{2}\right) \ge \frac{M-1}{2M} \ge \frac{1}{4}. \tag{94}$$

Proof intuition. When the visible context identifies one mode with high confidence, bounded reweighting shifts the conditional distribution only through the small posterior mass on the remaining modes. Atypical contexts carry negligible probability under each mode. Averaging their probabilities under the true mode weights introduces no extra factor of M, establishing the uniform discrepancy bound.

Proof. Step 1: uniform constants. Since $w_j \ge c_0$ and $\lambda_j \ge c_0/2$, the log-ratios $t_j = \log(\lambda_j / w_j)$ satisfy $\max_{j,l} |t_j - t_l| \le 2B$ with $B := \log(2/c_0)$. Writing $y := e^{2B} = 4c_0^{-2}$, the constant

$$C := A_0 + B_0 = y - 1 + 2\log y \le 3(y-1) < 12 c_0^{-2}.$$

Step 2: fiberwise bound. Fix $\boldsymbol{\lambda} \in \Lambda$ and a mask $K \in \mathrm{supp}(\mu)$. Let $V := K^{\complement}$, ensuring the visible set satisfies

$$|V| \ge v_{\min}(\mu) \ge s_{\mathrm{pin}}.$$

By Lemma E.2, $\boldsymbol{\beta}_{\boldsymbol{\lambda},V}$ is the tilt of $\boldsymbol{\beta}_V$ by $\boldsymbol{t}$. Furthermore, the conditional divergence on the fiber over x_V equals $D_{\mathrm{KL}}(\boldsymbol{\beta}_V \| \boldsymbol{\beta}_{\boldsymbol{\lambda},V})$. Applying Lemma E.3 therefore guarantees that for every $x_V \in \mathrm{supp}(p_V)$,

$$D_{\mathrm{KL}}\big(\boldsymbol{\beta}_V(x_V) \| \boldsymbol{\beta}_{\boldsymbol{\lambda},V}(x_V)\big) \;\le\; \min\big\{B_0,\; A_0\,\gamma_V(x_V)\big\}. \tag{95}$$

Step 3: pooled confidence set. Define $G_V := \{x_V \in \mathrm{supp}(p_V) : \gamma_V(x_V) \le e^{-\kappa|V|}\}$. For $x_V \in \mathcal{E}_V^j$, the pinning assumption yields $1 - \beta_{V,j}(x_V) \le e^{-\kappa|V|}$. This implies $\gamma_V(x_V) \le e^{-\kappa|V|}$, meaning $x_V \in G_V$. We therefore have $\mathcal{E}_V^j \subseteq G_V$ for every j. Because $p_V = \sum_j w_j p_V^j$, we obtain

$$p_V(G_V^{\complement}) = \sum\nolimits_j w_j\, p_V^j(G_V^{\complement}) \le \sum\nolimits_j w_j\, p_V^j\big((\mathcal{E}_V^j)^{\complement}\big) \le e^{-c_1|V|},$$

which extends Lemma C.3 verbatim to M regions.

Step 4: averaging. We split the expectation of Eq. (95) over G_V and its complement. Applying the bound $\gamma_V \le e^{-\kappa|V|}$ on G_V alongside the uniform bound B_0 outside it yields

$$\mathbb{E}_{X_V\sim p_V}\big[D_{\mathrm{KL}}(\boldsymbol{\beta}_V \| \boldsymbol{\beta}_{\boldsymbol{\lambda},V})\big] \;\le\; A_0 e^{-\kappa|V|} + B_0\, p_V(G_V^{\complement}) \;\le\; (A_0 + B_0)\, e^{-c|V|}.$$

By Lemma E.2, the masked discrepancy $\mathfrak{D}_\mu(p\|q_{\boldsymbol{\lambda}})$ is the average of the left-hand side over $K \sim \mu$. Setting $C = A_0 + B_0$ and taking the supremum over $\boldsymbol{\lambda} \in \Lambda$ establishes Eq. (93). The final inequality relies on the minimum visible size constraint $|K^{\complement}| \ge v_{\min}(\mu)$.

Step 5: modulus lower bound. At the stated tolerance, every $q_{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda} \in \Lambda$ is admissible in the supremum defining $\mathfrak{A}_{\mathrm{TV}}^{\mu}(p,\varepsilon)$.

Choose $j^* \in \arg\min_j w_j$, so $w_{j^*} \le 1/M$, and define

$$\lambda_{j^*} := 1 - \frac{(M-1)c_0}{2}, \qquad \lambda_j := \frac{c_0}{2} \quad (j \ne j^*).$$

These weights sum to one and belong to Λ. For every $j \ne j^*$, we have $\lambda_j = c_0/2 \le w_j$. Hence Lemma E.1 gives

$$\begin{aligned} D_{\mathrm{TV}}(p, q_{\boldsymbol{\lambda}}) &= \sum_{j\ne j^*} \left(w_j - \frac{c_0}{2}\right) \\ &= 1 - w_{j^*} - \frac{(M-1)c_0}{2} \\ &\ge (M-1)\left(\frac{1}{M} - \frac{c_0}{2}\right) \\ &\ge \frac{M-1}{2M} \\ &\ge \frac{1}{4}, \end{aligned}$$

where the last two inequalities use $c_0 \le 1/M$ and $M \ge 2$. Since this distribution is admissible, the same lower bound holds for $\mathfrak{A}_{\mathrm{TV}}^{\mu}(p,\varepsilon)$. □

E.4 Directional Sensitivity Beyond Two Modes

For $M = 2$, the lower bound in Eq. (94) becomes $(1-c_0)/2$, matching Thm. 4.1. More generally, for every $M \ge 2$, Thm. E.1 establishes mode blindness over an $(M-1)$-dimensional reweighting region containing a distribution at total-variation distance at least $1/4$ from p, while every distribution in that region satisfies the stated discrepancy budget.

Nonetheless, the corresponding two-sided sensitivity characterization does not extend from Thm. 4.2 by simply replacing binary residual uncertainty with the one-hot MMSE. The binary case is special—after quotienting out the common log-weight shift, there is only one mode-contrast direction. For $M > 2$, sensitivity instead depends on direction in the mode-contrast subspace $\mathbf{1}^{\perp}$. Define the schedule-averaged posterior covariance

$$\Sigma_\mu := \mathbb{E}_{K\sim\mu}\mathbb{E}_p[\mathrm{Cov}_p(\mathbf{Z} \mid X_{K^{\complement}})]\,.$$

For a log-weight perturbation direction $\boldsymbol{h} \in \mathbb{R}^M$, consider the path $\boldsymbol{\lambda}(a)$ with components

$$\lambda_j(a) = \frac{w_j e^{a h_j}}{\sum_{\ell=1}^M w_\ell e^{a h_\ell}}, \qquad j \in [M].$$

By Lemma E.2, the fiberwise discrepancy is the KL divergence between a categorical posterior and its exponential tilt. A second-order expansion therefore gives

$$\mathfrak{D}_\mu\big(p\|q_{\boldsymbol{\lambda}(a)}\big) = \frac{a^2}{2}\,\boldsymbol{h}^\top \Sigma_\mu \boldsymbol{h} + o(a^2), \qquad a \to 0.$$

Moreover, the schedule-averaged residual uncertainty satisfies

$$\mathbb{E}_{K\sim\mu}\big[\mathrm{mmse}_p(\mathbf{Z} \mid X_{K^{\complement}})\big] = \mathrm{tr}(\Sigma_\mu).$$

Thus, the scalar MMSE controls the total posterior uncertainty, but not how that uncertainty is distributed across mode-reweighting directions. A positive trace need not imply positive sensitivity in every contrast direction. For example, suppose the scheduled visible contexts perfectly distinguish mode 3 from the group $\{1, 2\}$ but leave modes 1 and 2 unresolved. Then changing the weights from $(1/3, 1/3, 1/3)$ to $(1/4, 1/4, 1/2)$ preserves the relative weights of modes 1 and 2 and leaves all scheduled masked conditionals unchanged, despite positive residual mode uncertainty. The universal null direction $\mathbf{1}$ reflects only invariance to a common log-weight shift. Additional null or weak directions may remain after averaging over the schedule.

Consequently, a multimode analog of Thm. 4.2 requires quantitative directional control of Σ_μ on $\mathbf{1}^\perp$, rather than scalar residual uncertainty alone. This distinction does not affect the multimode blindness result above or the full-mask guarantee of Prop. 4.1.

F Low-Visibility Mass for Common Mask Schedules

This section computes $\pi_s(\mu) := \mu(|K^{\complement}| \le s)$ for several mask distributions. These masses enter the mode-weight recovery bound in Cor. 4.3. Tab. 2 summarizes their asymptotic rates for fixed nonnegative integers s as $N \to \infty$. All schedule parameters are independent of N.

