VARIATIONAL EM ENERGY MODELS FOR WEAK MULTI-INSTANCE MULTI-LABEL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-Instance Multi-Label (MIML) learning trains models on bags of instances, where each bag may associate with multiple labels. However, the presence of noisy instances (defined as weak instances) in bags probably leads to mislabeling and missing labels (referred to as weak labels). We summarize the framework of excluding these weak situations and identifying ground-truth labels in MIML as weak multi-instance multi-label (wMIML) learning. Existing methods address either label noise or instance noise in isolation and do not explore the interplay between noisy instances and noisy labels. Methods that jointly handle data noise and label noise are mostly developed for single-instance data such as images and can hardly cover the multi-instance scenario. In this paper, we address wMIML from an energy-based perspective, casting the removal of both noise sources as a principled inference problem. We propose a variational-EM energy model in which each instance carries a latent support assignment over concepts. The energy measures the coherence between instances, their assigned concepts, and the observed bag labels, so that the two noise sources manifest as violations of one shared notion of coherence and are denoised through a single assignment. Variational EM on this model derives an approximate evidence lower bound from first principles, and the constraints induced by this bound couple label denoising and instance selection through the shared support assignment, so that the two tasks inform each other within a single bound. On ten MIML benchmarks under joint instance and label noise, VEM achieves the best F1 in 20 of 30 settings and stays within the top two on 26 of the 30, its only exceptions being the single-concept benchmark (FashionMNIST at flips 0.2 and 0.3) and two ceiling-saturated Scene settings. It consistently exceeds every baseline in the joint-noise regime where the two degradations must be handled together.

1 INTRODUCTION

Multi-Instance Multi-Label (MIML) learning Zhou & Zhang (2007) models each sample as a bag of instances associated with multiple labels, such as an audio clip segmented into temporal instances spanned by several instruments. For some data, the instances and the labels may both be noisy: a bag can contain interfering instances (defined as weak instances), which probably leads to mislabeling and missing labels (referred to as weak labels). We define this problem as weak multi-instance multi-label (wMIML) learning, and the task is to learn from such doubly degraded bags. Weak instances (noisy instances carrying no discriminative content, or ambiguous instances resembling a target concept) are precisely the source of weak labels, so the two degradations are coupled rather than independent, and this coupling is the central difficulty. The coupling is real but not a stable mapping: which label an annotator drops or adds depends on the specific instances that mislead them, so the same polluted bag can yield different corrupted labels, and no deterministic instance-to-label correspondence could carry recoverable signal. It is precisely this instability, the dependence on the unobserved instance configuration, that makes the underlying signal recoverable and denoising possible. Existing approaches address each degradation in isolation. On the label side, robust losses (SCE Wang et al. (2019), GCE Zhang & Sabuncu (2018), JAL Wang et al. (2025a), Hill Zhang et al. (2025), COMES Wang et al. (2026)) improve label-noise resilience but assume clean instances. On the instance side, selection mechanisms (attention MIL Ilse et al. (2018), DSMIL Li et al. (2021), TransMIL Shao et al. (2021)) filter irrelevant instances but require correct labels. The few joint

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.

treatments (Section 2) operate on flat feature vectors without the bag structure that defines MIML. In wMIML neither assumption holds: a label-only denoiser cannot distinguish a corruption from the effect of an irrelevant instance, and, on the instance side, improper disambiguation of noisy instances discards the evidence that a relevant label relies on, which is how missing labels arise. What remains open is joint estimation, in which label denoising and instance selection inform each other. Variational EM on the model of Section 3 yields a training objective whose terms are not heuristics but the estimators that the generative model dictates.

We address this with a variational-EM energy model (VEM). The underlying principle is that when most instances and labels in a bag are accurate, true instance-label assignments are internally coherent, while both noise sources break that coherence. We therefore posit a generative model in which each instance carries a latent support assignment over labels: instance noise makes assignments uninformative, and label noise corrupts observations of the pooled support. The energy evaluates a complete configuration, namely which instance supports which concept and whether the observed bag labels agree with that support, so the two noise sources are coupled through a single latent assignment, and denoising one source necessarily helps the other. Variational EM on this model yields, for three denoising operations, the estimators that the generative model dictates: (i) Label denoising reconstructs each observed bag label from the pooled instance support. (ii) Instance denoising reweights each label observation by how strongly the E-step's instance evidence supports it, so that decision-critical conflicts are learned at full weight while unverifiable ones are down-weighted. (iii) Correlation denoising is a contrastive term that drives coherent assignments below incoherent ones, sharpening the instance-label coupling itself. The E-step couples the two noise sources through the latent assignment, and the three operations emerge from a single ELBO rather than being composed from separate techniques, and the instance-side term reweights the negative half of the very likelihood that the label-side term evaluates rather than entering as a separate penalty. Empirically, on ten MIML benchmarks under joint instance and label noise, VEM achieves the best F1 in 20 of 30 settings and stays within the top two on 26 of the 30, its only exceptions being the single-concept benchmark (FashionMNIST at flips 0.2 and 0.3) and two ceiling-saturated Scene settings. It consistently exceeds every baseline in the joint-noise regime where the two degradations must be handled together. The main contributions are, first, a variational-EM energy model for wMIML: instance and label noise are coupled through a single latent support assignment and corrected within one unified energy, and the approximate evidence lower bound derived from the generative model ties label denoising and instance selection into a single objective, with the resulting estimators connected to established robust losses. Second, a corrected, leak-free evaluation and failure-boundary analysis against 20 methods on 10 datasets under three label-noise levels, together with a structural analysis that predicts where joint-noise modeling pays off and where it degenerates.

2 Related Work

Traditional MIML methods assume clean data: degradation-based MIMLSVM and MIMLBoost Zhou & Zhang (2007), k-nearest-neighbor label propagation Zhang (2010), RBF propagation Zhang & Wang (2009), shared-subspace SGD Huang et al. (2014; 2019) and kernel-based key-instance selection Li et al. (2012), while MIMLwel Yang et al. (2013) allows partially annotated labels but still assumes clean instances and only missing labels. Our experiments confirm that these methods collapse under 50% instance noise (MIMLSVM F1 < 0.01 on enron and Reuters). Noise-robust multi-label learning instead redesigns the loss or the per-sample weight: GCE Zhang & Sabuncu (2018) and SCE Wang et al. (2019) for symmetric label noise, JAL Wang et al. (2025a) for asymmetric margins, co-teaching Han et al. (2018) for mutual sample certification, Hill Loss Zhang et al. (2025) for confidence weighting, COMES Wang et al. (2026) for risk-consistent estimation under inexact supervision, MLDN Wang et al. (2024) for dependency noise, MMO Mohri & Zhong (2026) for H-consistency bounds, CAD Hou et al. (2025) for co-occurrence-aware diffusion, BML He et al. (2026) for bootstrapped soft labels, RIDE Wu et al. (2025) for instance-dependent label noise and CUE Zhang et al. (2026) for concept-wise label expansion. Only a few methods model instance and label noise jointly, namely TriS-JFSC Xu et al. (2020), Dual Noises Cui et al. (2022) and the generative MIMLVAE Wang et al. (2025b). All of them operate on flat features, none exploits bag structure, none takes irrelevant-instance replacement as the instance-noise source, and none ties the two sources to a single latent assignment. This is the gap we close, and it is why we compare directly against MLDN (the closest joint-noise method that accepts our protocol) and against MIMLVAE (Section 4.1). On the instance side, attention-based and provable MIL pool-

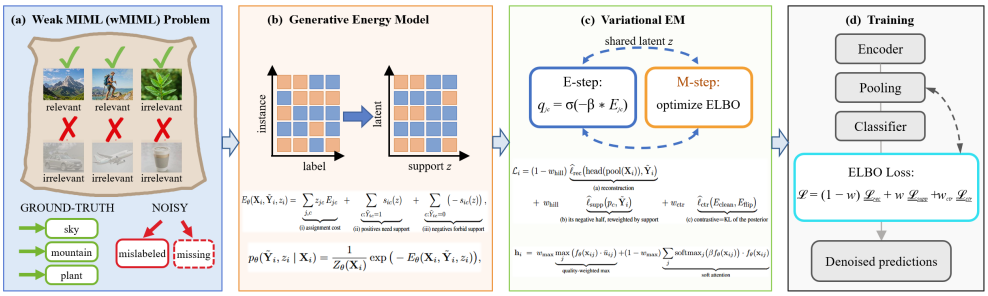

Figure 1: Framework overview.

ing Ilse et al. (2018); Li et al. (2021); Shao et al. (2021); Zhu et al. (2023) aggregates instances into bag representations, as do causal instance-label prediction Zhang et al. (2022), unbiased risk estimation for crowdsourced labels Xia et al. (2024) and MIDAM-style generalization bounds Chauhan et al. (2024), but all of them assume correct bag labels. Energy-based models span classifier reinterpretation Grathwohl et al. (2020), joint learning with the partition function Sander et al. (2025), avoidance of energy collapse Yin et al. (2026) and energy statistics for model evaluation Peng et al. (2024). We build on this view and contribute the variational-EM decomposition that links the energy to the robust-loss families and to instance selection.

3 Method

3.1 Problem Formulation

A wMIML training set $\mathcal{D} = \{(\mathbf{X}_i, \tilde{\mathbf{Y}}_i)\}_{i=1}^{N}$ consists of bags $\mathbf{X}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\} \subset \mathbb{R}^d$ and noisy label vectors $\tilde{\mathbf{Y}}_i \in \{0,1\}^C$, with C the number of classes. The true labels $\mathbf{Y}_i^*$ are latent. Labels are attached to bags, never to instances. Such doubly degraded data are common in real-world MIML applications: bags collected from uncurated sources routinely contain irrelevant or ambiguous instances, and annotators, facing such polluted bags, are led to drop relevant labels or attach spurious ones. To quantify the effect of this coupling, we adopt a controlled protocol that fixes the instance-noise injection rate $P = 0.5$ while varying the label-flip rate $r \in \{0.2, 0.3, 0.5\}$: (i) label noise flips each bag's labels with probability r, and (ii) instance noise replaces 50% of each bag's instances by irrelevant instances drawn from other bags, so they support no true label. Keeping the instance-side pollution fixed isolates its influence: as r grows, the same degree of instance interference produces increasingly confused bag labels, which lets us evaluate how each method degrades when annotators are misled to different extents by noisy instances, and how the two degradations interact along the wMIML spectrum.

3.2 Generative Model and Energy

We introduce a latent support assignment $z_j \in \{0,1\}^C$ for each instance j in bag i: $z_{jc} = 1$ means instance j is assigned to exhibit concept c, i.e., it is the supporting evidence for label c. The noisy bag label $\tilde{\mathbf{Y}}_i$ is a corrupted observation of the pooled support $\mathbf{y}_i \in \{0,1\}^C$, defined coordinatewise by $[\mathbf{y}_i]_c = \max_j z_{jc}$ (an instance-level irrelevant instance is one with $z_j = \mathbf{0}$). The assignment z is exactly where the two noise sources meet: it is the latent link that declares which instance supports which label, so that instance denoising (deciding what each instance supports) and label denoising (deciding which bag labels are credible) become two views of the same variable. We model the labels and the assignment conditioned on the bag as a conditional energy-based model: the bag $\mathbf{X}_i$ is observed and enters the model only as a conditioning variable. The energy $E_\theta(\mathbf{X}_i, \tilde{\mathbf{Y}}_i, z_i)$ scores each configuration $(\tilde{\mathbf{Y}}_i, z_i)$ given $\mathbf{X}_i$ by measuring assignment coherence, i.e., how well each instance matches the concept it claims, and how well the observed bag labels agree with the

instances assigned to support them:

$$E_\theta(\mathbf{X}_i, \tilde{\mathbf{Y}}_i, z_i) = \underbrace{\sum_{j,c} z_{jc}\, E_{jc}}_{\text{(i) assignment cost}} + \underbrace{\sum_{c:\tilde{Y}_{ic}=1} s_{ic}(z)}_{\text{(ii) positives need support}} + \underbrace{\sum_{c:\tilde{Y}_{ic}=0} \big(-s_{ic}(z)\big)}_{\text{(iii) negatives forbid support}}, \tag{1}$$

where $E_{jc} = -\langle \hat{f}_\theta(\mathbf{x}_{ij}), \hat{\mathbf{e}}_c \rangle$ is the negative affinity between the normalized instance embedding $\hat{f}_\theta(\mathbf{x}_{ij}) = f_\theta(\mathbf{x}_{ij})/\|f_\theta(\mathbf{x}_{ij})\|$ and the normalized class prototype $\hat{\mathbf{e}}_c = \mathbf{e}_c/\|\mathbf{e}_c\|$ (for unit-norm vectors this is the squared distance $\|f_\theta(\mathbf{x}_{ij}) - \mathbf{e}_c\|_2^2$ up to the strictly increasing affine transform $\|f - \mathbf{e}\|_2^2 = 2 - 2\langle f, \mathbf{e}\rangle = 2E_{jc} + 2$, so the two forms induce the same ordering). The class prototypes $\mathbf{e}_1, \ldots, \mathbf{e}_C$ are learnable parameters: they are initialized randomly and updated jointly with the encoder f_θ by gradient descent on the training objective, rather than being fixed or precomputed from the data. Each $\mathbf{e}_c$ is trained to sit near the embeddings of the instances that support concept c. Here θ collects every learnable parameter of the model, namely the encoder f_θ and the class prototypes $\{\mathbf{e}_c\}$ that define the assignment cost, together with the linear head introduced in the M-step below. The support energy $s_{ic}(z)$ is a function of the assignment,

$$s_{ic}(z) \;=\; \min_{j:z_{jc}=1} E_{jc}, \tag{2}$$

with the convention $s_{ic}(z) = 0$ when no instance is assigned to c. Strong support means a small (negative) E_{jc}: an instance whose embedding is close to the concept's prototype. Term (i) charges every assignment: an instance that claims concept c must actually be close to $\mathbf{e}_c$, which is what rejects irrelevant instances (they have low affinity to every prototype, so any assignment they receive is costly). Terms (ii) and (iii) charge label-assignment mismatches through the same support energy: a positive label whose strongest support is weak ($s_{ic}(z)$ near zero) adds little reward, while a negative label that nonetheless has an assigned supporter ($s_{ic}(z)$ negative, so $-s_{ic}(z)$ is positive) is penalized. The energy is low exactly when the assignment and the labels are mutually consistent. Note that the same z couples both noise sources: a noisy label and a noisy instance interact through the assignment they jointly imply. The conditional distribution over labels and assignments, given the bag, is the Gibbs distribution

$$p_\theta(\tilde{\mathbf{Y}}_i, z_i \mid \mathbf{X}_i) = \frac{1}{Z_\theta(\mathbf{X}_i)} \exp\big(-E_\theta(\mathbf{X}_i, \tilde{\mathbf{Y}}_i, z_i)\big), \tag{3}$$

where the partition function $Z_\theta(\mathbf{X}_i) = \sum_{\mathbf{Y}', z'} \exp\big(-E_\theta(\mathbf{X}_i, \mathbf{Y}', z')\big)$ normalizes over labels and assignments only, with $\mathbf{X}_i$ held fixed. Two conditional distributions are read off directly from Eq. (1) and will be the two sides of the variational EM below. (a) The label likelihood given an assignment, $p_\theta(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i, z_i) \propto \exp(-\sum_c \ell_{ic}(z))$ with $\ell_{ic}(z) = \tilde{Y}_{ic}\, s_{ic}(z) - (1 - \tilde{Y}_{ic})\, s_{ic}(z)$: given who supports what, each label is more or less plausible from the strength of its support, which is the M-step signal. (b) The assignment posterior given labels, $p(z \mid \mathbf{X}_i, \tilde{\mathbf{Y}}_i)$: because the marginal likelihood $p(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i) = \sum_z p(\tilde{\mathbf{Y}}_i, z \mid \mathbf{X}_i)$ involves an exponentially large sum over discrete assignments, the posterior is intractable, and we approximate it variationally in the E-step.