Random masking probabilities. Absorbing-state diffusion uses independent coordinate masking conditional on a noise level (Austin et al., 2021; Shi et al., 2024). To describe the resulting mask patterns, let $t \in [0, 1]$ denote the masking probability and draw it from a continuous probability density g. Conditional on t, mask each coordinate independently with probability t. Then $|K^{\complement}| \mid t \sim \mathrm{Bin}(N, 1-t)$, and

$$\pi_s(\mu_{\mathrm{diff}}) = \int_0^1 g(t)\ \mathbb{P}(\mathrm{Bin}(N, 1-t) \le s)\ \mathrm{d}t. \tag{96}$$

Here g specifies the mask-sampling distribution. Substituting $u = 1 - t$ and expanding the binomial tail gives

$$\begin{aligned}\pi_s(\mu_{\mathrm{diff}}) &= \sum_{j=0}^{s} \binom{N}{j} \int_0^1 g(1-u)\, u^j (1-u)^{N-j}\ \mathrm{d}u \\ &= \frac{1}{N+1} \sum_{j=0}^{s} \mathbb{E}[g(1 - U_{N,j})]\,,\end{aligned}$$

where $U_{N,j} \sim \mathrm{Beta}(j+1, N-j+1)$. The second equality follows from the Beta density normalization. For fixed j, $\mathbb{E}[U_{N,j}] = (j+1)/(N+2)$, so $U_{N,j} \to 0$ in probability. Since g is bounded and continuous,

$$\mathbb{E}[g(1 - U_{N,j})] \longrightarrow g(1).$$

Consequently, if $g(1) > 0$,

$$\pi_s(\mu_{\mathrm{diff}}) \sim \frac{(s+1)g(1)}{N}.$$

For the uniform density $g \equiv 1$, the exact expression is $\pi_s(\mu_{\mathrm{diff}}) = (s+1)/(N+1)$. The N^{-1} scale comes from masking probabilities within order N^{-1} of full masking.

Table 2: Low-visibility mass for fixed s as $N \to \infty$, under the conditions specified in this section.

Mask schedule	$\pi_s(\mu)$
Continuous density g, with $g(1) > 0$	$\Theta(N^{-1})$
Continuous clipped density, positive at $t_{\max} < 1$	$\Theta(N^{s-1} t_{\max}^N)$
Independent masking at fixed $t_0 \in (0,1)$	$\Theta(N^s t_0^N)$
Visible count at least αN for every mask	0
Full-mask probability at least a fixed $\eta > 0$	$\Theta(1)$

Masking probabilities bounded away from one. Clipped ranges of masking probabilities are used in block diffusion to reduce training variance (Arriola et al., 2025). Consider independent coordinate masking conditional on a masking probability t, with $t \le t_{\max}$ almost surely for a fixed $0 < t_{\max} < 1$. Monotonicity of the binomial lower tail gives

$$\pi_s(\mu_{\text{diff}}^{t_{\max}}) \le \mathbb{P}(\text{Bin}(N, 1 - t_{\max}) \le s)\,.$$

For fixed s, the $j = s$ term dominates this tail, so

$$\mathbb{P}(\text{Bin}(N, 1 - t_{\max}) \le s) = \Theta\big(N^s t_{\max}^N\big)\,.$$

Thus

$$\pi_s(\mu_{\text{diff}}^{t_{\max}}) \le e^{-rN + O(\log N)}, \qquad r := \log(1/t_{\max}) > 0.$$

Suppose the masking probability has a density g_{clip} continuous on $[0, t_{\max}]$, with $g_{\text{clip}}(t_{\max}) > 0$. Expanding the binomial tail gives

$$\pi_s(\mu_{\text{diff}}^{t_{\max}}) = \sum_{j=0}^{s} \binom{N}{j} \int_0^{t_{\max}} g_{\text{clip}}(t)(1-t)^j t^{N-j}\ \mathrm{d}t.$$

For each fixed j, substituting $t = t_{\max} v$ yields

$$\int_0^{t_{\max}} g_{\text{clip}}(t)(1-t)^j t^{N-j}\ \mathrm{d}t = \frac{t_{\max}^{N-j+1}}{N-j+1} \mathbb{E}\big[g_{\text{clip}}(t_{\max} Y_{N,j})(1 - t_{\max} Y_{N,j})^j\big]\,,$$

where $Y_{N,j} \sim \text{Beta}(N-j+1, 1)$. Since $Y_{N,j} \to 1$ in probability, boundedness and continuity give

$$\mathbb{E}\big[g_{\text{clip}}(t_{\max} Y_{N,j})(1 - t_{\max} Y_{N,j})^j\big] \longrightarrow g_{\text{clip}}(t_{\max})(1 - t_{\max})^j.$$

Using $\binom{N}{j} \sim N^j / j!$, the j-th term therefore has order $N^{j-1} t_{\max}^N$. The $j = s$ term dominates, giving

$$\pi_s(\mu_{\text{diff}}^{t_{\max}}) \sim \frac{g_{\text{clip}}(t_{\max})(1 - t_{\max})^s t_{\max}^{1-s}}{s!} N^{s-1} t_{\max}^N.$$

In particular,

$$\pi_s(\mu_{\text{diff}}^{t_{\max}}) = \Theta(N^{s-1} t_{\max}^N) = e^{-rN + O(\log N)}.$$

The additional factor N^{-1} relative to masking at a fixed probability comes from averaging over probabilities near $t_{\max}$.

Independent masking at a fixed probability. If each coordinate is masked independently with a fixed probability $t_0 \in (0,1)$, then

$$\pi_s(\mu_{\text{ind}}) = \mathbb{P}(\text{Bin}(N, 1 - t_0) \le s) = \Theta\big(N^s t_0^N\big)\,.$$

Low-visibility masks therefore have exponentially small probability.

Exact-count and structured masking. Fixed-ratio masking is used in masked language modeling and masked autoencoding (Devlin et al., 2019; Liu et al., 2019; He et al., 2022). Consider the exact-count case, in which $\lfloor t_0 N \rfloor$ coordinates are masked for a fixed $t_0 \in [0, 1)$. Then

$$|K^{\complement}| = N - \lfloor t_0 N \rfloor \ge (1 - t_0) N.$$

More generally, suppose every admissible mask retains at least αN coordinates for some fixed $\alpha > 0$. Then

$$v_{\min}(\mu) \ge \alpha N, \qquad \pi_s(\mu) = 0 \quad \text{whenever } s < \alpha N.$$

This applies to exact-count masking and to structured object or block schedules that retain a fixed positive fraction of the coordinates, including object-masked prediction settings satisfying this condition (Nam et al., 2026). In particular, $\pi_s(\mu) = 0$ eventually for every sequence $s = o(N)$.

A full-mask component. If the schedule assigns probability at least $\eta > 0$ to the full mask, with η independent of N, then

$$\pi_s(\mu) \geq \eta \qquad \text{for every } s \geq 0.$$

Hence $\pi_s(\mu) = \Theta(1)$. Under joint masked-block log loss, this component also provides the joint-distribution control of Prop. 4.1.

G Interpretation and Instantiation of Assumptions

Asms. 4.1 and 4.2 describe how visible context informs the mode label. The first bounds residual mode uncertainty from above for large contexts; the second bounds it from below at low visibility. We interpret these conditions and construct a distribution satisfying both on an overlapping range of visible-set sizes.

G.1 Large-Context Mode Pinning

Statistical-mechanics motivation. At zero external field and below the critical temperature, the Curie–Weiss distribution concentrates near two opposite magnetizations (Ellis & Newman, 1978). Visible spins then provide evidence about the active phase. This motivates the mode-pinning condition: with high probability under either mode, a sufficiently large visible set assigns exponentially small posterior probability to the other mode. The construction in Appx. G.3 verifies this behavior through explicit finite-N bounds.

Natural-data interpretation. For natural data, a mode may represent a global attribute such as language, document format, or code-versus-prose structure. Evidence about such attributes can accumulate across a document, as illustrated by topic and mixture models (Blei et al., 2003; Hofmann, 1999). Asm. 4.1 makes this interpretation quantitative: the posterior concentration bound must hold with high probability under each mode, for every visible set of sufficient size.

Effect on masked prediction. When $x_V \in \mathcal{E}_V^\tau$, the conditional distribution of the masked variables assigns at most $e^{-\kappa|V|}$ probability to completions in the opposite mode. Reweighting the modes while preserving the within-mode distributions changes this conditional by at most a constant multiple of $e^{-\kappa|V|}$ in total variation, provided both reweighted mode probabilities remain bounded away from zero. This weak sensitivity arises because the visible context already resolves the mode, even though both modes have nonvanishing probability in the data.

G.2 Low-Visibility Uncertainty

Bounded evidence from small contexts. Asm. 4.2 requires $\mathrm{mmse}_p(Z \mid X_V) \geq u_0 e^{-L|V|}$ for $|V| \leq s_{\mathrm{ov}}$. This bound allows residual mode uncertainty to decrease with visibility. For fixed $|V|$ and constants independent of N, it prevents this uncertainty from vanishing as the dimension grows.

A sufficient condition is that the absolute log-likelihood ratio between the two mode-conditional visible distributions is bounded linearly in $|V|$ (Lemma C.6). Product distributions with bounded evidence per coordinate satisfy this condition directly. The construction below shows that the bound also survives conditioning the components onto disjoint mode regions.

Natural-data interpretation. Short contexts can remain compatible with several languages, topics, or document types. Studies of language identification and short-text classification report difficulties associated with limited context (Baldwin & Lui, 2010; Phan et al., 2008). These observations motivate the assumption; its formal requirement is a lower bound on the Bayes uncertainty about the mode. Under that condition, low-visibility masks preserve sensitivity to mode weights (Thm. 4.2).

G.3 An Explicit Distribution Satisfying Both Assumptions

We construct two components by conditioning biased product distributions on opposite signs of the total magnetization, and combine them with a true mixing weight $w \in (0, 1)$. Their visible log-

likelihood ratio separates into a product-distribution term and a conditioning correction. Bounds on this correction establish both assumptions.

Construction. Let $\mathcal{A} = \{-1, +1\}$, let $\mathcal{X} = \mathcal{A}^N$ with N odd, and fix $\theta, w \in (0, 1)$. For every nonempty $V \subseteq [N]$, define

$$S_V(x_V) := \sum\nolimits_{i \in V} x_i, \qquad m_V(x_V) := \frac{S_V(x_V)}{|V|},$$

and write $m(x) := m_{[N]}(x)$. Since N is odd, $m(x) \neq 0$ for every $x \in \mathcal{X}$. We use $\tau \in \{+, -\}$ to label the modes, interpreting these labels as $+1$ and -1 in arithmetic expressions. Define the partition

$$\mathcal{X}_\tau := \{x \in \mathcal{X} : \tau\, m(x) > 0\}, \qquad \tau \in \{+, -\}.$$

Let ν_τ be the product distribution under which the coordinates are i.i.d. with $\nu_\tau(X_i = \tau) = (1 + \theta)/2$, and set

$$p^\tau := \nu_\tau(\cdot \mid \mathcal{X}_\tau), \qquad p := wp^+ + (1 - w)p^-.$$

We use the notation

$$L_0 := \log \frac{1+\theta}{1-\theta}, \qquad \alpha := \frac{\theta}{2(1+\theta)}, \qquad \eta_N := 2e^{-\theta^2 N/8}, \qquad Z_\theta := \nu_+(\mathcal{X}_+).$$

The components have disjoint supports, and p has full support on $\mathcal{X}$. Reflection gives

$$p^-(x) = p^+(-x), \qquad \nu_-(\mathcal{X}_-) = Z_\theta,$$

while the mode probabilities are $p(\mathcal{X}_+) = w$ and $p(\mathcal{X}_-) = 1 - w$.

Lemma G.1 (Exact visible log-likelihood ratio)**.** *For every nonempty $V \subseteq [N]$ and every $x_V \in \mathcal{A}^V$,*

$$\log \frac{p_V^+(x_V)}{p_V^-(x_V)} = L_0 S_V(x_V) + T_V(x_V),$$

where

$$T_V(x_V) := \log \frac{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = x_V)}{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = -x_V)}.$$

Consequently,

$$\operatorname{logit} \beta_V(x_V) = \operatorname{logit}(w) + L_0 S_V(x_V) + T_V(x_V).$$

The identities are interpreted in the extended real sense when a mode marginal vanishes.

Proof. Marginalizing the conditioned distribution gives

$$p_V^+(x_V) = \frac{\nu_{+,V}(x_V)}{Z_\theta}\, \mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = x_V).$$

By reflection,

$$p_V^-(x_V) = \frac{\nu_{+,V}(-x_V)}{Z_\theta}\, \mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = -x_V).$$

The normalizers cancel, giving

$$\frac{p_V^+(x_V)}{p_V^-(x_V)} = \frac{\nu_{+,V}(x_V)}{\nu_{+,V}(-x_V)} \frac{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = x_V)}{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = -x_V)}.$$

The logarithm of the second factor is precisely the conditioning term

$$T_V(x_V) = \log \frac{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = x_V)}{\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = -x_V)}.$$

For the first factor, independence and $x_i \in \{-1, +1\}$ give

$$\begin{aligned} \frac{\nu_{+,V}(x_V)}{\nu_{+,V}(-x_V)} &= \prod_{i \in V} \frac{\nu_+(X_i = x_i)}{\nu_+(X_i = -x_i)} = \prod_{i \in V} \left(\frac{1+\theta}{1-\theta}\right)^{x_i} \\ &= \exp\left(\log \frac{1+\theta}{1-\theta} \sum_{i \in V} x_i\right) \\ &= e^{L_0 S_V(x_V)}. \end{aligned}$$

Taking logarithms of the factorization therefore yields

$$\log \frac{p_V^+(x_V)}{p_V^-(x_V)} = L_0 S_V(x_V) + T_V(x_V).$$

To obtain the posterior-logit identity, recall that $\beta_V(x_V) = p(\mathcal{X}_+ \mid X_V = x_V)$. Since the mixture components are supported on their respective mode regions, Bayes' rule gives

$$\beta_V(x_V) = \frac{w\, p_V^+(x_V)}{w\, p_V^+(x_V) + (1-w) p_V^-(x_V)}, \qquad 1 - \beta_V(x_V) = \frac{(1-w) p_V^-(x_V)}{w\, p_V^+(x_V) + (1-w) p_V^-(x_V)}.$$

Dividing these expressions and taking logarithms,

$$\begin{aligned} \operatorname{logit} \beta_V(x_V) &= \log \frac{\beta_V(x_V)}{1-\beta_V(x_V)} = \log \frac{w}{1-w} + \log \frac{p_V^+(x_V)}{p_V^-(x_V)} \\ &= \operatorname{logit}(w) + L_0 S_V(x_V) + T_V(x_V). \end{aligned}$$

When either mode marginal vanishes, the corresponding identities hold under the stated extended-real convention. □

The term $T_V(x_V)$ accounts for conditioning ν_τ on $\mathcal{X}_\tau$ when constructing p^τ. For mode pinning, we use its sign on contexts whose magnetization favors the corresponding mode. For low-visibility uncertainty, we bound its magnitude uniformly over small visible sets.

Lemma G.2 (Bounds on the conditioning term)**.** *In the setting above, for every nonempty $V \subseteq [N]$, the following hold:*

(i) If $S_V(x_V) \geq 0$, then $T_V(x_V) \geq 0$. Moreover,

$$T_V(-x_V) = -T_V(x_V).$$

(ii) If $|V| \leq \alpha N$ and $\eta_N \leq 1$, then

$$|T_V(x_V)| \leq \eta_N \qquad \text{for every} \quad x_V \in \mathcal{A}^V.$$

Proof. Under ν_+, let

$$M := \sum_{i \notin V} X_i.$$

This sum is independent of X_V and has mean $\theta(N - |V|)$. Since $\mathcal{X}_+ = \{S_V + M > 0\}$,

$$\mathbb{P}_{\nu_+}(\mathcal{X}_+ \mid X_V = x_V) = \mathbb{P}(M > -S_V(x_V)).$$

Part (i). If $S_V(x_V) \geq 0$, then

$$\mathbb{P}(M > -S_V(x_V)) \geq \mathbb{P}(M > S_V(x_V)),$$

which gives $T_V(x_V) \geq 0$. Replacing x_V by $-x_V$ interchanges the numerator and denominator in the definition of T_V, proving the reflection identity.

Part (ii). For $|V| \leq \alpha N$, the distance from the mean of M to the threshold $-S_V(x_V)$ satisfies

$$\theta(N - |V|) + S_V(x_V) \geq \theta N - (1+\theta)|V| \geq \frac{\theta N}{2}.$$

Hoeffding's inequality therefore gives

$$\mathbb{P}(M \leq -S_V(x_V)) \leq \exp\left(-\frac{(\theta N/2)^2}{2(N - |V|)}\right) \leq e^{-\theta^2 N/8}.$$

The same bound holds with x_V replaced by $-x_V$. Both probabilities in the ratio defining T_V thus lie in $[1 - e^{-\theta^2 N/8}, 1]$. Using $-\log(1-u) \leq 2u$ for $0 \leq u \leq 1/2$ yields $|T_V(x_V)| \leq \eta_N$. □

Proposition G.1 (Compatibility of mode pinning and low-visibility uncertainty)**.** *Let p be the distribution constructed above and set*

$$c_0 := \min\{w, 1-w\}, \qquad \kappa := \frac{L_0\theta}{4}, \qquad c_1 := \frac{\theta^2}{16}, \qquad u_0 := w(1-w),$$
$$s_{\mathrm{pin}} := \left\lceil \max\left\{ \frac{16\ln 2}{\theta^2}, \frac{4|\operatorname{logit}(w)|}{L_0\theta} \right\} \right\rceil, \qquad L := L_0 + 1, \qquad s_{\mathrm{ov}} := \lfloor \alpha N \rfloor.$$

There exists $N_0 = N_0(\theta, w)$ such that, for every odd $N \geq N_0$:

(i) Asm. 4.1 holds with constants $c_0, \kappa, c_1, s_{\mathrm{pin}}$.

(ii) Asm. 4.2 holds with $u_0 = w(1-w)$, rate L, and scale s_{ov}.

(iii) The combined rate is $c = \min\{c_1, \kappa\} = \theta^2/16 < L_0 < L$, and the visibility ranges overlap: $s_{\mathrm{pin}} \leq s_{\mathrm{ov}}$.

Proof intuition. With sufficiently many visible coordinates, magnetization concentrates in the direction favored by the active mode. The resulting likelihood ratio easily overcomes the prior log-odds, pinning the mode posterior. For small visible sets, the conditioning correction is uniformly small, so the log-likelihood ratio is bounded in magnitude by a constant multiple of the visible count. This limits how quickly posterior uncertainty can decay, establishing the MMSE lower bound.

Proof. Choose $N_0 = N_0(\theta, w)$ sufficiently large that $\eta_N \leq 1$ and $\alpha N \geq s_{\mathrm{pin}}$ for every $N \geq N_0$.