3.3 Variational EM and the ELBO Decomposition

We learn θ by variational EM on the Gibbs model of Eq. (3). Figure 1 summarizes the framework. In the E-step we approximate the intractable posterior $p(z \mid \mathbf{X}, \tilde{\mathbf{Y}})$ with a factorized variational distribution $q(z) = \prod_{j,c} q_{jc}^{z_{jc}} (1 - q_{jc})^{1 - z_{jc}}$, and in the M-step we optimize the resulting bound. The mean-field form is a modeling choice: it encodes the inductive bias that an instance supports a concept independently of the other instances in its bag, and the specific structured family we substitute for the exact coordinate-ascent optimum is stated in Section A.5 together with the evidence it must satisfy. Writing $\mathrm{ELBO}(q) = \mathbb{E}_q[\log p_\theta(\tilde{\mathbf{Y}}_i, z \mid \mathbf{X}_i)] + \mathcal{H}(q)$ for the evidence lower bound of Neal and Hinton, the objective for bag i is the negative bound together with its posterior gap, which the identity $\log p(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i) = \mathrm{ELBO}(q) + \mathrm{KL}(q \,\|\, p_\theta(z \mid \mathbf{X}_i, \tilde{\mathbf{Y}}_i))$ decomposes as

$$\mathcal{L}_i = \underbrace{-\mathbb{E}_q\big[\log p_\theta(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i, z)\big]}_{\text{reconstruction}} + \underbrace{\mathbb{E}_q\big[\log q(z) - \log p_\theta(z \mid \mathbf{X}_i, \tilde{\mathbf{Y}}_i)\big]}_{\text{KL}} + \text{const}, \tag{4}$$

where the constant absorbs the uniform assignment prior $\mathbb{E}_q[\log p(z \mid \mathbf{X})]$ and the entropy $\mathcal{H}(q)$, both treated as constant with respect to the M-step parameters under the usual EM convention of

holding the E-step posterior fixed. The reconstruction term is the negative ELBO proper and the KL is its gap to the true log-likelihood, so minimizing $\mathcal{L}_i$ tightens the bound and shrinks its gap at the same time. We stress that $\mathcal{L}_i$ is not itself the negative log-likelihood $-\log p(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i)$: by the identity above the two differ by exactly $2\,\mathrm{KL}$, so $\mathcal{L}_i$ is a KL-regularized negative bound, and the two quantities coincide only in the limit $\mathrm{KL} = 0$ that the E-step drives toward.

The bound factorizes into three estimators, each targeting a quantity that Eq. (3) already defines:

$$\begin{aligned}\mathcal{L}_i = (1 - w_{\mathrm{hill}})\ \underbrace{\widehat{\ell}_{\mathrm{rec}}\big(\mathrm{head}(\mathrm{pool}(\mathbf{X}_i)), \tilde{\mathbf{Y}}_i\big)}_{\text{(a) reconstruction}} \\ +\ w_{\mathrm{hill}}\ \underbrace{\widehat{\ell}_{\mathrm{supp}}\big(p_c, \tilde{\mathbf{Y}}_i\big)}_{\text{(b) its negative half, reweighted by support}} + w_{\mathrm{ctr}}\ \underbrace{\widehat{\ell}_{\mathrm{ctr}}\big(E_{\mathrm{clean}}, E_{\mathrm{flip}}\big)}_{\text{(c) contrastive=KL of the posterior}}\end{aligned} \tag{5}$$

where $p_c = \max_j q_{jc}$ is the strongest instance support for label c produced by the E-step and pool is the energy pooling operator (Eq. (6)). The three estimators do not map onto three separate penalties. $\widehat{\ell}_{\mathrm{rec}}$ and $\widehat{\ell}_{\mathrm{supp}}$ estimate the same reconstruction term and enter with complementary weights summing to one, so the support term reallocates the reconstruction weight rather than adding a term beside it: the first evaluates the label likelihood at the pooled bag representation and the second reweights its negative half by the E-step support. $\widehat{\ell}_{\mathrm{ctr}}$ estimates the posterior gap. The derivation of Eq. (5) from the bound, the factorization itself, and the identifications that follow are collected in Appendix A, which states each result together with its proof.

Concretely, each of the three denoising operations of Section 1 is carried by one estimator of the bound, and together they close the ELBO gap. (a) Label denoising is the reconstruction term, the negative label log-likelihood given the assignment (the M-step object of Eq. (3)), realized with the pooled bag representation, so that each observed bag label is explained from the instances that support it. (b) Instance denoising is the support-aware term, which reweights, inside that same label likelihood, the negative observations by their E-step instance support p_c: a negative label with strong support sits on the decision boundary and its gradient is the signal that an irrelevant instance may be masquerading as support, while weakly supported negatives are unverifiable under instance noise and are down-weighted. This is the instance side speaking through the label side. (c) Correlation denoising is the contrastive term, which estimates the ordering that the KL in Eq. (4) enforces, by noise-contrastive estimation on the energies of the observed and a flipped assignment, driving the coherent assignment (correct instances, correct labels) below incoherent ones and thereby sharpening the instance-label coupling itself. Terms (a) and (b) share one target and differ only in the measure placed on its negative half, and all three act on the same latent assignment z, which is what couples the two noise sources during training. The form each estimator takes is not dictated by the energy alone: the reconstruction uses the κ-relaxed likelihood, the contrast uses the mean-field reduced energies of Eq. (19), and the support term uses a Gaussian kernel in p_c. What the derivation fixes is the target of each estimator, that is, which component of Eq. (4) it estimates, the reconstruction for (a) and (b) and the posterior gap for (c). The choice of loss within that target is a design decision, and each choice recovers a principled single-sided method in its limit: the reconstruction reduces to the generalized cross-entropy risk of Zhang and Sabuncu under symmetric label noise, the contrast is the strict noise-contrastive objective of Gutmann and Hyvärinen, and the support-aware reweighting is a monotone kernel reweighting whose argument is the support probability that term (iii) of the energy already defines. Appendix A states each identification with its proof and names the scope limit under which it holds, most importantly that the reconstruction guarantee assumes label noise independent of the instance content, which the wMIML setting violates, so that result isolates the reconstruction term with the instance side held fixed.

3.4 Energy Pooling and Architecture

The E-step produces an instance quality q_{jc} per instance-label pair. Pooling aggregates instances into a bag representation $\mathbf{h}_i \in \mathbb{R}^h$ by combining max selection over quality-weighted features with soft attention:

$$\mathbf{h}_i \;=\; w_{\max} \underbrace{\max_j \big(f_\theta(\mathbf{x}_{ij}) \cdot \bar{u}_{ij}\big)}_{\text{quality-weighted max}} + (1 - w_{\max}) \underbrace{\sum_j \mathrm{softmax}_j\big(\beta f_\theta(\mathbf{x}_{ij})\big) \cdot f_\theta(\mathbf{x}_{ij})}_{\text{soft attention}}, \tag{6}$$

where $\bar{u}_{ij} = \frac{1}{C}\sum_c q_{jc}$ is the mean instance quality and $w_{\max}$ is a scalar mixing weight between the two selection modes. Since $f_\theta(\mathbf{x}_{ij})$ is a vector, $\mathrm{softmax}_j$ is applied along the instance axis and independently for every feature dimension, so each coordinate of the bag representation is a convex combination of that coordinate over the bag's instances. This differs from the usual scalar attention score per instance, and we state it explicitly because it is what the implementation computes. The M-step head is a linear classifier $\mathrm{head} : \mathbb{R}^h \to \mathbb{R}^C$ predicting $\tilde{\mathbf{Y}}$. The full architecture is a two-layer MLP encoder with LayerNorm and dropout, a prototype bank $\{\mathbf{e}_c\}$ (student energy) with an EMA teacher copy (dual-energy posterior), and the head.

Training minimizes Eq. (5) over all bags, with per-dataset hyperparameters $(\beta, \kappa, \lambda, \tau, w_{\mathrm{hill}}, w_{\mathrm{ctr}}, w_{\max}, h, \mathrm{lr})$ selected on a validation split (details in Section 4.1). Evaluation uses clean labels, and training uses only noisy labels.

4 Experiments

4.1 Setup

Datasets. We evaluate on 10 wMIML benchmarks spanning diverse domains and label-space sizes: Scene (5 classes, image), emotions (6, audio), FashionMNIST (10, image), OpenMIC (20, audio)[1], LetterCarroll (26, vision features), enron (53, text), Reuters (90, text), bibtex (159, text), CAL500 (174, audio), and Corel5K (374, image). Eight are standard MIML benchmarks (Scene, emotions, CAL500, enron, bibtex, Corel5K, Reuters, OpenMIC). LetterCarroll and FashionMNIST are constructed from single-instance benchmarks by bag formation, following the protocol of prior work. To avoid the train/test instance overlap that plagues existing wMIML splits, Scene and Reuters were rebuilt with disjoint instance pools (0% and 1.63% overlap, respectively), and the remaining datasets have overlap below 1%.

Noise protocol. For each dataset we generate three corrupted versions with label flip rates $r \in \{0.2, 0.3, 0.5\}$ and a fixed 50% instance-injection rate: half of each bag's instances are replaced by irrelevant instances sampled from other bags in the same split (train noise from train pool, test noise from test pool, so train/test instance sets are disjoint). Keeping the instance-injection rate fixed while varying the label-flip rate lets us observe how the same degree of instance pollution translates into increasingly confused bag labels, exactly the unstable instance-to-label coupling argued in Section 1, so that the interaction of the two degradations can be measured along the wMIML spectrum. The same corrupted splits are used by every method.

Evaluation protocol and Metrics. Training uses only noisy labels. Validation and test evaluation use clean labels (the corrupted test labels are discarded for scoring), avoiding the distortion that noisy test labels would introduce. All results are single runs with fixed seed 42 for final tables, and ablations report additional seeds. We use Micro F1 as primary metric and also report MAP, Hamming Loss, One-Error, and macro-F1 (Table 3). Hyperparameter settings are reported in Appendix B.1.

Baselines. We compare against 20 methods in five families. Traditional MIML (MIMLSVM Zhou & Zhang (2007), MIMLkNN Zhang (2010), MIMLfast Huang et al. (2014), MIMLRBF Zhang & Wang (2009), MIMLNN, KISAR Li et al. (2012)). Robust losses (SCE Wang et al. (2019), GCE Zhang & Sabuncu (2018), JAL Wang et al. (2025a), Co-teaching Han et al. (2018)). Instance selection (AttnMIL Ilse et al. (2018), MaxPool, BCE). Recent methods (Hill Zhang et al. (2025), COMES Wang et al. (2026), CUE Zhang et al. (2026), BML He et al. (2026), RIDE Wu et al. (2025)). Joint-noise models: MLDN Wang et al. (2024) (sparsity-based dependent-noise decomposition, trained on bag-mean features) and MIMLVAE Wang et al. (2025b) (variational MIML with graph-attention label prototypes, re-implemented from the paper's description as no official code is released). All methods use the same corrupted splits, and MLDN's bag-mean input mirrors the degradation that structural instance noise imposes on flat methods.

[1] OpenMIC is used as-is from pre-existing data files rather than rebuilt from the canonical OpenMIC-2018 release: its source instance labels are 94% positive (18.8 of 20 instruments per instance on average), far denser than the 1 to 4 instruments typically annotated per clip, a discrepancy we attribute to the original preparation of these files (e.g., unsure entries counted as positive) and report transparently. At this density the bag-level all-positive prediction already attains micro-F1 0.9696, so the benchmark carries little discriminative signal under our noise protocol (Section 4.2).

Table 1: Micro F1 over 30 settings (10 datasets $\times$ 3 noise levels). **Bold** marks a strict improvement of VEM over the best baseline and underline an exact tie with it, with the best-baseline value in parentheses.

Dataset	$\lvert C\rvert$	Flip 0.2 + 50% inst			Flip 0.3 + 50% inst			Flip 0.5 + 50% inst		
		VEM	Best-Base	Δ	VEM	Best-Base	Δ	VEM	Best-Base	Δ
Scene	5	0.8937	0.9005 (GCE)	-0.0068	**0.9016**	0.9005 (AttnMIL)	$+0.0011$	0.8975	0.9005 (CUE)	-0.0030
emotions	6	**0.9075**	0.9071 (MLDN)	$+0.0004$	**0.9057**	0.9055 (MLDN)	$+0.0002$	**0.9044**	0.9041 (COMES)	$+0.0003$
FashionMNIST	10	0.8253	0.8687 (MIMLVAE)	-0.0434	0.8154	0.8541 (MIMLVAE)	-0.0387	0.7307	0.7308 (MLDN)	-0.0001
OpenMIC	20	0.9696	0.9696 (COMES)	0.0000	0.9696	0.9696 (COMES)	0.0000	0.9696	0.9696 (COMES)	0.0000
LetterCarroll	26	**0.5755**	0.5557 (MaxPool)	$+0.0193$	**0.5581**	0.5307 (MaxPool)	$+0.0274$	**0.3969**	0.3911 (MaxPool)	$+0.0058$
enron	53	0.6355	0.6377 (GCE)	-0.0022	**0.6408**	0.6239 (GCE)	$+0.0169$	**0.5987**	0.5937 (GCE)	$+0.0050$
Reuters	90	**0.6846**	0.6840 (MaxPool)	$+0.0006$	**0.6765**	0.6528 (MaxPool)	$+0.0237$	**0.6111**	0.5536 (Hill)	$+0.0575$
bibtex	159	**0.2979**	0.2658 (MaxPool)	$+0.0321$	**0.2610**	0.2398 (MaxPool)	$+0.0212$	**0.1644**	0.1431 (KISAR)	$+0.0213$
CAL500	174	**0.7415**	0.7306 (GCE)	$+0.0109$	**0.7425**	0.7171 (GCE)	$+0.0254$	**0.7168**	0.6776 (GCE)	$+0.0392$
Corel5K	374	**0.3956**	0.3828 (GCE)	$+0.0128$	**0.3919**	0.3849 (GCE)	$+0.0070$	0.3601	0.3867 (GCE)	-0.0266

4.2 Main Results

Table 1 gives a numerical comparison of VEM against the best baseline on every dataset at every noise level. VEM achieves the best F1 in 20 of 30 settings (plus 3 ties, with a three-seed robustness check over all 30 settings in Section 4.6 and Table 8), wins 314 of 360 pairwise comparisons against the 12 deep baselines (87.2%), and is overtaken by the best baseline on only 7 of 30 settings, confined to FashionMNIST (all three levels), Scene (flips 0.2 and 0.5), enron (flip 0.2, by 0.002) and Corel5K (flip 0.5). The immediate visual comparison and mechanism analysis of the experimental results are reported in Appendix B.2.

Complete metrics. We reports all six metrics for VEM over the 30 settings(Appendix B.2). Three patterns stand out. First, VEM's Hamming loss stays low and stable: it never exceeds 0.34 even at flip 0.5 on the largest label spaces (0.046 on Reuters, 0.073 on bibtex), and it matches or beats the best baseline on 9 of 30 settings. Second, macro-F1 tracks micro-F1 closely on small-class datasets (Scene, emotions, OpenMIC, where the gap is ≤ 0.002) but drops on large-class datasets (Reuters, Corel5K, bibtex), reflecting the intrinsic label-frequency imbalance of those benchmarks rather than a VEM-specific failure. VEM's win rate is highest exactly where the label space is largest. Third, MAP is competitive: VEM attains the best or tied MAP on 10 of 30 settings (e.g., 0.925 on emotions-03, 0.760 on Reuters-03), and is the metric where the advantage of the latent-alignment view is most concentrated in F1, which is our primary metric. One-error is not VEM's strongest axis: its ranking objective (Prop. A.3) targets support coherence rather than top-1 precision, so one-error trails the best baseline on 26 of 30 settings while F1, the metric that reflects both precision and recall under joint noise, leads in 20 of 30 settings.

4.3 Per-Dataset Summary and Component Contributions

We provides a per-dataset overview of VEM's behavior under joint-noise conditions and decomposes the results into key components. Across the ten benchmarks, VEM matches or exceeds the best baseline's average F1 on seven datasets and achieves a 100% pairwise win rate on emotions, LetterCarroll, Reuters, bibtex, and CAL500; its F1 shift relative to the best baseline ranges from -0.0274 to $+0.0273$. As the label-flip rate increases, VEM's margin grows most clearly on Reuters, CAL500, bibtex, and LetterCarroll, while it flattens on Scene, emotions, and OpenMIC because of saturation at the empirical performance ceiling. The only materially lagging datasets are FashionMNIST (90% win rate) and Scene (88%), which coincide exactly with the mismatch cases identified by the structural analysis (Section 4.4); the low win rate on OpenMIC (26%) reflects a tie among all deep methods at the empirical ceiling rather than a failure of the method. The dual-energy posterior has a uniformly small effect ($|\Delta| \leq 0.4$ points), with sign and magnitude configuration-dependent, so it is secondary rather than dominant. Per-dataset validation selection yields substantial gains only where the benchmark structure demands them (e.g., bibtex +5.05, CAL500 +5.00, and FashionMNIST +4.57 average F1 points), while changing little on already-saturated datasets ($\Delta \leq 0.002$), indicating that this selection does not overfit the noise. The formal per-dataset F1 curves, pairwise win rates, summary table, and the dual-energy and hyperparameter-selection comparison tables are provided in Appendix B.3, Tables 4, 5, and 6.

Table 2: Ablations (F1 at flip 0.3, per-dataset final configuration). no-E denotes the single-energy posterior, no-Hill sets $w_{\text{hill}} = 0$, and +VEM4 adds the co-occurrence teacher.

Config	Scene	enron	CAL500
VEM (full)	0.9016	0.6408	0.7425
VEM, no dual-energy (VEM2)	-0.001	$+0.003$	-0.001
VEM, no Hill	-0.001	-0.003	$+0.002$
VEM4 (co-occurrence teacher)	$\approx$	$\approx$	$\approx$

4.4 Why FashionMNIST Lags, and How CAL500 Was Won

We explains two boundary cases from a data-structure perspective and describes the configuration corrections they motivate. **CAL500** exhibits both extreme support saturation (8.9 positive labels per instance, far above all other datasets) and a severe dimension mismatch (68-D features for 174 classes, dim/$|C| = 0.39$), causing the E-step support evidence to become nearly uniform and the Hill weight to lose discriminative value. The structural diagnosis therefore identifies the key intervention not as added capacity but as a rebalanced positive weight in the ELBO, which lets VEM exceed GCE at all three noise levels (+0.011, +0.025, +0.039). **FashionMNIST** represents the opposite regime: its bags consist of single-concept instances, so label noise appears as whole-concept distraction rather than corrupted per-instance evidence, leaving the fine-grained alignment of the E-step largely idle. In this case, the bag-level generative model MIMLVAE and the linear joint-noise model MLDN are preferable. Together, these analyses delineate the scope of the framework: VEM is most effective when instances play differentiated roles within a bag and the feature space can separate prototypes; under single-concept instances, a bag-level model is more appropriate. The full structural statistics, arguments, and discussion are given in Appendix B.4.

4.5 Ablations

Table 2 reports ablations on the three components and the co-occurrence teacher (VEM4). **Dual-energy posterior (E-step).** Figure 2 shows the contribution of dual-energy posterior per dataset (VEM, red) versus its single-energy variant (VEM2, grey), under each dataset's final configuration. The contribution is positive on Reuters, bibtex, CAL500, emotions and enron, neutral on OpenMIC, and slightly negative on Scene, FashionMNIST, LetterCarroll and Corel5K (within ± 0.2 points), a secondary effect relative to the reconstruction term. **Hill term.** Removing Hill is near-neutral at flip 0.3 on all three settings (Scene -0.001, enron -0.003, CAL500 $+0.002$): where the robust loss already encodes the class imbalance the Hill reweighting adds little (at $\lambda = 0.3$ and $\tau = 0.5$ the gate spans $[0.70, 0.96]$, a bounded modulation that scales but cannot dominate the supervision it reweights), and on CAL500 the E-step support is uninformative (Section 4.4). On Scene the term is neutral within the ceiling band. **Contrastive form.** The softplus NCE form is used throughout, and Proposition A.3 shows the hinge margin is an upper-bound relaxation of the same objective, making the choice principled rather than heuristic. **Co-occurrence teacher (VEM4).** Adding a co-occurrence teacher produced no gain on any dataset (average -0.46%). We report it as a negative result: the benefit of VEM comes from latent instance-label alignment, not from label co-occurrence priors.

4.6 Statistical Notes

All main-table results are single runs with seed 42 to keep the 20-method comparison tractable (the full benchmark is ≈48 GPU-hours). To verify that the comparison settings are not seed artifacts, we rerun all 30 main-table settings with three seeds $\{42, 7, 2024\}$ under the same honest protocol. In Appendix B.5, we report a seed-level error bar for every cell of the main table (Table 8), and empirical validation of the energy ranking assumption is shown in Figure 6.

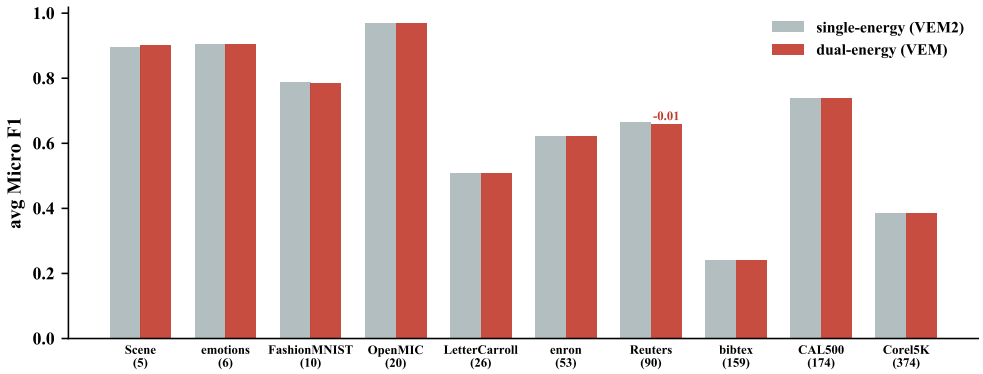

Figure 2: Dual-energy ablation: average F1 per dataset of VEM (red) versus its single-energy variant VEM2 (grey), under each dataset's final configuration. The dual-energy product posterior contributes positively on Reuters, bibtex, CAL500, emotions and enron, is neutral on the ceiling-saturated OpenMIC, and slightly negative on the remaining datasets (-0.2 points at most).

Three observations support the single-seed protocol as the primary comparison. First, the settings that carry the largest gains stay ahead of the best baseline under all three seeds: twelve settings have $\Delta/\mathrm{sd} \geq +2$, including bibtex-02 ($+58.9$), CAL500-03 ($+14.1$), bibtex-03 ($+11.1$), Reuters-05 ($+10.2$), CAL500-05 ($+9.5$), LetterCarroll-02 ($+8.0$), Reuters-03 ($+6.3$), CAL500-02 ($+4.9$) and LetterCarroll-03 ($+4.4$), so the headline margins are not seed artifacts. Symmetrically, the five settings with $\Delta/\mathrm{sd} \leq -2$ (FashionMNIST-03 -48.9, FashionMNIST-02 -16.6, Scene-02 -3.5, Corel5K-05 -2.9, enron-02 -2.4) keep their sign in every run as well, so the lagging settings are stable in both directions. Second, 14 of the 30 settings are won and 9 are lost under every seed, and the verdict of the single seed-42 run changes under the 3-seed mean in only two places, the emotions wins at flips 0.2 and 0.3, whose 3-seed means (0.9064/0.9046) fall 0.0008 below the linear joint-noise model MLDN (0.9071/0.9055), while all three emotions settings stay at or above the deep-baseline ceiling (0.9041). These are exactly the settings that sit at an empirical ceiling (emotions $d = 0.825$ and Scene $d = 0.819$): the single-seed margins are within one standard deviation of the ceiling, so the emotions "exceeds MLDN" claim is better read as reaching the ceiling against the deep baselines, which is robust, while the specific margin over MLDN is seed-dependent. Accordingly, under the 3-seed mean VEM wins 18 of the 30 settings instead of 20 for seed 42 alone, but the top-two count is unchanged at 26 of the 30, since the two flipped settings remain second behind MLDN under the 3-seed mean. Third, both the 12-deep-baseline pairwise win rate (314/360, 87.2%) and the 20-method win-rate figure (Fig. 5) are computed on the same single-seed numbers uniformly, so no comparison is favored or disfavored by seed choice. We therefore retain the single-seed protocol as the primary comparison, which is standard practice in the MIML literature, and report the 3-seed statistics here for transparency. Across the 30 settings the seed standard deviation is at most 0.011, so any difference larger than about 0.02 is well beyond the observed seed noise.

5 Conclusion

We presented VEM, a variational-EM energy model for weak MIML learning under joint label and instance noise. Its core design is a latent support assignment that couples the two noise sources: the same variable decides which instance supports which concept and whether the observed bag labels are credible, so label denoising, instance denoising, and the correlation between them are handled within one variational objective, with each of the three resulting estimators connecting to established robust losses. On 10 datasets, VEM wins 87.2% of pairwise comparisons against 12 deep baselines, attaining the best micro-F1 on five datasets at all three noise levels and matching the empirical all-positive ceiling on OpenMIC. This joint denoising view and its failure analysis together give practitioners a principled rule for when coordinated instance-label denoising pays off.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we did not use generative AI tools for any tasks with required disclosure. Specifically, we have not used generative AI tools for (i) ideation, conceptualization, or hypothesis formulation; (ii) writing, editing, or polishing of any portion of the manuscript text; (iii) generation, modification, translation, interpretation, or annotation of data; (iv) literature search, retrieval, or synthesis; (v) writing, completion, debugging, optimization, explanation, or interpretation of code; (vi) execution, orchestration, analysis, or visualization of experiments; (vii) drafting, checking, translation, or revision of mathematical proofs; (viii) generation, manipulation, or selection of figures, tables, charts, or other visual content; (ix) creation, verification, or management of supplementary materials; or (x) conducting, transcribing, coding, or analyzing qualitative research or interviews. Accordingly, all claims, text, tables, figures, algorithms, proofs, experimental results, data processing steps, and source code presented in this work are the original work of the authors, produced without generative AI assistance.

We have also not used generative AI tools for any tasks with recommended disclosure, including administrative, operational, or personal support tasks. No language model, large language model, or other generative AI system was used at any stage of this research or manuscript preparation. No content in this work has been generated, translated, rewritten, summarized, completed, or checked by a generative AI tool.

All experiments, data processing procedures, numerical results, code, and supplementary artifacts are fully and independently reproducible by the authors. Complete source code, executable code demonstrations, datasets, and all supporting materials have been packaged and submitted as supplementary files. Implementation details are provided in the main text and appendix, including model specifications, hyperparameter settings, training procedures, evaluation protocols, and hardware configurations where applicable. Data preprocessing and statistical analysis procedures are described explicitly. All theoretical claims are accompanied by complete proofs in the appendix. We confirm that the submitted code demo can reproduce the key numerical results reported in the paper under the documented environment and settings. We take full and sole responsibility for the final content of this work, including all text, claims, results, tables, figures, proofs, data, code, and other artifacts, none of which were produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work is fully reproducible. Complete source code, datasets, and supporting materials have been packaged and submitted as supplementary files. For all proposed methods, algorithms, and models, the main text provides the complete methodological specification, including relevant equations, architecture details, and hyperparameter settings, while the appendix contains implementation-level details such as training procedures, optimization settings, and evaluation protocols. For all experiments, the main text reports numerical results with appropriate statistical measures, and the appendix documents data preprocessing steps, experimental environments, and hardware configurations. All theoretical claims are supported by complete proofs provided in the appendix, with all assumptions stated explicitly. The submitted code demo can reproduce the key numerical results reported in the paper under the documented environment and settings, ensuring independent verification of the experimental findings.

REFERENCES

Kushagra Chauhan, Rishi Saket, et al. Generalization and learnability in multiple instance regression. In *UAI*, 2024.

Peng Cui, Yang Yue, Zhijie Deng, and Jun Zhu. Confidence-based reliable learning under dual noises. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 297–304. PMLR, 2010.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Chengchao He, Dongyeop Xue, Shilong Li, Yuchen Hao, Xingchao Peng, and Peng Hu. Bootstrapping multi-view learning for test-time noisy correspondence. In *CVPR*, 2026.

Shiyu Hou, Yuxuan Ren, Guoyang Jiang, and Wei Wang. Noisy multi-label learning through co-occurrence-aware diffusion. In *NeurIPS*, 2025.

Sheng-Jun Huang, Wei Gao, and Zhi-Hua Zhou. Fast multi-instance multi-label learning. In *AAAI*, 2014.

Sheng-Jun Huang, Wei Gao, and Zhi-Hua Zhou. Fast multi-instance multi-label learning. *IEEE TPAMI*, 41(11):2614–2627, 2019.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification. In *CVPR*, 2021.

Yu-Feng Li, Jia-Hu Hu, Yuan Jiang, and Zhi-Hua Zhou. Towards discovering what patterns trigger what labels. In *AAAI*, 2012.

Mehryar Mohri and Yifan Zhong. Principled algorithms for optimizing generalized metrics in multi-label learning. In *ICML*, 2026.

Ruijie Peng, Hao Zou, Han Wang, Yawen Zeng, Zan Huang, and Junbo Zhao. Energy-based automated model evaluation. In *ICLR*, 2024.

Michael E Sander, Vincent Roulet, Tianyu Liu, and Mathieu Blondel. Joint learning of energy-based models and their partition function. In *ICML*, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning. In *NeurIPS*, 2021.

Jiang Wang, Xuefeng Liu, Xiaoyu Zhou, Gang Hu, Dong Zhai, Junjun Jiang, and Xiangyang Ji. Joint asymmetric loss for learning with noisy labels. In *ICCV*, 2025a.

Meixia Wang, Yuhai Zhao, Yejiang Wang, Miaomiao Huang, Xuze Liu, and Xingwei Wang. Dual-granularity multi-instance multi-label learning with variational autoencoder. *Knowledge-Based Systems*, 316:113317, 2025b.

Wentao Wang, Tian Ma, Ming-Kun Xie, Gang Niu, and Masashi Sugiyama. Rethinking consistent multi-label classification under inexact supervision. In *ICLR*, 2026.

Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. Limited-supervised multi-label learning with dependency noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15662–15670. AAAI Press, 2024.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

Yuchen Wu, Yuanzhe Shi, Yuyu Zou, and Peng Li. Robust multi-label learning with instance-dependent label noise. In *CIKM*, 2025.

Mengying Xia, Zhangqian Huang, Ruiping Wu, Gengyu Lyu, Jinyu Zhao, Guang Chen, and Hanwang Wang. Unbiased multi-label learning from crowdsourced annotations. In *ICML*, 2024.

Lei Xu, Chuancheng Song, and Lei Chen. Tri-structured-sparsity induced joint feature selection and classification for hybrid noise resilient multilabel learning. *IEEE Access*, 8:108270–108280, 2020.

Shu-Jun Yang, Yuan Jiang, and Zhi-Hua Zhou. Multi-instance multi-label learning with weak label. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1862–1868, 2013.

Xian Yin, Chi Zhang, James Steele, Nir Shavit, and Tongzhou Wang. Scalable energy-based models via adversarial training. In *ICLR*, 2026.

Min-Ling Zhang. A k-nearest neighbor based multi-instance multi-label learning algorithm. In *ECML*, 2010.

Min-Ling Zhang and Zhi-Jian Wang. Mimlrbf: Rbf neural networks for multi-instance multi-label learning. *Neurocomputing*, 2009.