Part (i): mode pinning. For $|V| \geq s_{\mathrm{pin}}$ and $\tau \in \{+, -\}$, define

$$G_V^\tau := \{x_V \in \operatorname{supp}(p_V^\tau) : \tau\, m_V(x_V) \geq \theta/2\}.$$

For $x_V \in G_V^\tau$, Lemma G.2(i) gives $\tau T_V(x_V) \geq 0$. Hence

$$\begin{aligned}\tau \operatorname{logit} \beta_V(x_V) &= \tau \operatorname{logit}(w) + L_0 \tau S_V(x_V) + \tau T_V(x_V) \\ &\geq -|\operatorname{logit}(w)| + \frac{L_0\theta}{2}|V| \\ &\geq \frac{L_0\theta}{4}|V| = \kappa|V|.\end{aligned}$$

Here, the last inequality uses $|V| \geq s_{\mathrm{pin}} \geq 4|\operatorname{logit}(w)|/(L_0\theta)$. Therefore

$$p(\mathcal{X}_{-\tau} \mid X_V = x_V) = \frac{1}{1 + \exp\big(\tau \operatorname{logit} \beta_V(x_V)\big)} \leq e^{-\kappa|V|},$$

so $G_V^\tau \subseteq \mathcal{E}_V^\tau$ for both modes. For every visible event A,

$$p_V^\tau(A) \leq \frac{\nu_{\tau,V}(A)}{Z_\theta}.$$

Under $\nu_{\tau,V}$, the variables τX_i are independent with mean θ, so

$$\nu_{\tau,V}(\tau m_V < \theta/2) \leq e^{-\theta^2|V|/8}.$$

Also,

$$Z_\theta = \mathbb{P}_{\nu_+}\left(\sum_i X_i > 0\right) \geq 1 - e^{-\theta^2 N/2} \geq \tfrac{1}{2}.$$

It follows that

$$p_V^\tau\big((G_V^\tau)^{\mathsf{c}}\big) \leq 2e^{-\theta^2|V|/8} \leq e^{-\theta^2|V|/16} = e^{-c_1|V|},$$

where the second inequality uses $|V| \geq 16\ln 2/\theta^2$. Thus $p_V^\tau(\mathcal{E}_V^\tau) \geq 1 - e^{-c_1|V|}$. Since $p(\mathcal{X}_\tau) \geq c_0$ for both modes, Asm. 4.1 holds.

Part (ii): low-visibility uncertainty. For $V = \varnothing$,

$$\mathrm{mmse}_p(Z \mid X_V) = \mathrm{Var}_p(Z) = w(1-w).$$

For $1 \le |V| \le s_{\mathrm{ov}}$, Lemmas G.1 and G.2(ii) give

$$\left|\log \frac{p_V^+(x_V)}{p_V^-(x_V)}\right| \le L_0|V| + \eta_N \le (L_0+1)|V| = L|V|, \tag{97}$$

where $\eta_N \le 1$ for $N \ge N_0$. Applying Lemma C.6 yields

$$\mathrm{mmse}_p(Z \mid X_V) \ge w(1-w)e^{-L|V|}.$$

Part (iii): rates and overlap. Since $L_0 = 2\,\mathrm{artanh}(\theta) \ge 2\theta$, we have $\kappa \ge \theta^2/2 > c_1$. Therefore

$$c = c_1 = \frac{\theta^2}{16} < L_0 < L.$$

The choice of N_0 ensures $\alpha N \ge s_{\mathrm{pin}}$. Since s_{pin} is an integer, this gives

$$s_{\mathrm{pin}} \le \lfloor \alpha N \rfloor = s_{\mathrm{ov}}.$$

□

Remark G.1 (Compatible bounds and a reference decay rate). On the overlap window, the two assumptions give

$$w(1-w)e^{-L|V|} \le \mathrm{mmse}_p(Z \mid X_V) \le \frac{5}{4}e^{-c|V|}, \qquad s_{\mathrm{pin}} \le |V| \le s_{\mathrm{ov}},$$

with $c < L$. The mixing weight enters the lower-bound prefactor and the pinning threshold, while the rates c and L depend only on θ. For comparison, nearly balanced visible configurations under the unconditioned product distributions have the large-deviation rate

$$r^\star(\theta) := \mathrm{kl}\big(\tfrac{1}{2} \,\big\|\, \tfrac{1+\theta}{2}\big) = -\tfrac{1}{2}\log(1-\theta^2).$$

This gives the reference rate used in the experiments. It satisfies $c = \frac{\theta^2}{16} < r^\star(\theta) < L_0 < L$.

H Experimental Protocols and Additional Results

This section evaluates the quantitative consequences of mode-weight sensitivity while separating population geometry from optimization. Appx. H.2 computes discrepancy profiles and residual uncertainty for finite reweightings, examining the blindness and sensitivity predictions of Thms. 4.1 and 4.2. Appx. H.3 varies low-visibility mass and measures the resulting largest admissible mode-weight error, testing the scaling in Cor. 4.3. Appx. H.4 optimizes the same coherent family to examine how the measured curvature predicts recovery under a fixed gradient-descent protocol. Appx. H.5 expresses the excess-loss signal in sampling units. Throughout the controlled experiments, the within-mode distributions are known and fixed, so changes in the mode-weight signal are separated from component-estimation error and conditional incompatibility. Appx. H.6 provides a complementary, exploratory measurement of label uncertainty in natural text. Tab. 3 summarizes the experimental settings.

H.1 Controlled Distributions, Estimands, and Computation

Conditioned-product mixture. For $\tau \in \{-1,+1\}$, let ν_τ be the product distribution on $\{-1,+1\}^N$ satisfying $\nu_\tau(X_i = \tau) = (1+\theta)/2$, and define

$$\mathcal{X}_\tau = \{x : \tau \sum\nolimits_i x_i > 0\}, \qquad p^\tau = \nu_\tau(\cdot \mid \mathcal{X}_\tau).$$

The data distribution is $p = wp^+ + (1-w)p^-$, with $w \in (0,1)$ and binary mode indicator $Z := \mathbb{1}\{X \in \mathcal{X}_+\}$. Setting $c_0 := \min\{w, 1-w\}$, we vary only the global mode weight over a compact interval:

$$q_\lambda = \lambda p^+ + (1-\lambda)p^-, \qquad \lambda \in I := [c_0/2,\, 1 - c_0/2]\,.$$

The within-mode distributions are identical under p and q_λ, and, for each λ, every masked conditional is induced by the same joint distribution q_λ. Because the components have disjoint sign-of-magnetization supports, $D_{\mathrm{TV}}(p, q_\lambda) = |w - \lambda|$. This separates an interpretable joint-distribution error from the visibility-dependent signal seen by the masked objective.

The main experiments use $\theta = 0.8$ and $w = 1/2$, giving $I = [1/4, 3/4]$. Additional robustness evaluations cover $\theta \in \{0.6, 0.8\}$ and $w \in \{1/2, 0.9\}$; for $w = 0.9$, the corresponding interval is $I = [0.05, 0.95]$. Direct optimization uses the main setting $(\theta, w) = (0.8, 1/2)$.

Curie–Weiss model. We use the zero-field Curie–Weiss model (Ellis & Newman, 1978),

$$p_{\beta,N}(x) \propto \exp\left(\frac{\beta}{2N}\left(\sum_{i=1}^{N} x_i\right)^2\right), \qquad x \in \{-1, +1\}^N,$$

where β is the inverse temperature. We include this model to distinguish the effect of macroscopic mode structure from artifacts of the conditioned-product construction. At $\beta = 1.5$, the spins interact and the magnetization has two macroscopic modes, providing an interacting test of whether mode blindness persists beyond the conditioned-product mixture we constructed. At $\beta = 0.5$, the model is in the high-temperature unimodal regime and serves as a null control without comparable large-context mode pinning. Both settings use the same sign-of-magnetization partition and mode-reweighting construction defined above. Note that the closed-form reference rate of Remark G.1 applies only to the conditioned-product mixture. Any Curie–Weiss reference obtained by substituting its spontaneous magnetization is used only as a heuristic visual guide.

Estimands. For the MMSE, we encode the positive and negative mode labels as 1 and 0, respectively. Since p and q_λ are exchangeable, the quantities below depend on the visible set $V = K^{\complement}$ only through its size $m = |V|$. We define

$$\mathfrak{D}_{N,m}(\lambda) := \mathfrak{D}_K(p\|q_\lambda), \qquad \Delta_{N,m} := \sup_{\lambda \in I} \mathfrak{D}_{N,m}(\lambda), \qquad U_{N,m} := \mathrm{mmse}_p(Z \mid X_V). \quad (98)$$

Here $\Delta_{N,m}$ is the largest masked discrepancy over the allowed mode weights, and $U_{N,m}$ is the residual uncertainty about the mode after observing m coordinates. The maximum defining $\Delta_{N,m}$ is attained at an endpoint of I: the discrepancy is nonincreasing for $\lambda \le w$ and nondecreasing for $\lambda \ge w$. For $U_{N,m} > 0$ and $\lambda \ne w$, we also define

$$\rho_{N,m}(\lambda) := \frac{\mathfrak{D}_{N,m}(\lambda)}{U_{N,m}\mathrm{kl}(w\|\lambda)}. \quad (99)$$

This ratio compares the discrepancy with the sensitivity scale $U_{N,m}\mathrm{kl}(w\|\lambda)$ in Lemma 4.1.