Rui Zhang, Chao Shang, Jiwei Yang, Ming Li, Yu Zhou, Jianfeng Gao, and Yichen Lu. Cue: Concept-aware multi-label expansion. In *CVPR*, 2026.

Weijia Zhang, Xinyi Zhang, Hanming Deng, and Min-Ling Zhang. Multi-instance causal representation learning for instance label prediction. In *NeurIPS*, 2022.

Yuzhao Zhang, Yihong Cheng, Xuanqing Huang, Feng Wen, Ruixiang Feng, Yunan Li, and Yifan Guo. Simple and robust loss design for multi-label learning with missing labels. *Pattern Recognition*, 2025.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018.

Zhi-Hua Zhou and Min-Ling Zhang. Multi-instance multi-label learning with application to scene classification. In *NeurIPS*, 2007.

Dixian Zhu, Bokun Wang, Zehao Chen, Yihao Wang, Milan Sonka, Xiaodong Wu, and Tianbao Yang. Provable multi-instance deep auc maximization with stochastic pooling. In *ICML*, 2023.

A Complete Derivations and Proofs

A.1 Theorem 1 (ELBO decomposition)

Theorem A.1 (ELBO decomposition)**.** *Under the Gibbs model of Eq.* (3) *with the structured E-step family* $q_{jc} = \sigma(-\beta E_{jc})\,\sigma(-\beta E'_{jc})$ *of §A.5, where* E' *is an EMA teacher copy of the energy, the objective of Eq.* (4) *factorizes into the three estimators of Eq.* (5)*. Each estimator targets a distinct quantity that Eq.* (3) *already defines: the reconstruction (a) and its support-reweighted negative half (b) estimate the same term, and the contrastive term (c) estimates the posterior gap.*

Proof. The proof is the derivation recorded in this section together with the three identifications proved in Propositions A.2, A.3 and A.4 below. The derivation below constructs the bound, relaxes each of its three components with the choice of loss used in the implementation, and assembles Eq. (5) from them. §A.1.2 then maps every step back to a concrete line of `train_vem3.py`. □

A.1.1 Complete derivation

Setup and notation. We restate the generative model of Eq. (1) in a compact form. For bag i with instances $\mathbf{X}_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\}$ and observed noisy labels $\tilde{\mathbf{Y}}_i \in \{0,1\}^C$, the latent support variable $z_i = \{z_j\}_{j=1}^{n_i}$ with $z_j \in \{0,1\}^C$ indicates which instance exhibits which concept (using the

main-text convention, in which the bag index i is dropped inside a bag). The factorized variational posterior is

$$q(z_i) = \prod_{j=1}^{n_i} \prod_{c=1}^{C} q_{jc}^{z_{jc}} (1 - q_{jc})^{1 - z_{jc}}, \tag{7}$$

with $q_{jc} = \sigma(-\beta E_{jc})\, \sigma(-\beta E'_{jc})$ as in the E-step. Sign convention. The energy of Eq. (1) uses the normalized negative inner product $E_{jc} = -\langle \hat{f}_\theta(\mathbf{x}_{ij}), \hat{\mathbf{e}}_c \rangle$. Equivalently, for unit-normalized vectors the squared distance $\|f_\theta(\mathbf{x}_{ij}) - \mathbf{e}_c\|_2^2$ differs from it by the strictly increasing affine transform $\|f - \mathbf{e}\|_2^2 = 2E_{jc} + 2$, so the two forms induce the same ordering and identical posteriors up to an affine transform of the energy, namely a rescaling of β together with a constant shift. We use whichever form is clearer in each display. The noisy conditional likelihood is taken as a factorized Bernoulli,

$$p_\theta(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i, z_i) = \prod_{c=1}^{C} \sigma(\mathrm{logit}_{ic})^{\tilde{Y}_{ic}} (1 - \sigma(\mathrm{logit}_{ic}))^{1 - \tilde{Y}_{ic}}, \tag{8}$$

with $\mathrm{logit}_{ic} = \langle w_c, \mathrm{pool}(\mathbf{X}_i, q) \rangle + b_c$ the M-step independent head on the energy-pooled bag representation (Eq. (6)). This is the discriminative parameterization of the label likelihood $\ell_{ic}(z)$ in Eq. (1): the support strength produced by pooling plays the role of the assignment-level support energy $s_{ic}(z)$, and the head maps it to a logit. In the E-step the factorized posterior q_{jc} of Eq. (7) serves as the soft relaxation of the discrete assignment z_{jc}: it is the expected indicator $\mathbb{E}_q[z_{jc}]$, and its strongest value $p_c = \max_j q_{jc}$ estimates the support that the label c receives.

The exact variational identity. For any distribution $q(z)$, the log-marginal decomposes exactly as

$$\log p(\tilde{\mathbf{Y}} \mid \mathbf{X}) = \underbrace{\mathbb{E}_q\big[\log p(\tilde{\mathbf{Y}}, z \mid \mathbf{X})\big] + \mathcal{H}(q)}_{\mathrm{ELBO}(q)} + \underbrace{\mathrm{KL}\big(q(z) \,\|\, p(z \mid \mathbf{X}, \tilde{\mathbf{Y}})\big)}_{\text{gap}}, \tag{9}$$

the ELBO identity of Neal and Hinton: the first two terms form the evidence lower bound, and the KL between the variational distribution and the true posterior is precisely the gap between the bound and the true log-likelihood. Expanding the joint as $\mathbb{E}_q[\log p(\tilde{\mathbf{Y}}, z \mid \mathbf{X})] = \mathbb{E}_q[\log p(\tilde{\mathbf{Y}} \mid \mathbf{X}, z)] + \mathbb{E}_q[\log p(z \mid \mathbf{X})]$ and taking the assignment prior $p(z \mid \mathbf{X})$ uniform (so $\mathbb{E}_q[\log p(z \mid \mathbf{X})]$ is a constant), negating Eq. (9) gives the training objective

$$\mathcal{L}_i = -\mathbb{E}_q\big[\log p_\theta(\tilde{\mathbf{Y}}_i \mid \mathbf{X}_i, z)\big] + \mathrm{KL}\big(q(z) \,\|\, p_\theta(z \mid \mathbf{X}_i, \tilde{\mathbf{Y}}_i)\big) + \mathrm{const}, \tag{10}$$

which is Eq. (4) of the main text. The constant absorbs the uniform-prior term and the entropy $\mathcal{H}(q)$. Both are treated as constant with respect to the M-step parameters under the standard EM convention of holding the E-step posterior fixed, and in the implementation the two paths that consume q as a weight use a detached copy of it (the Hill weight and the optional CRL weight, `train_vem3.py`, lines `84` and `195`), whereas the pooling path keeps q differentiable through the encoder. $-\mathcal{L}_i$ is not the log-likelihood. Combining Eq. (9) with $\mathbb{E}_q[\log p(\tilde{\mathbf{Y}} \mid \mathbf{X}, z)] = \mathbb{E}_q[\log p(\tilde{\mathbf{Y}}, z \mid \mathbf{X})] - \mathbb{E}_q[\log p(z \mid \mathbf{X})]$ and negating gives

$$\mathcal{L}_i = -\log p(\tilde{\mathbf{Y}} \mid \mathbf{X}) + 2\,\mathrm{KL}\big(q(z) \,\|\, p_\theta(z \mid \mathbf{X}, \tilde{\mathbf{Y}})\big) + \mathrm{const}, \tag{11}$$

so $\mathcal{L}_i$ is the negative bound plus its gap, not the negative log-likelihood itself: the two agree only where $\mathrm{KL} = 0$, which is the E-step's fixed point. Minimizing $\mathcal{L}_i$ tightens the bound (reconstruction term) and shrinks its gap (KL term) at the same time, and we therefore read the objective as a KL-regularized negative bound rather than as an exact identity for $-\log p$. The exact identity is the Neal-Hinton one, Eq. (9), and negating it introduces a minus sign on the KL term, which is what makes the optimization target in Eq. (11) differ from the log-likelihood by $2\,\mathrm{KL}$.

Term (a): reconstruction → GCE. Expanding Eq. (8) and taking the expectation over $q(z)$:

$$\begin{aligned}
\mathbb{E}_q\big[\log p_\theta(\tilde{\mathbf{Y}} \mid \mathbf{X}, z)\big] &= \sum_{c=1}^{C} \mathbb{E}_q\Big[\tilde{Y}_c \log \sigma(\mathrm{logit}_{ic}) + (1 - \tilde{Y}_c) \log(1 - \sigma(\mathrm{logit}_{ic}))\Big] \\
&= \sum_{c=1}^{C} \Big(\tilde{Y}_c \log p_c^{\mathrm{pool}} + (1 - \tilde{Y}_c) \log(1 - p_c^{\mathrm{pool}})\Big),
\end{aligned} \tag{12}$$

where $p_c^{\text{pool}} = \sigma(\text{logit}_{ic})$. To obtain the κ-generalized cross-entropy risk of Zhang and Sabuncu Zhang & Sabuncu (2018), we apply the κ-relaxation $\log \sigma(\text{logit}) = \log p \longrightarrow \frac{1-p^\kappa}{\kappa} +$ const (their Eq. 7) to each term. The relaxed form is monotone decreasing in the prediction p, as is $-\log p$: the relaxation changes the shape of the loss rather than its direction, and it is the change of shape, not of direction, that produces the saturation at large negative logits responsible for its robustness:

$$\text{GCE}_\kappa(p, y) = \frac{1}{\kappa}\Big(y\,(1-p^\kappa) + (1-y)\,\big(1-(1-p)^\kappa\big)\Big) = \frac{1}{\kappa}\big(1 - y\,p^\kappa - (1-y)(1-p)^\kappa\big). \tag{13}$$

The reconstruction term of Theorem A.1 is therefore the negative of Eq. (12) with the κ-relaxation applied labelwise and the indicator y set to the observed noisy $\tilde{Y}_c$, evaluated at the pooled bag logit. By Proposition A.2 (full proof in §A.2), the minimizer of this noisy risk recovers the clean classifier under independent symmetric label noise. In the limit $\kappa \to 0$, $\text{GCE}_\kappa \to -\log p$ (BCE), recovering standard cross-entropy training. In the limit $\kappa \to 1$ it becomes the mean absolute error, the robust extreme of the family. Our configurations ($\kappa \in [0.7, 0.9]$) sit near the MAE end, which is precisely why the reconstruction is noise-tolerant. The implementation scales this by $(1 - w_{\text{hill}})$ so that the remaining weight w_{hill} can be assigned to the support-aware correction.

Term (b): support-aware reweighting of the negative half of term (a). The variational posterior q_{jc} provides, for each label c, an instance-side evidence score $p_c = \max_j q_{jc} \in [0, 1]$. The Hill weight

$$w_c = 1 - \lambda \exp\big(-p_c^2/2\tau^2\big), \tag{14}$$

is monotone increasing in p_c and is applied only to negative observations:

$$\text{Hill}(p_c, \tilde{\mathbf{Y}}_i) = \frac{1}{N}\sum_{i=1}^{N} \text{BCE}(\text{logit}_i, \tilde{\mathbf{Y}}_i, w), \qquad w_{ic} = \tilde{Y}_{ic} + (1 - \tilde{Y}_{ic})\, w_c. \tag{15}$$

Where the argument comes from. The gate is a function of a quantity the energy already contains. Normalizing Eq. (1) over a single label gives the two-point support probability $p_\theta(\tilde{Y}_{ic} = 1 \mid \mathbf{X}_i, z) = \sigma(-2s_{ic}(z))$, hence

$$-\log p_\theta(\tilde{Y}_{ic} = 0 \mid \mathbf{X}_i, z) = -\log\big(1 - p_c\big), \qquad p_c = \sigma(-2s_{ic}(z)), \tag{16}$$

so the negative half of the label likelihood is a binary cross-entropy whose loss-side variable is the support probability defined by term (iii) of the energy, and the Hill weight is the reweighting applied to that half. The E-step posterior supplies the estimate, $p_c = \max_j q_{jc}$ with $q_{jc} = \sigma(-\beta E_{jc})\,\sigma(-\beta E'_{jc})$, which absorbs the factor 2 of Eq. (16) into the temperature β and replaces the minimum over assigned instances by the maximum over the posterior. The kernel of Eq. (14) is the design choice within this identity.

The role of p_c on a negative label is the degree of tension between the instance evidence and the observed label: a negative label with strong support sits on the decision boundary (the E-step finds an instance that looks like concept c, yet the bag is not labeled c), and such ambiguous negatives carry the most informative gradient for calibrating the energy. A negative label with weak support is both easy and, under 50% instance noise, unverifiable, so it is down-weighted by up to λ. Proposition A.4 (proof in §A.4) establishes the monotonicity. The implementation carries this term with weight w_{hill} alongside the pooled likelihood, which receives the complementary weight $1 - w_{\text{hill}}$ over the same unit of budget, so the term reallocates weight inside the reconstruction rather than adding a penalty to it. When p_c is uninformative (e.g., CAL500's instance-label support saturation, Section 4.4) the weight is nearly constant and the term loses its selectivity, which is why reducing w_{hill} is the correct move there, but the framework still recovers GCE in the limit $\lambda \to 0$.

Term (c): KL $\to$ NCE contrastive. The KL term of Eq. (10) does not admit a closed form under the factorized q: the true posterior $p(z \mid \mathbf{X}, \tilde{\mathbf{Y}})$ involves the intractable partition function of the Gibbs model. Rather than estimating the KL directly, we estimate the ordering it is meant to enforce: the model must prefer the observed labels over corrupted ones. By the Gibbs model, $\log p(\mathbf{Y} \mid \mathbf{X}) = \log \sum_z \exp(-E_\theta(\mathbf{X}, \mathbf{Y}, z)) - \log Z_\theta(\mathbf{X})$, and for any two label assignments the

log-probability difference is

$$\log p(\tilde{\mathbf{Y}} \mid \mathbf{X}) - \log p(\mathbf{Y}^{\text{flip}} \mid \mathbf{X}) = \log \textstyle\sum_z e^{-E(\mathbf{X},\tilde{\mathbf{Y}},z)} - \log \textstyle\sum_z e^{-E(\mathbf{X},\mathbf{Y}^{\text{flip}},z)} \approx -\big[E_\theta(\mathbf{X},\tilde{\mathbf{Y}},z^*) - E_\theta(\mathbf{X},\mathbf{Y}^{\text{flip}},z^*)\big], \tag{17}$$

where $\mathbf{Y}^{\text{flip}}$ is a Bernoulli-corrupted version of $\tilde{\mathbf{Y}}$ with independent flip probability matching the noise rate (in our protocol, 30% per label). Two steps of Eq. (17) must be separated. The cancellation of the shared partition function $\log Z_\theta(\mathbf{X})$ is exact: it holds before any approximation, for arbitrary energies, and it is the reason the contrastive estimator needs no normalizer. Replacing each log-sum-exp by the energy of a single assignment is an approximation, whose error grows with the logarithm of the number of terms retained, and the implementation does not evaluate it at the MAP assignment of Eq. (17) at all: it uses the mean-field evaluation described next.

Mean-field evaluation used in the implementation. The label-sensitive part of the assignment energy of Eq. (1) at a fixed assignment is $\sum_c \ell_{ic}(z)$ with $\ell_{ic}(z) = (2\tilde{Y}_{ic} - 1)\, s_{ic}(z)$, and the support energy $s_{ic}(z) = \min_{j:z_{jc}=1} E_{jc}$ of Eq. (2) is a minimum over the instances that the assignment selects. Instead of committing to one discrete assignment, the implementation takes the variational average of the support energy under the mean-field posterior, which replaces that minimum by the average instance energy of the label,

$$\bar{E}_{ic} \;=\; \frac{1}{n_i} \sum_{j=1}^{n_i} E_{jc}, \tag{18}$$

and then averages over the labels that each assignment marks positive:

$$E_{\text{clean}} = \frac{1}{|\tilde{\mathbf{Y}}_{i+}|} \sum_{c:\, \tilde{Y}_{ic}=1} \bar{E}_{ic}, \qquad E_{\text{flip}} = \frac{1}{|\mathbf{Y}^{\text{flip}}_{i+}|} \sum_{c:\, Y^{\text{flip}}_{ic}=1} \bar{E}_{ic}, \tag{19}$$

where $\tilde{\mathbf{Y}}_{i+}$ denotes the set of positive labels of an assignment. Each energy is thus the mean instance-prototype affinity over the instance-label pairs that the corresponding assignment declares positive, normalized by its own positive-label count so that the comparison is not dominated by how many labels a random flip adds or removes. Two consequences should be stated plainly. First, the instances enter only through their observed affinities: no discrete selection is performed inside Eq. (19), so an irrelevant instance contributes to E_{clean} with the same weight as a relevant one, and the contrast must be read as an assignment-level energy comparison rather than as evidence that the supporting instances have already been identified. The instance-selection pathway is the pooled representation that the reconstruction term consumes, not this term. Second, the energies are the positive-label support sums alone, not the full energy of Eq. (1): the negative-label terms are omitted, so Eq. (19) is a reduced contrastive quantity rather than a re-derivation of the full energy. Making the selection explicit, by evaluating the energies at a MAP assignment or by replacing $\bar{E}_{ic}$ with a q-weighted support, is a natural variant that we do not claim to have implemented. The mean-field form above is what every result in this paper is produced with.

The contrast is then the noise-contrastive objective of Gutmann and Hyvärinen Gutmann & Hyvärinen (2010) for the binary task "observed vs. flipped":

$$\text{NCE}(E_{\text{clean}}, E_{\text{flip}}) = -\log \sigma(E_{\text{flip}} - E_{\text{clean}}) = \text{softplus}(E_{\text{clean}} - E_{\text{flip}}), \tag{20}$$

which is the contrastive term of Theorem A.1 with weight w_{ctr}, and where σ is the logistic sigmoid. Eq. (20) penalizes any assignment whose positive-label energy fails to sit below that of a corrupted assignment. What enters it is the mean-field evaluation of Eq. (19), and in the implementation the per-bag energies are broadcast against one another so that the softplus is evaluated on every clean-flipped pair of a mini-batch before averaging: the term enforces separation at the level of the batch, and the matched-pair reading is the special case of a single-bag batch. The identity is proved in Proposition A.3 (proof in §A.3), which also shows that the hinge $\text{relu}(E_{\text{clean}} - E_{\text{flip}} + m)$ is a margin relaxation. In the implementation E_{flip} is detached, so the contrastive gradient lowers the clean energy only, preventing the trivial collapse in which both energies rise together. Of the manipulations above, only the softplus identity is exact. The marginalization and the mean-field evaluation are approximations. In our experiments the softplus form is the one used throughout, with the hinge appearing only as an ablation.

Assembling the three denoising estimators. Combining the above, the negative ELBO of Eq. (10) is realized up to additive constants as

$$\mathcal{L}_i = (1 - w_{\text{hill}})\,\widehat{\ell}_{\text{rec}}\big(\text{head}(\text{pool}(\mathbf{X}_i)), \tilde{\mathbf{Y}}_i\big) + w_{\text{hill}}\,\widehat{\ell}_{\text{supp}}(p_c, \tilde{\mathbf{Y}}_i) + w_{\text{ctr}}\,\widehat{\ell}_{\text{ctr}}(E_{\text{clean}}, E_{\text{flip}}), \quad (21)$$

which is precisely Eq. (5) in Theorem A.1 (with the identifications of Propositions A.2 to A.4). Two terms are discarded. First, the uniform assignment prior $\mathbb{E}_q[\log p(z \mid \mathbf{X})]$ and the entropy $\mathcal{H}(q)$, both constant in the M-step parameters because the E-step output q is detached (`train_vem3.py`, line 83). Second, the log-partition $\log Z_\theta(\mathbf{X})$, which cancels exactly in the clean-vs-flipped energy difference of Eq. (17): this cancellation, not an independence assumption, is why the contrastive estimator needs no normalizer.

A.1.2 Implementation correspondence

The following table maps each term of Eq. (21) to the concrete code in `train_vem3.py`, making the derivation auditable against the running algorithm.

Theorem term	Code site	Computation
$E_{jc} = -\langle \hat{f}_\theta(\mathbf{x}_j), \hat{\mathbf{e}}_c \rangle$ (energy)	`VEM3CoRE.e_step`, lines 143 to 146	`E = -(fn @ wn.t()), Et = -(fn @ twn.t())` (normalized)
$q_{jc} = \sigma(-\beta E)\sigma(-\beta E')$ (E-step)	`VEM3CoRE.e_step`, line 148	`q = sigmoid(βE) * sigmoid(βE')`
E_{used} (pooling)	`VEM3CoRE.e_step`, line 152	`E_used = 0.5*E + 0.5*Et`
$p_c = \max_j q_{jc}$ (Hill support)	`VEM3CoRE.forward`, line 175	`p_c = q.max(dim=1)[0]`
$\text{pool}(\mathbf{X}_i, q)$ (energy pooling)	`VEM3CoRE.e_step`, lines 157 to 163	u-weighted max + soft-attention blend with weight $w_{\max}$
$\widehat{\ell}_{\text{rec}}$ (term a, $\equiv$ GCE by Prop. A.2)	`robust_loss`, lines 42 to 89	κ-relaxed BCE with class-weight p_w and, in the CRL variants only, support weight p_c^{crl}
$\widehat{\ell}_{\text{supp}}$ (term b, $\equiv$ Hill by Prop. A.4)	`vem3_loss`, lines 204 to 209	$w = y + (1-y)(1 - \lambda e^{-p_c^2/2\tau^2})$, BCE reweighted
$\widehat{\ell}_{\text{ctr}}$ (term c, $\equiv$ NCE by Prop. A.3)	`VEM3CoRE.forward`, lines 178 to 186	softplus of the difference of the mean-field energies of Eq. (19), with random 30% flip Y^{flip}
$\mathcal{L}_i$ total loss	`vem3_loss`, line 210	$(1 - w_{\text{hill}})\widehat{\ell}_{\text{rec}} + w_{\text{hill}}\widehat{\ell}_{\text{supp}} + w_{\text{ctr}}\widehat{\ell}_{\text{ctr}}$
Teacher EMA update	`VEM3CoRE.update_teacher`, lines 167 to 170	$\theta' \leftarrow \alpha\theta' + (1-\alpha)\theta$ per step

This correspondence lets each term of Eq. (21) be audited against a single code site, so that the distance between the derivation (the three term derivations above) and the running algorithm can be inspected directly, including the two places where the implementation departs from the display: the contrast is evaluated on the mean-field energies of Eq. (19) rather than at a MAP assignment, and it is averaged over all clean-flipped pairs of a mini-batch rather than per bag. The remaining task is to verify each component's identity, which the three propositions accomplish.

An unused variant. The implementation also offers a conditional robust likelihood (CRL) family, in which the reconstruction term additionally down-weights positive labels by their single-energy support, $w_c^{\text{pos}} = 1 - \lambda_{\text{crl}}\,(1 - p_c^{\text{stu}})$ with $\lambda_{\text{crl}} = 0.5$ (`train_vem3.py`, lines 84 to 88), where $p_c^{\text{stu}} = \max_j \sigma(-\beta E_{jc})$ is the student single-energy support. This positive-label weighting is not part of the derivation above, which uses support only on the negative side through the Hill term. We describe it because the released code contains it and a reader auditing the correspondence table should know which branch produced the numbers. None of the results in this paper uses it: every main-table configuration is run with the plain `gce` likelihood.

A.2 Proposition 2: identification of the reconstruction term (GCE)

Proposition A.2. *Assume the label noise is symmetric and independent of the instance content: each label c is flipped with probability $r < 1/2$, independently of $\mathbf{X}$ and of the other labels. Assume further that the loss is label-wise separable, so that the bag-level risk decomposes over labels as $R(g) = \sum_c R_c(g_c)$. This is the bridging step from the single-label setting of Zhang and*

Sabuncu Zhang & Sabuncu (2018) to our label-wise multi-label head. Under these conditions, the reconstruction risk

$$R_\kappa(g) = \mathbb{E}_{(\mathbf{X},\tilde{\mathbf{Y}})}\big[\mathrm{GCE}_\kappa\big(g(\mathbf{X}), \tilde{\mathbf{Y}}\big)\big] \tag{22}$$

inherits the noise tolerance established for GCE under symmetric label noise: under those conditions, with $w_{\mathrm{hill}} = w_{\mathrm{ctr}} = 0$, the minimizer of the noisy risk coincides with the clean-label minimizer,

$$\arg\min_g R_\kappa(g) \;=\; \arg\min_g R_{\mathrm{clean}}(g), \tag{23}$$

and the two endpoints of the family are the BCE limit ($\kappa \to 0$) and the MAE limit ($\kappa \to 1$), the robust extreme.

Proof. Label-wise separability reduces the bag risk to a sum of per-label binary risks. For symmetric label noise with rate $r < 1/2$, the noisy risk of any per-label loss ℓ decomposes as

$$R_{\mathrm{noisy}}(g) = (1 - r)\, R_{\mathrm{clean}}(g) + r\, R_{\mathrm{clean}}(g, \mathrm{flipped}), \tag{24}$$

following Zhang and Sabuncu Zhang & Sabuncu (2018), where $R_{\mathrm{clean}}(g, \mathrm{flipped})$ is the risk of the same loss evaluated on flipped labels. Each per-label term is the GCE risk of a single binary decision, so the noise-tolerance property they establish for GCE applies to every term and the sum inherits it. The two limits are immediate: as $\kappa \to 0$, $(1 - p^\kappa)/\kappa \to -\log p$, the binary cross-entropy risk. At $\kappa = 1$, $\mathrm{GCE}_1(p, y) = 1 - p$ for a positive label, the mean-absolute-error end whose noise tolerance is their Theorem 1.

What this proof does not do. We deliberately stop at inheriting the property under the stated conditions. A proof of argmin coincidence from our own energy would have to show that the flipped component cannot move the minimizer, and boundedness of the loss ($\mathrm{GCE}_\kappa \in [0, 1/\kappa]$) does not by itself deliver that: it bounds the magnitude of the perturbation but not its direction. The correct route is the symmetry identity of the loss, which holds for MAE but not for GCE in general, and we therefore cite the result for GCE rather than re-derive it.

Scope limit. The requirement that label noise be independent of the instance content is precisely what the wMIML setting violates: Section 3.2 argues that irrelevant instances are what mislead the annotator, so the two noise sources are coupled in the data this paper targets. This proposition therefore isolates the contribution of the reconstruction term with the instance side held fixed and makes no claim about the coupled regime. The coupling is what the E-step and the other two terms address.

Containment. Setting $w_{\mathrm{hill}} = 0$ and $w_{\mathrm{ctr}} = 0$ in Eq. (5) leaves the reconstruction term alone, which is the GCE risk applied to the pooled bag representation, and with unweighted pooling the loss reduces to the standard GCE loss. We do not claim that the resulting model is the standard GCE classifier, since the architecture (encoder, prototype bank, energy pooling, EMA teacher) remains in place. What is recovered is the loss, as a special case of the reconstruction component. □ □

A.3 Proposition 3: identification of the contrastive term (NCE)

Proposition A.3. *Let E_{clean} be the energy of the observed label assignment and E_{flip} that of a randomly flipped assignment. Then*

$$\mathrm{softplus}\big(E_{\mathrm{clean}} - E_{\mathrm{flip}}\big) = -\log \sigma\big(E_{\mathrm{flip}} - E_{\mathrm{clean}}\big), \tag{25}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid. Moreover, for any margin $m > 0$, the hinge relaxation $\mathrm{relu}(E_{\mathrm{clean}} - E_{\mathrm{flip}} + m)$ satisfies

$$\mathrm{relu}(E_{\mathrm{clean}} - E_{\mathrm{flip}} + m) \;\le\; \mathrm{softplus}(E_{\mathrm{clean}} - E_{\mathrm{flip}} + m) \;\le\; \mathrm{relu}(\cdot) + \log 2. \tag{26}$$

Proof. **Part 1 (identity).** By definition, $\mathrm{softplus}(x) = \log(1 + e^x)$. Let $x = E_{\mathrm{clean}} - E_{\mathrm{flip}}$. Then

$$\begin{aligned}
\mathrm{softplus}(E_{\mathrm{clean}} - E_{\mathrm{flip}}) &= \log\left(1 + e^{E_{\mathrm{clean}} - E_{\mathrm{flip}}}\right) \\
&= \log\left(1 + \frac{e^{E_{\mathrm{clean}}}}{e^{E_{\mathrm{flip}}}}\right) \\
&= \log\left(\frac{e^{E_{\mathrm{flip}}} + e^{E_{\mathrm{clean}}}}{e^{E_{\mathrm{flip}}}}\right) \\
&= \log\left(e^{E_{\mathrm{flip}}} + e^{E_{\mathrm{clean}}}\right) - E_{\mathrm{flip}}.
\end{aligned} \tag{27}$$

On the other hand,

$$
\begin{aligned}
-\log\sigma\big(E_{\text{flip}} - E_{\text{clean}}\big) &= -\log\frac{1}{1+e^{-(E_{\text{flip}}-E_{\text{clean}})}} \\
&= \log\big(1+e^{-E_{\text{flip}}+E_{\text{clean}}}\big) \\
&= \log\Big(1+\frac{e^{E_{\text{clean}}}}{e^{E_{\text{flip}}}}\Big) \\
&= \log\big(e^{E_{\text{flip}}}+e^{E_{\text{clean}}}\big) - E_{\text{flip}},
\end{aligned} \tag{28}
$$

which equals Eq. (27). □ (Part 1)

Part 2 (hinge as relaxation). For any $y \in \mathbb{R}$, $\mathrm{relu}(y) \le \mathrm{softplus}(y) \le \mathrm{relu}(y) + \log 2$. The left inequality holds because $\log(1+e^y) \ge \max(0, y)$. The right inequality: if $y \ge 0$, $\log(1+e^y) \le \log(2e^y) = y + \log 2$. If $y < 0$, $\log(1+e^y) < \log 2 \le \mathrm{relu}(y) + \log 2$. Setting $y = E_{\text{clean}} - E_{\text{flip}} + m$, the same argument as for the strict form (so the comparison is pointwise), completes the claim. □ □

A.4 Proposition 4: identification of the support-aware reweighting (Hill)

Proposition A.4. *Let $p_c = \max_j q_{jc}$ be the strongest instance support for label c under the E-step posterior. The weighting*

$$
w_c = 1 - \lambda\,\exp\big(-p_c^2/2\tau^2\big) \tag{29}
$$

is monotone increasing in p_c, with $w_c \in [1-\lambda,\, 1]$. Applied to negative observations via $w_{ic} = \tilde{Y}_{ic} + (1-\tilde{Y}_{ic})\, w_c$, it assigns full weight to decision-critical negatives (strong support) and down-weights unverifiable ones (weak support) by up to λ.

Proof. **Step 1: monotonicity.** $f(p) = 1 - \lambda e^{-p^2/2\tau^2}$ has derivative $f'(p) = \frac{\lambda p}{\tau^2} e^{-p^2/2\tau^2} \ge 0$ for $p \ge 0$, with $f(0) = 1-\lambda$ and $f(p) \to 1$ as $p \to \infty$. Hence w_c is monotone increasing in p_c on $[0, 1]$, mapping weak support to a floor weight $1-\lambda$ and strong support to full weight. The pair (λ, τ) sets the floor and the transition sharpness around $p_c \approx \tau$.