For both controlled model families, the mode posteriors depend on the visible configuration only via its magnetization $r = 2j - m$, where j is the number of visible $+1$ spins. Write $\beta_V = p(Z = 1 \mid X_V)$ and $\beta_{\lambda,V} = q_\lambda(Z = 1 \mid X_V)$. Conditional on X_V, the positive- and negative-mode distributions of the masked block have disjoint supports and are identical under p and q_λ; only their posterior mixing weights differ. Consequently, the conditional block KL divergence is exactly $\mathrm{kl}(\beta_V\|\beta_{\lambda,V})$. Each expectation therefore reduces to a sum over the $m+1$ possible visible magnetizations, weighted by their probabilities under p. We evaluate these sums using log-domain arithmetic.

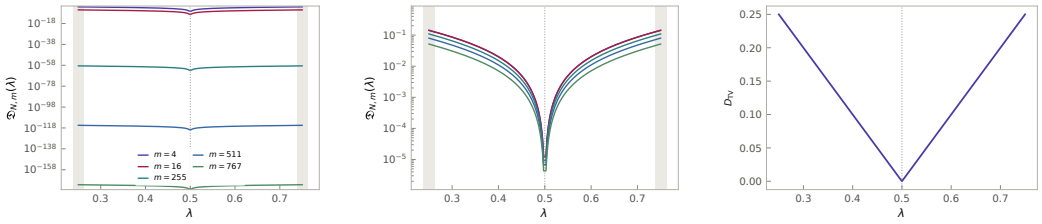
Numerical validation. The log-domain implementation and a 50-digit reference differ by at most 8.2×10^{-14} in $\log \mathfrak{D}$ on representative cells. Exhaustive enumeration at small N also checks the posterior tilt, risk identity, gradient, curvature, schedule linearity, and full-mask endpoint.

H.2 Exact Computation of Mode-Blindness Geometry

Discrepancy profiles. Fig. 3a shows the conditioned-product profiles at $\theta = 0.8$, $w = 1/2$, and $N = 1023$. With little visible context, uncertainty about Z makes the masked conditional sensitive to the mixture weight. As visibility increases, the context identifies the mode more reliably, so changing its prior weight has less effect on conditional prediction. The discrepancy profiles consequently flatten across the reweighting interval. The Curie–Weiss control at $\beta = 0.5$ uses the same sign-of-magnetization partition and reweighting construction (Fig. 3b). Its magnetization distribution is unimodal, so the sign partition does not correspond to two separated macroscopic modes. Its

Table 3: Experimental settings and evaluation protocols.

Component	Settings
Exact population geometry	Conditioned-product mixture: $\theta \in \{0.6, 0.8\}$. Curie–Weiss model: $\beta \in \{1.5, 0.5\}$. Sequence lengths $N \in \{127, 255, 511, 1023\}$ and mode weights $w \in \{1/2, 9/10\}$. Visible sizes $m = \lvert K^{\complement}\rvert$ include $\{0, 1, 2, 4\}$, dyadic sizes, and $\lfloor \varrho N \rfloor$ for $\varrho \in \{1/4, 1/2, 3/4\}$. Exact summation over visible magnetization.
Schedule intervention	Two-point visible-size distribution $(1-\pi)\delta_{m_{\mathrm{base}}} + \pi\delta_s$, with $m_{\mathrm{base}} \in \{\lfloor N/4 \rfloor, \lfloor N/2 \rfloor\}$, $s \in \{0, 1, 2, 4\}$, and $\pi \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. Evaluation at four excess-risk budgets.
Optimization within q_λ	Conditioned-product mixture with $\theta = 0.8$ and $w = 1/2$. Calibration over $N \in \{31, 63, 127\}$ selects $N = 63$ and $m_{\mathrm{base}} = 31$. Learning rate $\eta = 0.1$ and a budget of 10^5 steps. Stochastic runs use batch size 512 and five seeds.
Sampling cost	Conditioned-product mixture with $N = 1023$, $m_{\mathrm{base}} = 511$, and $\lambda = 3/4$. Exact first and second moments of the per-example excess loss.
Natural-text exploration	Code versus prose (`github-code-clean` Python (CodeParrot, 2022) against C4 `en` (Raffel et al., 2020)) and German versus English (OPUS-100 `de-en` (Tiedemann, 2012; Zhang et al., 2020)). Disjoint sets of packed sequences for fitting, calibration, and evaluation; GPT-2 tokenization (Radford et al., 2019). A short-context posterior estimator and calibrated long-context classifiers. Uncertainty estimated using 1,000 bootstrap replicates obtained by resampling packed sequences.



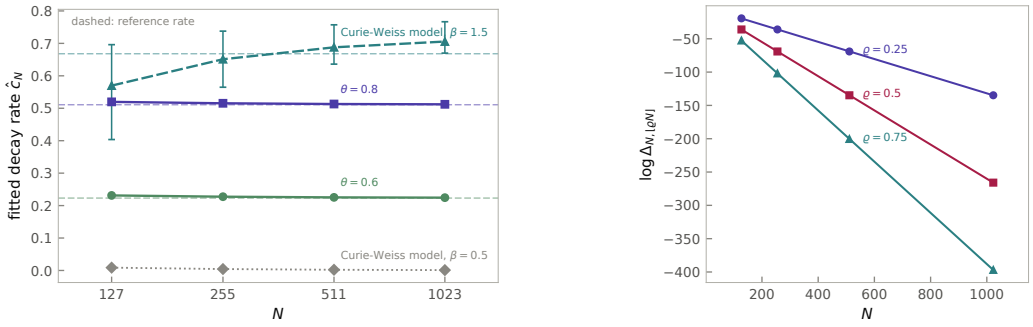
(a) Conditioned-product mixture. (b) Curie–Weiss null control. (c) Joint total-variation error.

Figure 3: Mode-reweighting geometry at $N = 1023$ and $w = 1/2$. (a–b) Discrepancy profiles $\lambda \mapsto \mathfrak{D}_{N,m}(\lambda)$ at different visible sizes m for the conditioned-product mixture ($\theta = 0.8$) and Curie–Weiss control ($\beta = 0.5$). Shaded strips mark the endpoints of $I = [1/4, 3/4]$. (c) Joint TV error, common to both constructions and all visible sizes.

profiles retain appreciable curvature and show much weaker suppression over the displayed visibility range. For both constructions, the disjoint component supports give $D_{\mathrm{TV}}(p, q_\lambda) = |w - \lambda|$, independently of visibility (Fig. 3c). Increasing visibility can therefore suppress the masked discrepancy while leaving the joint-distribution error unchanged.

Decay rates. At $\theta = 0.8$ and $N = 1023$, the maximal discrepancy $\Delta_{N,m}$ decreases from 1.44×10^{-1} at $m = 0$ to 4.33×10^{-173} at $m = 767$, a reduction of 171.52 orders of magnitude. Across the evaluated system sizes, the fitted conditioned-product rates $\hat{c}_N$ approach the closed-form reference

$$r^\star(\theta) = -\frac{1}{2}\log(1 - \theta^2)$$



(a) Fitted and reference decay rates.

(b) Decay at fixed visible fractions.

Figure 4: Decay rates across system sizes. (a) Fitted rates for the conditioned-product mixture and both Curie–Weiss settings. The conditioned-product reference is given in closed form; the interacting Curie–Weiss reference is a heuristic visual guide. (b) Log discrepancy $\log \Delta_{N,\lfloor \varrho N \rfloor}$ versus N at fixed visible fractions $\varrho \in \{1/4, 1/2, 3/4\}$.

from above (Fig. 4a). At $N = 1023$, the relative errors are 0.27% for $\theta = 0.8$ and 0.58% for $\theta = 0.6$. This comparison concerns the reference rate $r^\star(\theta)$ in Remark G.1, rather than the sharpness of the conservative mode-pinning exponent $c = \theta^2/16$ certified by Prop. G.1. The theorem establishes exponential suppression. Agreement with $r^\star(\theta)$ is an additional numerical observation for this construction. The unimodal control's fitted rate instead decreases with N. At each fixed visible fraction $\varrho \in \{1/4, 1/2, 3/4\}$, $\log \Delta_{N,\lfloor \varrho N \rfloor}$ decreases approximately linearly with N (Fig. 4b), consistent with the predicted exponential suppression in system size.

Residual uncertainty governs reweighting sensitivity. The normalized ratio $\rho_{N,m}(\lambda)$ in Eq. (99) separates the effects of mode-weight divergence and residual uncertainty. Lemma 4.1 bounds this ratio uniformly over a fixed weight interval, whereas Eq. (24) identifies its limiting value only as $\lambda \to w$. For $w = 1/2$, across the 98 evaluated (N, m) settings—49 for each value of θ—and all retained λ-grid points with $|\lambda - w| \geq 0.01$, we obtain $\rho_{N,m}(\lambda) \in [4.000, 4.408]$. Thus, the normalization remains informative across finite reweightings, beyond the local limit $1/(w(1-w)) = 4$. The asymmetric setting $\theta = 0.8$, $w = 0.9$ gives ratios in $[6.550, 13.825]$, with local reference $1/(w(1-w)) \simeq 11.1$. Together, these evaluations show that residual uncertainty accounts for the large visibility-dependent changes in discrepancy in both symmetric and asymmetric mixtures.