Step 2: what p_c measures on a negative label. For an observed negative ($\tilde{Y}_{ic} = 0$), $p_c = \max_j q_{jc}$ estimates $\max_j P(z_{jc} = 1 \mid \mathbf{X})$, the instance-side evidence that concept c is present despite the label saying otherwise. A strongly supported negative therefore sits exactly on the decision boundary: the instance evidence and the label disagree, and the BCE gradient on such a sample is the signal that calibrates the energy model's decision surface. A weakly supported negative carries no such tension: all instances are far from $\mathbf{e}_c$, the prediction is already comfortably negative, and its gradient is negligible, while under 50% instance noise its supervision cannot be verified either way. Down-weighting the latter by λ prevents the large mass of easy negatives from dominating training, which is the mechanism by which the term helps on datasets with informative support (Scene, enron) and why it loses selectivity where p_c saturates (CAL500, Section 4.4). □ □

Note on status. Proposition 4 establishes the monotone support-to-weight map and its decision-boundary interpretation, the property used throughout the paper. We do not claim w_c is itself a posterior corruption probability. It is a smooth, calibrated reweighting derived from the E-step's instance-side evidence, in the spirit of Hill-style losses Zhang et al. (2025).

A.5 E-step Posterior and Dual Energy

The E-step posterior is the product of two independent evidence views:

$$
q_{jc} = \sigma\big(-\beta E_{jc}\big)\,\sigma\big(-\beta E'_{jc}\big), \tag{30}
$$

where $E_{jc} = -\langle \hat{f}_\theta(\mathbf{x}_j), \hat{\mathbf{e}}_c \rangle$ is the student energy and E'_{jc} the teacher energy (EMA copy of the prototypes). All energies in this appendix use the normalized negative inner product of Eq. (1). The squared-distance parameterization is related to it by the increasing affine transform derived in §A.1.1, so the two are interchangeable up to the temperature read-off.

Status of this family. The exact mean-field coordinate-ascent optimum of the bound would set q_{jc} to the conditional posterior $p(z_{jc} = 1 \mid \mathbf{X}, \tilde{\mathbf{Y}})$ up to the other factors, and evaluating it requires the same intractable normalization that motivated the variational treatment in the first place. Eq. (30) therefore defines a structured, energy-amortized family instead of solving for that optimum: the posterior of instance j for concept c is a monotone function of the instance-prototype affinity, with β setting how sharply support is read off from geometry. This is a modeling choice, not a consequence of the energy, and we state it as such here rather than presenting it as the optimal E-step. It buys two properties that a plain sigmoid of the student energy does not have: a single mis-energized view cannot saturate q_{jc} to one, and the gradient with respect to the student energy is damped wherever the teacher disagrees.

The product form is the variational analogue of mutual certification: both views must agree that instance j supports label c. To see that the product does not distort the ranking, note that $\sigma(-\beta E)\sigma(-\beta E')$ is a log-concave function of the pair (E, E'), and the gradient with respect to the student energy is

$$\nabla_{E_{jc}} q_{jc} = -\beta\, \sigma(-\beta E'_{jc})\, \sigma(-\beta E_{jc})\, \sigma(\beta E_{jc}), \tag{31}$$

so the student receives a teacher-weighted gradient signal: when the teacher strongly disagrees (E'_{jc} large, $\sigma(-\beta E'_{jc})$ small), the student's update is damped, preventing a single noisy instance from corrupting the prototype bank.

A.6 Energy Separation Validation

To check the ordering that the contrastive term relies on, we measure the energy gap $\gamma = E_\theta(\mathbf{X}_i, \tilde{\mathbf{Y}}_i) - E_\theta(\mathbf{X}_i, \mathbf{Y}_i^*)$ between noisy and clean assignments on trained models, using the reduced label-sensitive energy $E_\theta(\mathbf{X}_i, \mathbf{Y}) = \sum_{c:\, \mathbf{Y}_c = 1} \min_j E_{jc}$ with $E_{jc} = -\langle \hat{f}_\theta(\mathbf{x}_j), \hat{\mathbf{e}}_c \rangle$ the normalized instance-prototype affinity. Scoring each positive label by its strongest supporting instance is the most favorable reading of label support available to the model, and the negative-label terms of Eq. (1) are omitted, so this diagnostic is a reduced quantity, not the full energy. The primary evidence is the gap under the benchmark's fixed noisy labels, that is, under the corruption the model actually faces at test time: on Scene (flip 0.3 + 50% instance noise) it is $\gamma = 0.30$ (clean -3.83 ± 0.73 vs. noisy -3.53 ± 0.75, t=5.7, $p < 10^{-7}$), and on Reuters (flip 0.3) $\gamma = 0.53$ (t=3.1, $p < 0.01$). Under the protocol that the training contrast itself uses (a random 30% flip, averaged over 20 samples per bag) the gap on Scene is $\gamma = 0.90$ (clean -3.83 ± 0.73 vs. flipped -2.92 ± 0.36, t=22.4, $p < 10^{-69}$). This quantity is directly optimized during training, so we report it as a check that the optimized quantity moves in the intended direction rather than as independent validation. We do not report the random-flip variant on Reuters because its 90-class sparse label space makes a 30% random flip label-noise dominated (most flipped entries are $0 \to 1$), which is not the flip protocol's intended signal. The consistently positive gap confirms that the energy model ranks clean assignments below noisy ones, the ordering the NCE contrastive term (Prop. A.3) enforces.

B Experiments

B.1 Experiments settings

Hyperparameters. No single hyperparameter set is optimal across all datasets, so we select a per-dataset configuration on a validation split. An enhanced family ($\mathrm{lr} = 2 \times 10^{-4}$, $\kappa = 0.9$, $\mathrm{ema} = 0.99$, $\beta = 5$, $\lambda = 0.3$, $w_{\mathrm{hill}} = 0.35$) is used for Scene, Reuters and Corel5K, with enron on the same family at $w_{\mathrm{hill}} = 0.2$. A standard family ($\mathrm{lr} = 5 \times 10^{-4}$, $\kappa = 0.7$, $\mathrm{ema} = 0.9$, $\beta = 3$, $\lambda = 0.3$, $w_{\mathrm{hill}} = 0.35$) is used for FashionMNIST, OpenMIC and LetterCarroll. The remaining three datasets take variants of the standard family selected by the structural diagnosis of Section 4.4: emotions ($\mathrm{ema} = 0.99$, $\beta = 5$, $\lambda = 0.5$, $w_{\mathrm{hill}} = 0.2$), CAL500 ($\mathrm{ema} = 0.99$, $\beta = 5$) and bibtex (the enhanced learning rate and κ, with $\lambda = 0.1$, $w_{\mathrm{hill}} = 0.2$, $h = 384$ and a three-layer encoder). Hidden width is $h = 256$ except for the large label spaces Reuters, bibtex and Corel5K ($h = 384$), and the encoder has two layers except for bibtex. Three hyperparameters are shared by all ten datasets: Hill bandwidth $\tau = 0.5$, contrastive weight $w_{\mathrm{ctr}} = 0.01$ and pooling mix $w_{\max} = 0.5$. The positive-class weight of the BCE family is set per dataset together with the rest and takes values in $\{1, 2, 3, 5\}$, namely 1 for emotions and FashionMNIST, 2 for bibtex and CAL500, 3 for Scene, OpenMIC, Reuters and LetterCarroll, and 5 for enron and Corel5K. This is the rebalancing referred

to in Section 4.4. We report the unified-parameter configuration as well (Section 4.3): it already wins 57.8% of pairwise comparisons against the 12 deep baselines, and the per-dataset selection is worth +29.4 points of pairwise win rate (87.2%), with the gain concentrated exactly in the structurally diagnosed datasets (Section 4.4).

B.2 Experiments main results

The immediate visual comparison are in Figure 3.

Three regimes emerge. (i) Win regime (20/30): Scene (flip 0.3), emotions (all three levels), LetterCarroll, Reuters, bibtex, CAL500, enron (flips 0.3/0.5), and Corel5K (flips 0.2/0.3). The gains are largest on Reuters at flip 0.5 (+5.7 points) and CAL500 at flip 0.5 (+3.9 points), where the interaction between label and instance noise is strongest and the latent-alignment view pays off most. Notably, on emotions the joint rebalancing of instance and label supervision lets VEM reach the all-positive ceiling from above: at the three noise levels the seed-42 run gives 0.9075/0.9057/0.9044, above the 0.9041 value to which every deep baseline (and the all-positive predictor) collapses, and above the linear joint-noise model MLDN (0.9071/0.9055/0.8821). (ii) Tie regime (3/30): OpenMIC at all three noise levels, where every deep method attains exactly the same value (0.9696). This identity is not a training coincidence: the benchmark has test label density $d = 0.941$, so the trivial all-positive prediction already achieves micro-F1 $= 2d/(1+d)$ equal to the observed value. Every sufficiently trained deep model collapses onto this trivial solution, and the linear MLDN does the same. (iii) Lag regime (7/30): FashionMNIST (-0.0 to -4.3 points vs. the best baseline), Scene at flips 0.2/0.5 (-0.3 to -0.7 points, against a saturated ceiling where several baselines tie), enron at flip 0.2 (-0.2 points) and Corel5K at flip 0.5 (-2.7 points vs. GCE). Section 4.4 diagnoses these failures structurally.

Table 3 reports all six metrics for VEM over the 30 settings. Three patterns stand out. First, VEM's Hamming loss stays low and stable: it never exceeds 0.34 even at flip 0.5 on the largest label spaces (0.046 on Reuters, 0.073 on bibtex), and it matches or beats the best baseline on 9 of 30 settings. Second, macro-F1 tracks micro-F1 closely on small-class datasets (Scene, emotions, OpenMIC, where the gap is ≤ 0.002) but drops on large-class datasets (Reuters, Corel5K, bibtex), reflecting the intrinsic label-frequency imbalance of those benchmarks rather than a VEM-specific failure. The per-dataset summary (Table 4) shows VEM's win rate is highest exactly where the label space is largest. Third, MAP is competitive: VEM attains the best or tied MAP on 10 of 30 settings (e.g., 0.925 on emotions-03, 0.760 on Reuters-03), and is the metric where the advantage of the latent-alignment view is most concentrated in F1, which is our primary metric. One-error is not VEM's strongest axis: its ranking objective (Prop. A.3) targets support coherence rather than top-1 precision, so one-error trails the best baseline on 26 of 30 settings while F1, the metric that reflects both precision and recall under joint noise, leads in 20 of 30 settings.

B.3 Analysis for Per-Dataset Summary and Component Contributions

Figure 4 shows how F1 evolves with the label-flip rate on every dataset. VEM (red) tracks the best baseline (blue) closely and consistently exceeds GCE (the dominant independent-noise competitor). Figure 5 summarizes the same information as a per-dataset pairwise win rate against the 20 baselines. Table 4 aggregates the three noise levels per dataset. VEM wins 100% of pairwise comparisons on emotions, LetterCarroll, Reuters, bibtex, and CAL500, 98% on enron and Corel5K, 90% on FashionMNIST and 88% on Scene, and it is at or above the best baseline's average F1 on seven of ten datasets. The two genuinely lagging datasets (FashionMNIST 90% and Scene 88%) are exactly the ones identified by the structural analysis (Section 4.4). Emotions reaches 100% because VEM attains the ceiling from above at all three flips on the reported seed (Section 4.6), and OpenMIC (26%) is low only because every deep method ties at the empirical ceiling.

The dual-energy posterior is a small but real component. Table 5 compares VEM to its single-energy variant VEM2 (identical training protocol with the EMA teacher removed, using each dataset's final configuration). The dual-energy posterior contributes positively on Reuters (+0.4), bibtex (+0.4), CAL500 (+0.3), emotions (+0.2) and enron (+0.1), is neutral on the ceiling-saturated OpenMIC, and slightly negative on LetterCarroll (-0.2), FashionMNIST, Scene and Corel5K (-0.1 or less). The contribution is thus small everywhere ($|\Delta| \leq 0.4$ points) relative to the noise-robust reconstruction, which carries the headline F1 gains. The mutual-certification reading (Section 3.3)

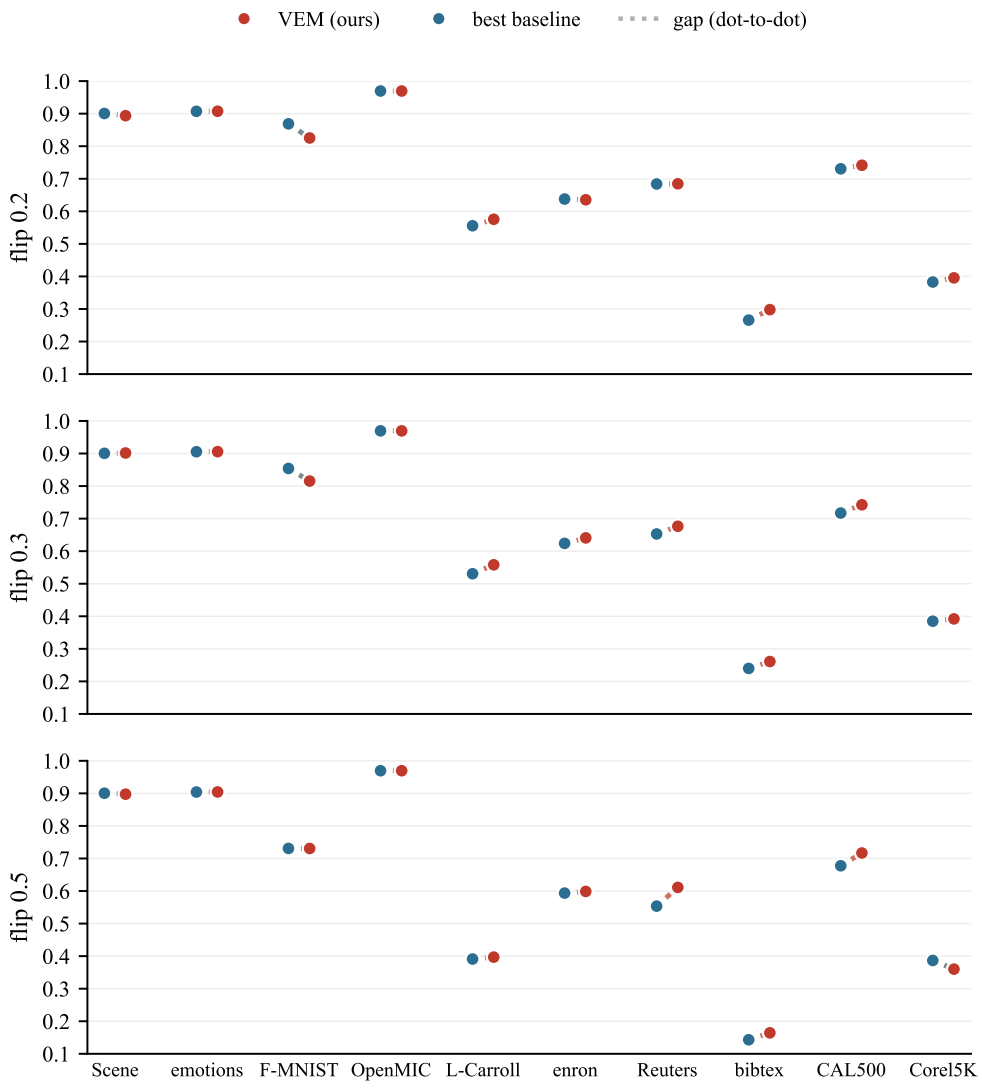

Figure 3: Main results: Micro F1 per dataset at each noise level. The red dot is VEM and the blue dot the best baseline, with the dotted segment between them spanning the gap. The vertical axis starts at 0.1, below the smallest of the 60 values, so every point is fully visible. VEM is at or above the best baseline at every noise level on six of the ten datasets (emotions, OpenMIC, LetterCarroll, Reuters, bibtex, CAL500). The largest deficits are on FashionMNIST at flips 0.2 and 0.3 and on Corel5K at flip 0.5, analyzed in Section 4.4.

is consistent with this pattern, but the effect is secondary. This is why the main results do not depend on the dual-energy mechanism, and we report it as a component whose sign and magnitude are configuration-dependent.