H.3 Low-Visibility Schedule Intervention

Mask schedule and estimands. We fix the sequence length at $N = 1023$. Let δ_m denote a schedule supported on masks leaving exactly m coordinates visible. For $s < m_{\text{base}}$, we evaluate

$$\mu = (1 - \pi)\delta_{m_{\text{base}}} + \pi\delta_s, \qquad \pi_s(\mu) = \pi. \tag{100}$$

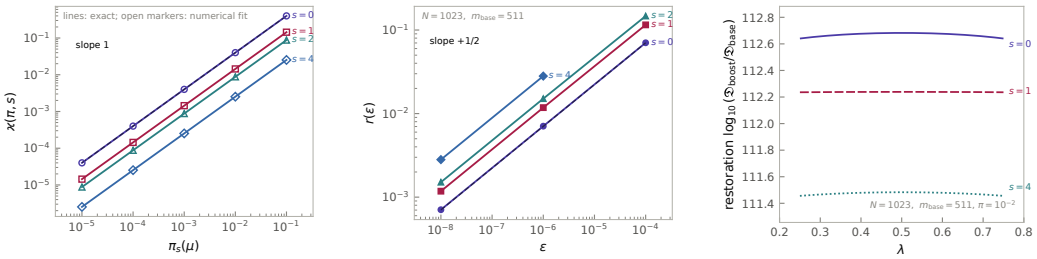
Here, since $s < m_{\text{base}}$, the schedule's low-visibility mass is exactly $\pi_s(\mu) = \pi$. A fraction π of examples therefore leave s coordinates visible, while the remainder use the baseline visibility m_{base}. We test $s \in \{0, 1, 2, 4\}$, including both the full mask ($s = 0$) and non-full masks.

We compute the curvature with respect to λ and the admissible mode-weight recovery radius:

$$\varkappa(\pi, s) := \partial_\lambda^2 \mathfrak{D}_\mu(p\|q_\lambda)\big|_{\lambda=w} = \frac{\mathbb{E}_{K\sim\mu}\left[\text{mmse}_p(Z \mid X_{K^c})\right]}{\big(w(1-w)\big)^2}, \tag{101}$$

$$r(\varepsilon) := \sup\Big\{|w - \lambda| : \lambda \in I,\ \mathfrak{D}_\mu(p\|q_\lambda) \leq \varepsilon\Big\}. \tag{102}$$

The discrepancy and curvature are exact weighted averages of their fixed-visibility values. The radius $r(\varepsilon)$ is evaluated from the discrepancy constraint in Eq. (102) and measures admissible errors within the tested reweighting interval I. Hollow markers indicate that this interval, rather than the risk constraint, limits the reported radius. In the symmetric setting, these points have $r(\varepsilon) = 1/4$ because every tested weight satisfies the tolerance; larger reweightings outside I are not evaluated. The resulting plateau is therefore an interval-saturation effect, not an optimization outcome.



(a) Curvature vs. $\pi_s(\mu)$. (b) Recovery radius vs. tolerance. (c) Discrepancy under intervention.

Figure 5: Low-visibility interventions at sequence length $N = 1023$ and the baseline number of visible coordinates $m_{\text{base}} = 511$, with $s \in \{0, 1, 2, 4\}$. (a) Curvature versus intervention mass π; lines show exact values and open markers show numerical fits. (b) Recovery radius versus tolerance at $\pi = 10^{-2}$. (c) Discrepancy profiles at $\pi = 10^{-2}$ compared with the blind baseline.

Recovery scaling. When the low-visibility contribution dominates the baseline, the curvature has a fitted log–log slope of 1.0000 against π (Fig. 5a), consistent with linear schedule averaging. For fixed s and sufficiently small tolerances with the interval boundary inactive, the local quadratic approximation gives

$$r(\varepsilon) \simeq \sqrt{\frac{2\varepsilon}{\varkappa(\pi, s)}} \asymp \sqrt{\frac{\varepsilon}{\pi}}.$$

Both dependencies appear in the exact evaluations. At $\pi = 10^{-2}$, the radius has a fitted log–log slope of 0.4995 against ε (Fig. 5b). At $\varepsilon = 10^{-4}$, its slope against full-mask mass $\pi_0(\mu)$ is -0.4894 (Fig. 2b) and approaches $-1/2$ as the tolerance decreases.

Restored mode-weight sensitivity. The $s = 1$ and $s = 4$ interventions strengthen the discrepancy without assigning any probability to the full mask, directly illustrating directional recovery from nonempty visible contexts. The $s = 0$ intervention supplies the joint-KL endpoint: at $\pi_0(\mu) = 10^{-2}$, it increases the nonzero reweighting discrepancy by approximately 4.8×10^{112} relative to the blind baseline (Fig. 5c). Throughout these comparisons, the within-mode distributions remain fixed, and the measured radius concerns recovery along q_λ. For $s = 0$, Prop. 4.1 supplies the separate, stronger guarantee over all admissible joint distributions.

H.4 Direct Optimization within q_λ

Likelihood and curvature. We optimize only $a \in \mathbb{R}$, with $\lambda = \text{logit}^{-1}(a)$ ranging over $(0, 1)$, while keeping the within-mode distributions fixed. The binary label $Z = \mathbb{1}\{X \in \mathcal{X}^+\}$ is determined by the same complete sample X used for masked prediction; it is not an externally supplied annotation. For $V = K^{\complement}$, the shared-component structure gives

$$-\log q_{\lambda,K}(X_K \mid X_V) = -Z \log \beta_{\lambda,V}(X_V) - (1 - Z) \log\big(1 - \beta_{\lambda,V}(X_V)\big) - \log p(X_K \mid X_V, Z).$$

The final term is independent of a. Consequently, optimizing mode-label cross-entropy gives the same parameter gradients as the joint masked-block negative log-likelihood. The visible context still enters through $\beta_{\lambda,V}$, obtained by the posterior tilt in Lemma 3.1. The experiment does not replace conditional prediction by unconditional fitting of mode frequencies.

Let $R_\mu(a)$ denote the population masked-prediction risk of q_λ, and let $a^\star = \text{logit}(w)$ denote the true parameter. The masked discrepancy is the excess risk above $R_\mu(a^\star)$:

$$\mathfrak{D}_\mu(p \| q_\lambda) = R_\mu(a) - R_\mu(a^\star) = \mathbb{E}_{K \sim \mu,\, X \sim p} \text{kl}\big(\beta_{K^{\complement}} \| \beta_{\lambda,K^{\complement}}\big)\,. \tag{103}$$

The curvature at the truth is therefore $H_\mu := \partial_a^2 R_\mu(a)\big|_{a=a^\star} = \mathbb{E}_{K \sim \mu}\left[\text{mmse}_p(Z \mid X_{K^{\complement}})\right]$. Near $a^\star$, population gradient descent with learning rate $\eta > 0$ contracts the logit error by approximately $1 - \eta H_\mu$ per step. For $\eta H_\mu \ll 1$, the corresponding half-time is $t_{1/2} \simeq \log 2/(\eta H_\mu)$. Since $\lambda - w \simeq w(1 - w)(a - a^\star)$ locally, this also predicts the time to halve the mode-weight error. At each checkpoint, we record $\widehat{\lambda}_t = \text{logit}^{-1}(a_t)$ and the excess risk.

Table 4: Error-halving times in steps at $N = 63$, $\lambda_0 = 0.4$, and $\eta = 0.1$. Predictions use $\log 2/(\eta H_\mu)$. SGD entries report medians over five seeds, with bootstrap intervals in brackets. Entries marked $> 10^5$ indicate that the half-error criterion was not met within the step budget.

Schedule	H_μ	**Predicted**	**Population GD**	**SGD** median [interval]
Blind	6.23×10^{-9}	1.11×10^{9}	$> 10^5$	$> 10^5$
$s = 1,\ \pi = 3\times10^{-3}$	2.70×10^{-4}	2.57×10^{4}	2.59×10^{4}	$2.68\,[2.47, 3.06]\times10^{4}$
$s = 0,\ \pi = 3\times10^{-2}$	7.50×10^{-3}	9.24×10^{2}	9.45×10^{2}	$1.00\,[1.00, 1.10]\times10^{3}$
$s = 1,\ \pi = 10^{-1}$	9.00×10^{-3}	7.70×10^{2}	7.76×10^{2}	$9.00\,[8.00, 9.00]\times10^{2}$

Calibration and optimizer. Before running gradient descent, we perform an exact population calibration over $N \in \{31, 63, 127\}$. For each N, we set $m_{\text{base}} = \lfloor N/2 \rfloor$ and evaluate the same two-point schedule family used in the low-visibility intervention. We use $s \in \{0, 1, 4\}$ and the schedule-mass grid $\pi \in \{0, 10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}, 10^{-1}\}$, where $\pi_s(\mu) = \pi$ is the probability of leaving s coordinates visible. For every calibrated setting, we compute the exact population residual uncertainty, risk gradients, curvature, and risk profile, together with the resulting local prediction of the optimization half-time. The case $\pi = 0$ is the blind baseline $\delta_{m_{\text{base}}}$ and is independent of s.