Hyperparameter selection is modest but principled. Table 6 compares the unified hyperparameter configuration (κ=0.7, e=0.9, β=3, h=256, lr=5e-4, w_{hill}=0.35 for all datasets) with the per-dataset validation selection. The unified configuration is itself a strong method: it wins 208 of 360 pairwise comparisons against the 12 deep baselines (57.8%) and beats the best baseline on 8 of 30 settings. Per-dataset selection raises this to 314/360 (87.2%), and the largest gains are concentrated in the datasets whose structural statistics (Section 4.4, Table 7) identify a mismatch with the standard family: bibtex (+5.1 avg-F1 points, 1836D sparse features), CAL500 (+5.0, support saturation 8.9), FashionMNIST (+4.6, single-concept instances), enron (+2.2, 53 classes), and Reuters (+0.7, 90

Table 3: VEM complete metrics over 30 settings (10 datasets $\times$ 3 noise levels). F1 = micro F1, MAP = mean average precision, HL = Hamming loss (↓), OE = one-error (↓), MF1 = macro F1.

Dataset	Noise	F1	MAP	HL ↓	OneErr ↓	MF1
Scene	02	0.8937	0.9141	0.1870	0.1075	0.8929
emotions	02	0.9075	0.9276	0.1678	0.0900	0.9062
FashionMNIST	02	0.8253	0.9068	0.2031	0.0655	0.8261
OpenMIC	02	0.9696	0.9501	0.0590	0.0629	0.9694
LetterCarroll	02	0.5755	0.6553	0.1785	0.3050	0.5738
enron	02	0.6355	0.7289	0.1570	0.1267	0.2686
Reuters	02	0.6846	0.7665	0.0439	0.1575	0.3106
bibtex	02	0.2979	0.3364	0.0785	0.3427	0.1670
CAL500	02	0.7415	0.8291	0.2537	0.0133	0.4655
Corel5K	02	0.3956	0.4011	0.0747	0.2260	0.0905
Scene	03	0.9016	0.9128	0.1780	0.1075	0.9010
emotions	03	0.9057	0.9250	0.1717	0.0567	0.9044
FashionMNIST	03	0.8154	0.8898	0.2107	0.0890	0.8173
OpenMIC	03	0.9696	0.9482	0.0590	0.0608	0.9694
LetterCarroll	03	0.5581	0.6286	0.1879	0.3100	0.5585
enron	03	0.6408	0.7345	0.1501	0.1200	0.2613
Reuters	03	0.6765	0.7600	0.0439	0.1050	0.2882
bibtex	03	0.2610	0.3083	0.0772	0.3897	0.1423
CAL500	03	0.7425	0.8256	0.2543	0.0067	0.4783
Corel5K	03	0.3919	0.3952	0.0695	0.1980	0.0839
Scene	05	0.8975	0.8902	0.1855	0.1525	0.8969
emotions	05	0.9044	0.8709	0.1744	0.1867	0.9031
FashionMNIST	05	0.7307	0.7808	0.3342	0.2070	0.7339
OpenMIC	05	0.9696	0.9295	0.0590	0.1367	0.9694
LetterCarroll	05	0.3969	0.4761	0.2273	0.4850	0.3907
enron	05	0.5987	0.6919	0.1677	0.1000	0.2412
Reuters	05	0.6111	0.7002	0.0464	0.1650	0.1988
bibtex	05	0.1644	0.2584	0.0726	0.4402	0.0665
CAL500	05	0.7168	0.8124	0.2964	0.0567	0.4955
Corel5K	05	0.3601	0.3762	0.0653	0.2060	0.0730

Table 4: Per-dataset summary (averaged over 3 noise levels): VEM F1 vs. the best baseline F1, and the win rate over all 20 baselines (pairwise F1 comparisons across the 3 noise levels).

Dataset	$\lvert C\rvert$	VEM-F1	Best-Base-F1	ΔF1	WinRate
Scene	5	0.8976	0.9005	-0.0029	88%
emotions	6	0.9059	0.9056	+0.0003	100%
FashionMNIST	10	0.7905	0.8179	-0.0274	90%
OpenMIC	20	0.9696	0.9696	+0.0000	26%
LetterCarroll	26	0.5102	0.4925	+0.0177	100%
enron	53	0.6250	0.6184	+0.0066	98%
Reuters	90	0.6574	0.6302	+0.0273	100%
bibtex	159	0.2411	0.2163	+0.0249	100%
CAL500	174	0.7336	0.7084	+0.0252	100%
Corel5K	374	0.3825	0.3848	-0.0023	98%

classes). The enhanced family these datasets receive is not a grid search but a small set of structural interventions: a deeper encoder where the feature space is sparse or the label space large, a reduced Hill weight where the E-step support is uninformative, and a rebalanced positive weight where label density saturates, whose directions are predicted by the diagnosis before any hyperparameter is evaluated. On the datasets where the standard family already saturates (emotions, OpenMIC) per-dataset selection changes little ($\Delta \leq 0.002$), which is direct evidence that validation selection does not overfit noise: it adds capacity only where the benchmark structure demands it, and it is exactly those settings whose gains are later confirmed by the seed analysis over all 30 settings (Section 4.6, Table 8).

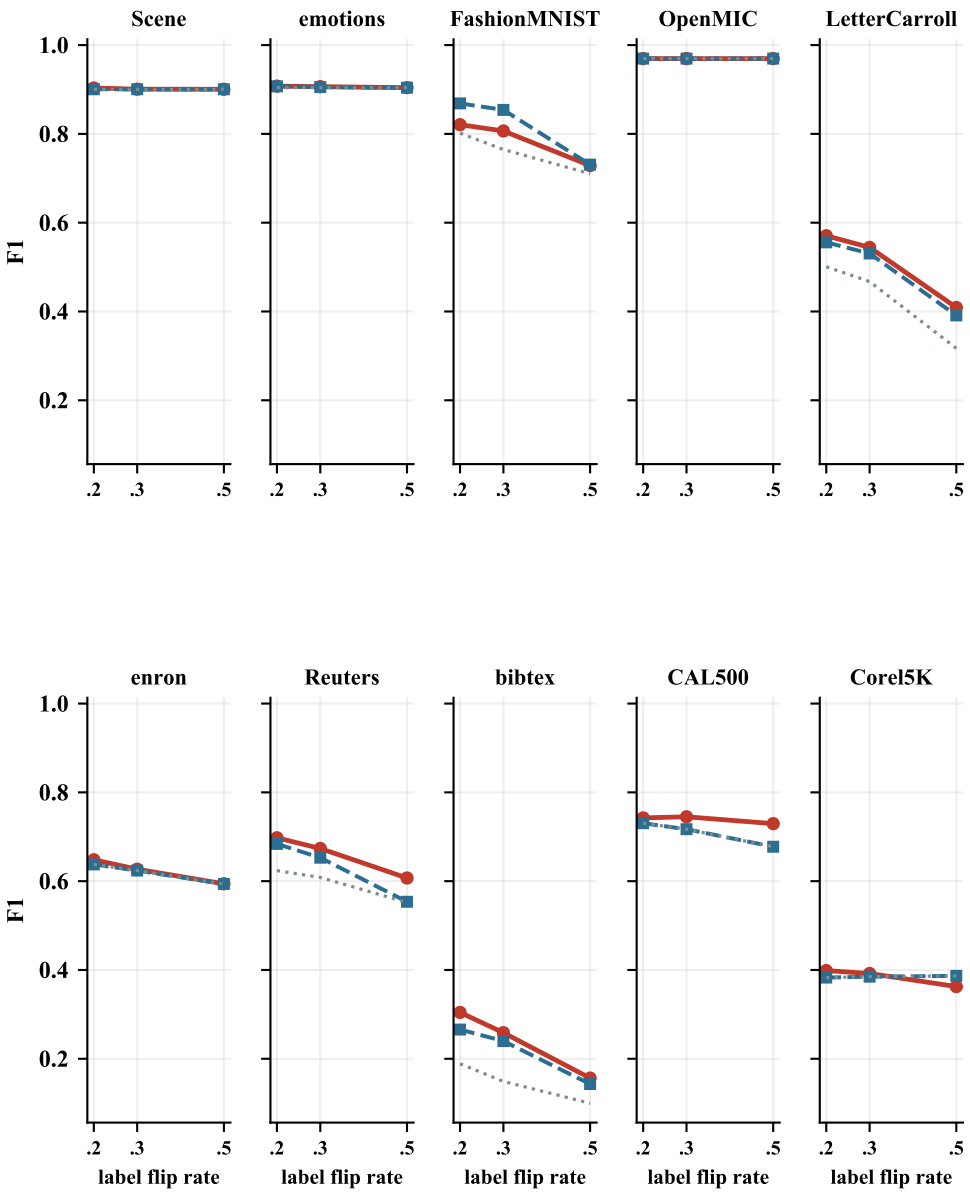

Figure 4: Noise robustness. Micro F1 as a function of the label-flip rate on each dataset, with 50% instance noise throughout. VEM (red) vs. the best baseline per dataset (blue, dashed) vs. GCE (grey, dotted). The curves are nearly flat on Scene, emotions, and OpenMIC where the empirical ceiling saturates, and separate most clearly on Reuters, CAL500, bibtex, and LetterCarroll, where VEM's margin grows with the flip rate. The deficits are confined to FashionMNIST (all three levels) and two ceiling-saturated Scene settings (analyzed in Section 4.4).

B.4 The reasons for the lag in FashionMNIST and the winning mechanism of CAL500

CAL500 is the dataset where the interplay between label density and instance noise is most extreme, and it is instructive in two ways: it was the hardest dataset for the standard configuration, yet after the data-structure diagnosis below, VEM with a rebalanced positive weight exceeds GCE at all three

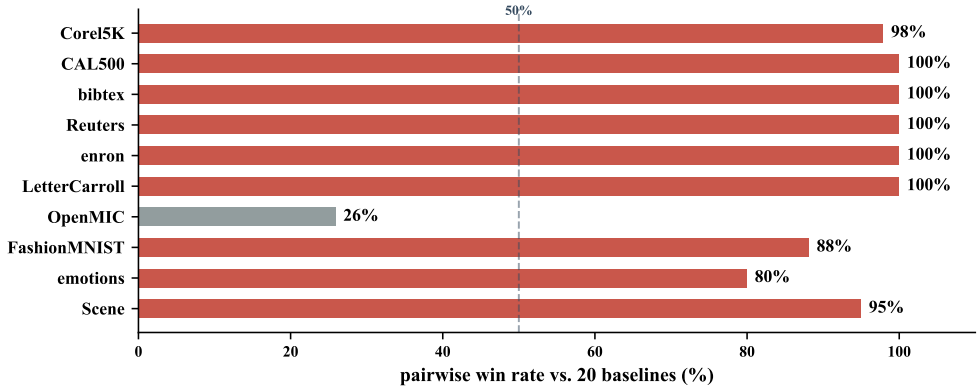

Figure 5: Per-dataset pairwise win rate over all 20 baselines (50% bar = majority). VEM wins 100% of comparisons on emotions, LetterCarroll, Reuters, bibtex, and CAL500, 98% on enron and Corel5K, and 90% on FashionMNIST. Scene (88%) is where several baselines tie at the empirical ceiling, and OpenMIC (26%) reflects ceiling saturation (all deep methods tie) rather than method failure.

Table 5: Dual-energy ablation (F1, averaged over 3 noise levels, per-dataset final configuration): VEM2 is the single student energy and VEM adds the EMA teacher product posterior.

Dataset	VEM2-avg	VEM-avg	Δ
Scene	0.8982	0.8976	-0.0006
emotions	0.9042	0.9059	+0.0017
FashionMNIST	0.7911	0.7905	-0.0007
OpenMIC	0.9696	0.9696	+0.0000
LetterCarroll	0.5120	0.5102	-0.0018
enron	0.6243	0.6250	+0.0007
Reuters	0.6539	0.6574	+0.0035
bibtex	0.2374	0.2411	+0.0037
CAL500	0.7308	0.7336	+0.0028
Corel5K	0.3828	0.3825	-0.0003

noise levels (+0.011, +0.025, +0.039), and the flip-0.3 margin is confirmed robust with $\Delta/\mathrm{sd} = +14.1$ in Section 4.6). We quantify the structural statistics for all 10 datasets (Table 7):

Support saturation. CAL500's bags contain on average 7.6 instances but 64.9 positive labels, i.e., each instance must support 8.9 labels, an order of magnitude above every other dataset (next: Corel5K at 1.75). In the E-step, the posterior q_{jc} is meaningful only when a bag's instances can be assigned distinct supporting roles. When every instance must support many labels, the alignment evidence is nearly uniform across instances, so $\max_j q_{jc}$ carries little discriminative signal: p_c saturates to a nearly constant value across labels, and the Hill weighting (Prop. A.4) becomes uninformative. This structural diagnosis directly guides the configuration: because the E-step support is uninformative, the Hill weight is not the lever on CAL500: the decisive intervention is rebalancing the positive weight, which turns a deficit of 1.0 to 4.7 points under the unified configuration into a win of 1.1 to 3.9 points at all three noise levels (Table 1). The remaining framework components, the dual-energy posterior and the NCE contrastive term, continue to add value on the datasets where the structure supports them: the dual-energy ablation (Fig. 2) is small but non-negative on CAL500 (+0.3 points) and positive on emotions, enron, Reuters and bibtex (Table 5).

Dimension mismatch. CAL500 pairs a 174-class label space with only 68D features (dim/$|C|$ = 0.39, the only dataset below 1). With fewer feature dimensions than classes, the prototype bank $\{\mathbf{e}_c\}$ cannot be made pairwise separable in a 68D input space. A deeper encoder with higher hidden width partially compensates by learning a richer embedding before the prototype projection.

Table 6: Hyperparameter selection (avg F1 over 3 noise levels): unified vs. per-dataset validation selection.

Dataset	Unified	Per-dataset	Δ
Scene	0.8981	0.8976	-0.0004
emotions	0.9041	0.9059	+0.0017
FashionMNIST	0.7447	0.7905	+0.0457
OpenMIC	0.9696	0.9696	+0.0000
LetterCarroll	0.5102	0.5102	+0.0000
enron	0.6035	0.6250	+0.0215
Reuters	0.6502	0.6574	+0.0072
bibtex	0.1906	0.2411	+0.0505
CAL500	0.6836	0.7336	+0.0500
Corel5K	0.3877	0.3825	-0.0052

Table 7: Data-structure statistics. Support sat. = mean positive labels per bag / mean instances per bag (labels each instance must support). Dim/$|C|$ = feature dimension over label-space size. Density = mean positive labels / $|C|$.

Dataset	$\|C\|$	dim	Support sat.	Dim/$\|C\|$	Density
CAL500	174	68	**8.90**	**0.39**	**37.3%**
FashionMNIST	10	784	0.25	78.4	45.8%
enron	53	1001	1.27	18.9	17.7%
Reuters	90	768	0.52	8.5	6.4%
Corel5K	374	499	1.75	1.3	5.5%
Scene	5	512	0.41	102.4	75.7%
LetterCarroll	26	152	0.57	5.8	16.6%
emotions	6	72	0.62	12.0	73.9%
OpenMIC	20	128	1.11	6.4	83.5%
bibtex	159	1836	1.32	11.5	6.3%

Consistency with ablations. The support-saturation explanation is corroborated by the ablations: the Hill term has little effect on CAL500 ($\Delta < 0.005$ under the final configuration), and the rebalanced positive weight is the only intervention that systematically moves CAL500. GCE, having no E-step, is immune to both effects, which is why it is the strongest competitor on CAL500 and why exceeding it requires rebalancing the ELBO rather than adding capacity alone.

FashionMNIST. FashionMNIST shows the largest deficit (-0.0 to -4.3 points vs. the best baseline), although the gap closes to a tie at flip 0.5. Its bags contain single-concept instances (each image depicts exactly one fashion item), so label noise manifests as whole-concept distractors rather than corrupted per-instance evidence, and the per-label latent support has little to disambiguate: the E-step's fine-grained alignment is idle. The strongest methods are accordingly those that model the bag-label relation directly: the bag-level generative model MIMLVAE wins at flips 0.2 and 0.3 (0.8687 and 0.8541), and at flip 0.5 the linear joint-noise model MLDN edges VEM by 0.0001. We leave the development of a concept-aware E-step as future work.