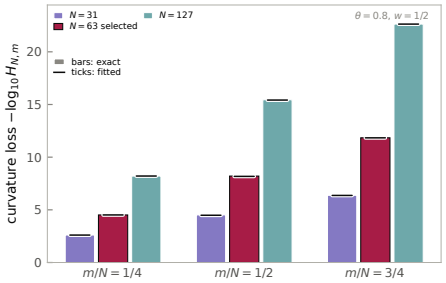


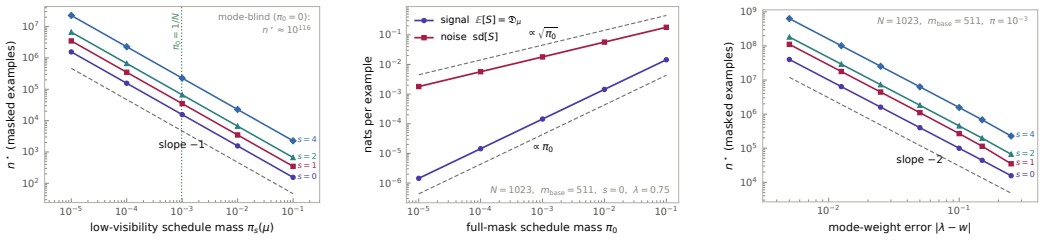
Figure 6: Curvature $H_{N,m} = U_{N,m}$ across visible fractions and system sizes, shown as $-\log_{10} H_{N,m}$. Filled bars give exact values; horizontal ticks show log-linear fits over the middle half of the visibility range. Outlined bars identify $N = 63$, selected for optimization before inspecting trajectories.

Before inspecting any optimization trajectory, a predeclared rule selects the largest numerically resolvable dimension for which the blind predicted half-time is at least ten times the 10^5-step budget. The rule also requires at least three $s = 1$ interventions with predicted half-times between 500 and 80,000 steps and spanning at least a factor of ten in curvature. It selects $N = 63$, with $m_{\text{base}} = 31$ visible coordinates under the blind baseline (Fig. 6).

We choose the learning rate from $\{0.025, 0.05, 0.1\}$ using fixed-seed, 1000-step pilot runs on the globally highest-curvature configuration, initialized at $\lambda_0 \in \{0.1, 0.9\}$. We retain the largest rate for which all pilot runs remain finite, the logit parameter satisfies $|a| \le 20$, and the exact excess risk does not exceed the predeclared 25% risk-growth bound. The resulting rate, $\eta = 0.1$, is fixed across all reported runs. Population gradient descent uses $\lambda_0 \in \{0.1, 0.3, 0.4, 0.6, 0.7, 0.9\}$. Minibatch SGD uses $\lambda_0 \in \{0.4, 0.6\}$, batch size 512, and five independent seeds. Each run has a maximum budget of 10^5 steps, with evaluation checkpoints every 100 steps. A stochastic crossing is accepted only when three consecutive checkpoints satisfy $|\widehat{\lambda}_t - w| \le \frac{1}{2}|\lambda_0 - w|$. The reported half-time is the first checkpoint in this sustained three-checkpoint crossing. Runs without a sustained crossing are reported as right-censored.

Recovery times. Tab. 4 compares predicted and observed half-times at $\lambda_0 = 0.4$. For the displayed interventions, population half-times differ from the local prediction by at most approximately 2.3%, and stochastic medians preserve the same ordering. Under the blind schedule, neither population gradient descent nor any of the five stochastic runs reaches the threshold within the budget. Calibration confirms that the blind gradient remains numerically resolvable, so the absence of recovery is not a floating-point underflow artifact.

The primary intervention uses $s = 1$, while the $s = 0$, $\pi = 3\times10^{-2}$ schedule provides a full-mask boundary control. Additional runs with $s = 4$ exhibit the same inverse-curvature ordering. Over the full 10^5-step budget, the $s = 1$, $\pi = 3\times10^{-3}$ population run moves from $\lambda_0 = 0.4$ to $\lambda = 0.493$, whereas the blind population run ends at $\lambda = 0.400006$. Uncensored stochastic runs terminate once the sustained half-error crossing has been confirmed. Their terminal iterates therefore indicate recovery to the prescribed threshold, not convergence to w. Stochastic runs that never satisfy the criterion continue to the full budget and remain right-censored.



(a) Cost versus low-visibility mass. (b) Signal and noise scaling. (c) Cost versus mode-weight error.

Figure 7: Reference sample size for mode-weight errors in the conditioned-product mixture at $N = 1023$, $m_{\text{base}} = 511$, $\theta = 0.8$, and $w = 1/2$. (a) At the endpoint $\lambda = 3/4$, the reference sample size scales as $n^\star \propto \pi_s(\mu)^{-1}$ in the regime dominated by the low-visibility contribution. The vertical reference marks $\pi_0 = 1/N$. (b) For the $s = 0$ channel, the mean excess loss is proportional to π_0, whereas its standard deviation is proportional to $\sqrt{\pi_0}$. (c) At $\pi_s(\mu) = 10^{-3}$, the reference sample size scales as $n^\star \propto |\lambda - w|^{-2}$ near the truth.

Curvature and recovery dynamics. Population gradient descent isolates recovery dynamics without sampling noise, while minibatch SGD tests whether the inverse-curvature ordering persists under stochastic updates. All schedule comparisons use the same parameterization and learning rate. The half-time expression is a local prediction for fixed-step population gradient descent. The stochastic comparison concerns measured threshold-crossing times.

H.5 Sampling Cost of Mode-Weight Errors

Reference sample size. To express the strength of the restored objective signal in sampling units, define the per-example excess loss

$$S := \log p_K(X_K \mid X_{K^\complement}) - \log q_{\lambda,K}(X_K \mid X_{K^\complement}), \qquad X \sim p, \quad K \sim \mu.$$

Its expectation is the schedule-averaged masked discrepancy $\mathbb{E}[S] = \mathfrak{D}_\mu(p\|q_\lambda)$. We define the reference sample size

$$n^\star := \frac{\mathrm{Var}(S)}{\mathbb{E}[S]^2}. \tag{104}$$

At the reference sample size $n^\star$, the mean excess loss equals one standard error of its sample average. This quantity is invariant to positive rescaling of S and depends only on its signal-to-noise ratio. We use this scale to compare the sampling cost of estimating the excess loss of a fixed mode-weight perturbation across mask schedules.

Sampling cost and per-example loss. Fig. 7a shows that, at $\pi_0 = 10^{-3} \approx 1/N$ and $\lambda = 3/4$, the reference sample scale is $n^\star = 1.56 \times 10^4$. Under the blind baseline $\pi_0 = 0$, it is 5.94×10^{115}. Assigning probability approximately $1/N$ to full-mask examples thus reduces $n^\star$ by more than 111 orders of magnitude for this conditioned-product mixture. Across the positive schedule masses, the contribution from low-visibility masks is much larger than the baseline contribution. In this range, the mean excess loss is proportional to $\pi_s(\mu)$, while its standard deviation is proportional to $\sqrt{\pi_s(\mu)}$ (Fig. 7b). Substitution into Eq. (104) gives $n^\star \propto \pi_s(\mu)^{-1}$. At fixed low-visibility mass, the sample scale increases as $n^\star \propto |\lambda - w|^{-2}$ when the mode-weight error decreases (Fig. 7c).

We also compute the Bayes conditional log loss $H(X_K \mid X_{K^\complement})$ for each mask size. At $N = 1023$, masks leaving $s = 1, 2, 4$ coordinates visible have conditional entropies only 0.21%, 0.35%, and 0.58% below the full-mask value $H(X)$, respectively. Thus, these non-full low-visibility masks retain nearly the full-mask prediction difficulty while preserving mode-weight sensitivity. The comparison separates overall conditional entropy from sensitivity to a particular global-distributional error. The $s = 0$ endpoint has the distribution-free guarantee of Prop. 4.1, whereas the measured $s > 0$ behavior concerns the conditioned-product mixture.

Table 5: Exponential and power-function fits over the same uncensored classifier cells. Parentheses report the coefficient of determination, R^2.

Corpus	Features	Exponential rate	Power exponent
Code–prose	unigram	0.0110 (0.861)	−0.942 (0.9998)
Code–prose	unigram+bigram	0.0126 (0.809)	−1.109 (0.993)
German–English	unigram	0.0143 (0.943)	−1.150 (0.974)
German–English	unigram+bigram	0.0142 (0.939)	−1.149 (0.977)

H.6 Exploratory Natural-Text Data Geometry

We examine how estimated residual uncertainty $\mathrm{mmse}_p(Z \mid X_V)$ changes with context visibility in two binary classification settings: code versus prose and German versus English. Here Z records dataset membership; its interpretation as an intrinsic partition into modes is not assumed. The study provides an exploratory measurement of data geometry.

Natural-text data. For code–prose, we compare permissively licensed Python files from `github-code-clean` (CodeParrot, 2022) with English C4 documents (Raffel et al., 2020). For German–English, we use the two sides of OPUS-100 (Tiedemann, 2012; Zhang et al., 2020). We use GPT-2 tokenization (Radford et al., 2019) for both classes and pack documents to 1,152 tokens, so the packed-document length cannot reveal the label. The realized datasets contain 16,000 packed sequences per class for code–prose and 14,718 per class for German–English. We divide the packed sequences into disjoint $80/10/10$ fitting, calibration, and evaluation splits. Each class contributes 15,000 sampled windows at every visibility and split used by its estimator.

Estimators and residual uncertainty. At visible size $|V| \in \{1, 2, 4, 8\}$, we construct a posterior proxy from the ratio of class-specific Stupid Backoff n-gram scores. At $|V| \in \{12, 16, 24, 32, 64, 128, 256, 512, 1024\}$, we fit logistic classifiers with unigram or unigram-plus-hashed-bigram features and calibrate their scores by isotonic regression on the calibration split. The bigram features allow us to assess the effect of including local token order.

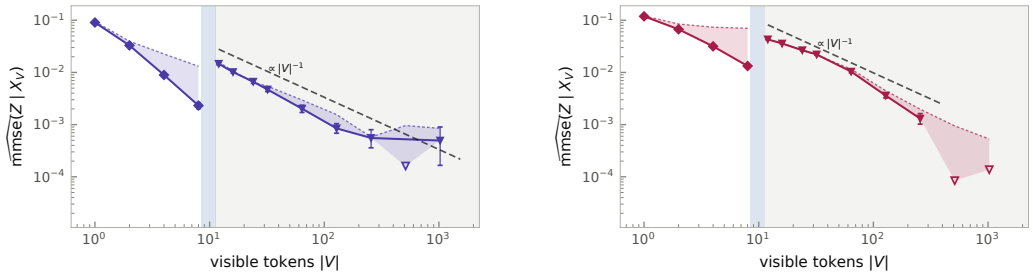
For each estimator and visibility, evaluation uses $n_{\mathrm{ev}} = 30{,}000$ windows, with 15,000 from each class. Each window is sampled by choosing a packed sequence uniformly from its class and then choosing a valid starting position uniformly within that sequence. Let $\widehat{\beta}_j$ denote the predicted positive-class probability for evaluation window j, and let z_j be its binary label. We report the plug-in uncertainty estimate and held-out Brier score

$$\widehat{U}_V := \frac{1}{n_{\mathrm{ev}}} \sum_{j=1}^{n_{\mathrm{ev}}} \widehat{\beta}_j(1-\widehat{\beta}_j), \qquad \widehat{B}_V := \frac{1}{n_{\mathrm{ev}}} \sum_{j=1}^{n_{\mathrm{ev}}} (z_j - \widehat{\beta}_j)^2.$$

Both averages weight sampled windows equally. The population Brier risk bounds Bayes residual uncertainty from above; $\widehat{B}_V$ estimates this risk, while $\widehat{U}_V$ uses the fitted posterior probabilities. We obtain 95% percentile intervals for $\widehat{U}_V$ from 1,000 bootstrap replicates. Each replicate resamples the represented packed sequences with replacement, retains all sampled windows belonging to each selected sequence, and recomputes the window average. The fitted predictors and calibration maps remain fixed during resampling.

The plug-in quantity $\widehat{U}_V$ provides a posterior-based proxy for residual mode uncertainty. The held-out Brier score $\widehat{B}_V$ estimates the population Brier risk, which upper-bounds Bayes MMSE. The shaded regions between the $\widehat{U}_V$ and $\widehat{B}_V$ curves show the gap between the posterior-based uncertainty proxy and held-out prediction error. Bootstrap intervals quantify evaluation-sample variability with the fitted predictors and calibration maps held fixed.

Calibration and censoring. For the classifier estimates, isotonic predictions are clipped to $[f, 1-f]$, where $f := \frac{1}{2n_{\mathrm{cal}}}$ and $u_{\mathrm{clip}} := f(1-f)$. Here $n_{\mathrm{cal}} = 30{,}000$ is the number of calibration windows, so $u_{\mathrm{clip}} \approx 1.66664 \times 10^{-5}$. We censor a classifier cell if $\widehat{U}_V \le 1.05\, u_{\mathrm{clip}}$ or $\widehat{B}_V > 2\widehat{U}_V$. The first criterion detects estimates near the imposed clipping floor; the second detects a large discrepancy between posterior variance and observed prediction error. The short-context



(a) Code–prose. (b) German–English.

Figure 8: Estimated residual label uncertainty in natural-text data. (a) Code–prose. (b) German–English. In both panels, short- and long-context measurements use separate estimators. Error bars show bootstrap intervals obtained by resampling packed sequences. Shaded bands extend from the estimated residual uncertainty to the held-out Brier score. Hollow markers indicate censored cells.

posterior proxies are not calibrated and are not subject to this censoring rule. The study contains 44 corpus–estimator–visibility cells: eight short-context cells and 36 classifier cells. We analyze the two estimator ranges separately and compare exponential and power-function fits over the common uncensored classifier range.

Uncertainty decay with visibility. The reported uncertainty decreases by approximately $183\times$ on code–prose between $|V| = 1$ and 1024, and by $91\times$ on German–English between $|V| = 1$ and its largest uncensored context, 256 (Fig. 8). These ratios compare short-context backoff estimates with long-context estimates from the unigram-plus-hashed-bigram classifier. Within the common uncensored classifier range, $12 \leq |V| \leq 256$, power-function fits have higher R^2 than exponential fits, with exponents ranging from -1.15 to -0.94 (Tab. 5). Adding hashed bigrams lowers Brier error on most cells while leaving the fitted decay exponents broadly similar. Seven of the 36 classifier cells are censored, all at $|V| \in \{512, 1024\}$ because their held-out Brier scores exceed twice their plug-in uncertainty estimates. No cell is censored by the clipping-floor criterion.

Relation to the controlled experiments. These measurements quantify dataset-membership uncertainty under fitted predictors rather than exact mode posteriors or joint-distribution recovery. The observed preference for power-function fits is merely descriptive across the uncensored classifier range, neither verifying nor precluding exponential decay in Bayes uncertainty. Furthermore, the plug-in estimate is not a certified bound on Bayes MMSE (unlike the population Brier risk, which yields an upper bound). Limited predictor capacity, calibration errors, and the omission of the largest-context regimes prevent strict asymptotic interpretations. In contrast, the controlled experiments compute Bayes uncertainty exactly and test its relation to masked discrepancy and mode-weight recovery.

I Discussion

I.1 Implications for Mask Schedule Design

Context visibility and global frequencies. Our results suggest that mask schedules should be evaluated not only by the difficulty of the conditional prediction problems they induce, but also by which aspects of the joint distribution remain visible to the objective. Within the two-mode reweighting family, Thm. 4.2 identifies schedule-averaged residual mode uncertainty as the governing sensitivity scale. Mode pinning provides one important regime where this quantity decays exponentially (Thm. 4.1), though the sensitivity characterization itself does not rely on exponential decay. More generally, the rate at which residual mode uncertainty vanishes with visibility determines how rapidly sensitivity to global mode weights is lost.

Low-visibility interventions. This perspective suggests a simple schedule-level intervention: retain positive probability mass on contexts for which the mode is not yet resolved. Under Asm. 4.2,

assigning mass η to masks leaving at most a fixed $s \leq s_{\mathrm{ov}}$ visible coordinates yields an $O(\sqrt{\varepsilon/\eta})$ mode-weight recovery guarantee within the reweighting family. Low-visibility mass therefore acts as a controllable source of global-frequency supervision. This also exposes a fundamental design trade-off: reducing visibility strengthens sensitivity to global structure, while potentially making the conditional prediction problem harder. The calculation in Appx. H.5 further shows how this restored signal translates into the sampling cost of detecting a fixed mode-weight perturbation.

I.2 Limitations and Open Directions

Our analysis isolates objective-level identifiability: which distributional directions are constrained by a masked-prediction objective, and at what excess-risk scale. Connecting this population geometry to end-to-end pretraining systems raises several natural questions.

Beyond the mode-reweighting family. Our low-visibility recovery guarantees govern a structured family in which the within-mode distributions are fixed and only their global weights change. Extending such guarantees to the full identifiability modulus requires controlling arbitrary directions in distribution space. Prop. 4.1 provides one distribution-free route via positive full-mask probability. Other structural assumptions may permit recovery without relying on this boundary case. Characterizing structural conditions for stable joint-distribution recovery without full masking is a natural next step.

Prediction heads and continuous objectives. Our positive guarantees are derived for the joint masked-block log loss, with conditionals induced by a common joint distribution. While the mode-blindness construction readily extends to token-factorized log loss, the full-mask recovery argument does not: a factorized full-mask loss scores only coordinate marginals rather than the joint distribution (Appx. C.7). An intriguing architectural question is whether richer masked-block prediction heads can recover some of these missing joint constraints while remaining computationally tractable at scale. In addition, continuous reconstruction and latent-regression objectives raise the analogous question of what geometric discrepancy they induce over the underlying data or latent distribution.

Data geometry and multiple modes. The mode-pinning assumption is deliberately schedule-independent, requiring every sufficiently large visible set to identify the mode with high probability. This can be restrictive when mode information is unevenly distributed across coordinates. By contrast, Thm. 4.2 depends directly on the residual uncertainty under the prescribed mask schedule, suggesting an average-case route to broader data geometries. Beyond two modes, sensitivity becomes inherently directional: scalar residual uncertainty need not control every weight contrast, and a multimode extension requires directional control of the posterior covariance (Appx. E.4). Overlapping and hierarchical mode structures provide further natural directions.

Estimating uncertainty in natural text. Our natural-text study measures uncertainty about dataset membership using fitted posterior proxies rather than exact Bayes posteriors. Over the common uncensored classifier range, power-function fits describe the estimated uncertainty better than exponential fits. These finite-range measurements do not determine the asymptotic Bayes decay rate, but motivate studying sensitivity through residual uncertainty without presupposing exponential pinning. Developing more expressive posterior estimators and scaling to broader semantic partitions will help further illuminate this data geometry in real-world pretraining.

Model class and optimization. Our theory is formulated directly over distributions and therefore decouples objective geometry from any specific parameterization or model architecture. In restricted or parametric model classes, representational constraints and inductive biases may exclude some blind alternatives or prevent the true data distribution from being represented exactly. Similarly, our controlled optimization experiments hold the within-mode distributions fixed to isolate the loss curvature governing mode-weight recovery. Understanding how this curvature interacts with jointly learned components, approximation error, and modern optimizer dynamics is an essential step toward translating objective-level identifiability into end-to-end pretraining behavior.