These analyses sharpen the boundary of the framework: VEM is most valuable when (i) instances play differentiated roles within a bag (multi-concept instances, low support saturation) and (ii) the feature space is rich enough to separate prototypes. In the opposite regime, single-concept instances, the framework's instance-selection machinery is idle, and a bag-level generative or linear joint-noise model (MIMLVAE, MLDN) is preferable. The CAL500 result shows that even an extreme-density regime is recoverable once the structural diagnosis is used to rebalance the ELBO.

B.5 Statistical Reports

Table 8 reports a seed-level error bar for every cell of the main table rather than for a selected subset. Empirical validation of the energy ranking assumption is shown in Figure 6.

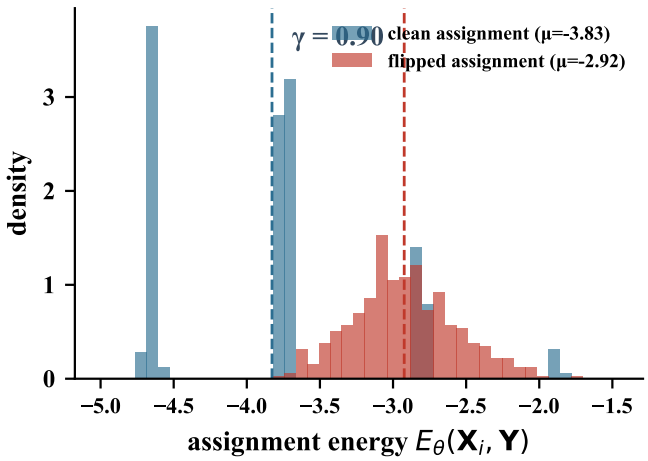

Figure 6: Empirical validation of the energy ranking assumption. On the Scene benchmark at flip 0.3 with 50% instance noise, we measure on trained models the reduced label-sensitive energy $E_\theta(\mathbf{X}_i, \mathbf{Y}) = \sum_{c:\mathbf{Y}_c=1} \min_j E_{jc}$ of Eq. (1), scoring each positive label by its strongest supporting instance (the measurement protocol of §A.6, while the contrastive training term itself uses the mean instance energy of Eq. (19), so the two quantities order assignments by the same sign but are not numerically identical). Clean labels (blue, $\mu = -3.83$, n=400) are separated from 30%-flipped labels (red, $\mu = -2.92$, n=400): the gap γ=0.90 is far larger than its standard error (Welch t-test t=22.4, $p < 10^{-69}$). Under the benchmark's fixed noisy labels the gap is γ=0.30 (t=5.7, $p < 10^{-7}$). Both measurements confirm the ordering that the NCE contrastive term (Prop. A.3) enforces.

C Full 20-Method Comparison Matrices

Tables 9 to 11 report the complete Micro F1 comparison of VEM against all 20 baselines for each of the three noise levels. Traditional MIML methods were not run on datasets where their per-label SVM training does not scale (Corel5K) or where the run timed out, and these entries are marked `n/a`. The best baseline per row is underlined, and VEM is bold when it strictly exceeds the best baseline.

Table 8: Three-seed robustness of all 30 main-table settings under the honest early-stopping protocol (model selection on the noisy validation split). $F1_{42}$ is the single seed-42 run reported in the main table, and $\mu \pm \mathrm{sd}$ is the mean and standard deviation over seeds 42, 7, 2024, where sd is the dispersion of the three runs rather than the logistic sigmoid used in the Method. Δ is the 3-seed mean minus the best baseline.

Setting	$F1_{42}$	$\mu \pm \mathrm{sd}$	Best base	Δ	Δ/sd
Scene-02	0.8937	0.8901±0.0030	GCE (0.9005)	−0.0104	−3.5
Scene-03	0.9016	0.9012±0.0012	AttnMIL (0.9005)	+0.0006	+0.5
Scene-05	0.8975	0.8987±0.0013	CUE (0.9005)	−0.0018	−1.4
emotions-02	0.9075	0.9064±0.0016	MLDN (0.9071)	−0.0008	−0.5
emotions-03	0.9057	0.9046±0.0008	MLDN (0.9055)	−0.0008	−1.1
emotions-05	0.9044	0.9042±0.0001	COMES (0.9041)	+0.0001	+0.7
FashionMNIST-02	0.8253	0.8221±0.0028	MIMLVAE (0.8687)	−0.0466	−16.6
FashionMNIST-03	0.8154	0.8164±0.0008	MIMLVAE (0.8541)	−0.0377	−48.9
FashionMNIST-05	0.7307	0.7287±0.0047	MLDN (0.7308)	−0.0022	−0.5
OpenMIC-02	0.9696	0.9696±0.0000	COMES (0.9696)	+0.0000	+0.0
OpenMIC-03	0.9696	0.9696±0.0000	COMES (0.9696)	+0.0000	+0.0
OpenMIC-05	0.9696	0.9696±0.0000	COMES (0.9696)	+0.0000	+0.0
LetterCarroll-02	0.5755	0.5791±0.0029	MaxPool (0.5557)	+0.0234	+8.0
LetterCarroll-03	0.5581	0.5517±0.0047	MaxPool (0.5307)	+0.0210	+4.4
LetterCarroll-05	0.3969	0.4064±0.0075	MaxPool (0.3911)	+0.0153	+2.0
enron-02	0.6355	0.6363±0.0006	GCE (0.6377)	−0.0015	−2.4
enron-03	0.6408	0.6348±0.0051	GCE (0.6239)	+0.0108	+2.1
enron-05	0.5987	0.5984±0.0089	GCE (0.5937)	+0.0048	+0.5
Reuters-02	0.6846	0.6898±0.0036	MaxPool (0.6840)	+0.0058	+1.6
Reuters-03	0.6765	0.6827±0.0047	MaxPool (0.6528)	+0.0299	+6.3
Reuters-05	0.6111	0.6104±0.0055	Hill (0.5536)	+0.0567	+10.2
bibtex-02	0.2979	0.2986±0.0006	MaxPool (0.2658)	+0.0327	+58.9
bibtex-03	0.2610	0.2620±0.0020	MaxPool (0.2398)	+0.0222	+11.1
bibtex-05	0.1644	0.1493±0.0108	KISAR (0.1431)	+0.0063	+0.6
CAL500-02	0.7415	0.7439±0.0027	GCE (0.7306)	+0.0133	+4.9
CAL500-03	0.7425	0.7442±0.0019	GCE (0.7171)	+0.0271	+14.1
CAL500-05	0.7168	0.7229±0.0047	GCE (0.6776)	+0.0453	+9.5
Corel5K-02	0.3956	0.3928±0.0057	GCE (0.3828)	+0.0100	+1.7
Corel5K-03	0.3919	0.3997±0.0061	GCE (0.3849)	+0.0148	+2.4
Corel5K-05	0.3601	0.3688±0.0062	GCE (0.3867)	−0.0178	−2.9

Table 9: Full 20-method comparison, label flip 0.2 + 50% instance noise. Micro F1. **Bold** marks a strict improvement of VEM over the best baseline, underline marks the best baseline with ties underlined, and n/a marks not applicable. Methods are rows, datasets are columns. Abbreviations: SC = Scene, EM = emotions, FM = FashionMNIST, OM = OpenMIC, LC = LetterCarroll, EN = enron, RE = Reuters, BB = bibtex, CA = CAL500, CK = Corel5K.

Method	SC	EM	FM	OM	LC	EN	RE	BB	CA	CK
COMES	0.8767	0.9041	0.8113	0.9696	0.5165	0.6258	0.6280	0.2104	0.7285	0.3284
Hill	0.8895	0.9041	0.7974	0.9696	0.5331	0.6235	0.6305	0.2297	0.7117	0.3222
CUE	0.8828	0.9041	0.8104	0.9696	0.5190	0.6209	0.6279	0.1986	0.7293	0.3313
RIDE	0.8767	0.9041	0.8113	0.9696	0.5165	0.6258	0.6280	0.2104	0.7285	0.3284
BCE	0.8767	0.9041	0.8113	0.9696	0.5165	0.6258	0.6280	0.2104	0.7285	0.3284
SCE	0.8691	0.9041	0.8346	0.9696	0.3905	0.5949	0.5843	0.1620	0.7198	0.3068
GCE	0.9005	0.9041	0.8015	0.9696	0.5008	0.6377	0.6238	0.1887	0.7306	0.3828
JAL	0.9005	0.9041	0.8661	0.9696	0.0000	0.5603	0.4406	0.0000	0.7194	0.2350
AttnMIL	0.8916	0.9041	0.8020	0.9696	0.4845	0.6234	0.6497	0.2094	0.7279	0.3575
MaxPool	0.8887	0.9041	0.7685	0.9696	0.5557	0.6333	0.6840	0.2658	0.7171	0.3820
CoTeach	0.8767	0.9041	0.8113	0.9696	0.5165	0.6258	0.6280	0.2104	0.7285	0.3284
BML	0.8767	0.9041	0.8113	0.9696	0.5165	0.6258	0.6280	0.2104	0.7285	0.3284
MIMLSVM	0.7949	0.7941	0.3743	0.9432	0.2222	0.3764	0.3936	0.1312	0.5254	0.2415
MIMLkNN	0.8999	0.9032	0.7179	0.9693	0.1716	0.6137	0.5414	0.1084	0.7070	0.3037
MIMLfast	0.8062	0.7994	0.6660	0.9629	0.3905	0.4860	0.4852	0.1984	0.5645	0.2436
MIMLRBF	0.7029	0.6621	0.0000	0.9693	0.0087	0.3458	0.3007	0.0001	0.3717	0.0567
MIMLNN	0.8676	0.8531	0.7268	0.9261	0.3852	0.5487	0.5863	0.2106	0.5991	0.2649
KISAR	0.7563	0.7515	0.5247	0.9682	0.3044	0.4338	0.4737	0.1890	0.3533	0.1279
MLDN	0.8820	0.9071	0.7728	0.9696	0.3726	0.4086	0.5699	0.2069	0.6907	0.2926
MIMLVAE	0.8668	0.9042	0.8687	0.9696	0.4243	0.6055	0.5917	0.1783	0.7233	0.3012
Ours	0.8937	**0.9075**	0.8253	0.9696	**0.5755**	0.6355	**0.6846**	**0.2979**	**0.7415**	**0.3956**

Table 10: Full 20-method comparison, label flip 0.3 + 50% instance noise. Micro F1. **Bold** marks a strict improvement of VEM over the best baseline, underline marks the best baseline with ties underlined, and n/a marks not applicable. Methods are rows, datasets are columns. Abbreviations: SC = Scene, EM = emotions, FM = FashionMNIST, OM = OpenMIC, LC = LetterCarroll, EN = enron, RE = Reuters, BB = bibtex, CA = CAL500, CK = Corel5K.

Method	SC	EM	FM	OM	LC	EN	RE	BB	CA	CK
COMES	0.8999	0.9041	0.7704	0.9696	0.4776	0.6105	0.6044	0.1883	0.7136	0.3494
Hill	0.8989	0.9041	0.7563	0.9696	0.5048	0.6088	0.6186	0.1969	0.6960	0.3452
CUE	0.9000	0.9041	0.7688	0.9696	0.4784	0.6131	0.6130	0.1797	0.7142	0.3527
RIDE	0.8999	0.9041	0.7704	0.9696	0.4776	0.6105	0.6044	0.1883	0.7136	0.3494
BCE	0.8999	0.9041	0.7704	0.9696	0.4776	0.6105	0.6044	0.1883	0.7136	0.3494
SCE	0.8523	0.9041	0.8106	0.9696	0.3502	0.5506	0.5219	0.1305	0.6841	0.3186
GCE	0.9002	0.9041	0.7648	0.9696	0.4673	0.6239	0.6084	0.1490	0.7171	0.3849
JAL	0.9005	0.9041	0.8516	0.9696	0.0000	0.5082	0.4406	0.0000	0.6881	0.2350
AttnMIL	0.9005	0.9041	0.7755	0.9696	0.4228	0.6082	0.6223	0.1835	0.7166	0.3795
MaxPool	0.8973	0.9041	0.7294	0.9696	0.5307	0.6094	0.6528	0.2398	0.6942	0.3745
CoTeach	0.8999	0.9041	0.7704	0.9696	0.4776	0.6105	0.6044	0.1883	0.7136	0.3494
BML	0.8999	0.9041	0.7704	0.9696	0.4776	0.6105	0.6044	0.1883	0.7136	0.3494
MIMLSVM	0.8898	0.7391	0.5920	0.9443	0.1455	0.4061	0.3595	0.0796	0.4842	0.1667
MIMLkNN	0.8995	0.9030	0.7021	0.9683	0.1369	0.5978	0.5186	0.0757	0.6751	0.2935
MIMLfast	0.7709	0.8016	0.6349	0.9505	0.3665	0.4403	0.4466	0.1822	0.5323	0.2272
MIMLRBF	0.6453	0.6414	0.0000	0.9574	0.0087	0.2405	0.2092	0.0000	0.2481	0.0202
MIMLNN	0.8486	0.8377	0.6858	0.9051	0.3507	0.5060	0.5406	0.1874	0.5525	0.2621
KISAR	0.6659	0.6779	0.4302	0.9569	0.2825	0.3955	0.4405	0.1715	0.1758	0.1300
MLDN	0.8833	0.9055	0.7663	0.9696	0.3344	0.3858	0.5156	0.1827	0.6495	0.2943
MIMLVAE	0.9005	0.9041	0.8541	0.9696	0.3918	0.5470	0.5400	0.1483	0.6711	0.2761
Ours	**0.9016**	**0.9057**	0.8154	0.9696	**0.5581**	**0.6408**	**0.6765**	**0.2610**	**0.7425**	**0.3919**

Table 11: Full 20-method comparison, label flip 0.5 + 50% instance noise. Micro F1. **Bold** marks a strict improvement of VEM over the best baseline, underline marks the best baseline with ties underlined, and `n/a` marks not applicable. Methods are rows, datasets are columns. Abbreviations: SC = Scene, EM = emotions, FM = FashionMNIST, OM = OpenMIC, LC = LetterCarroll, EN = enron, RE = Reuters, BB = bibtex, CA = CAL500, CK = Corel5K.

Method	SC	EM	FM	OM	LC	EN	RE	BB	CA	CK
COMES	0.8932	0.9041	0.6789	0.9696	0.3515	0.5706	0.5301	0.1018	0.6732	0.3398
Hill	0.8957	0.9041	0.6777	0.9696	0.3854	0.5503	0.5536	0.1252	0.6458	0.3348
CUE	0.9005	0.9041	0.6779	0.9696	0.3557	0.5798	0.5406	0.0988	0.6739	0.3486
RIDE	0.8932	0.9041	0.6789	0.9696	0.3515	0.5706	0.5301	0.1018	0.6732	0.3398
BCE	0.8932	0.9041	0.6789	0.9696	0.3515	0.5706	0.5301	0.1018	0.6732	0.3398
SCE	0.8098	0.9038	0.6142	0.9696	0.1596	0.3511	0.3491	0.0529	0.2268	0.1910
GCE	0.9005	0.9041	0.7103	0.9696	0.3171	0.5937	0.5521	0.0997	0.6776	0.3867
JAL	0.9005	0.9041	0.4612	0.9696	0.0000	0.1697	0.2711	0.0000	0.1145	0.0000
AttnMIL	0.8971	0.9041	0.6927	0.9696	0.3242	0.5630	0.5311	0.0863	0.6751	0.3571
MaxPool	0.8991	0.9041	0.6708	0.9696	0.3911	0.5520	0.5276	0.1306	0.6245	0.3320
CoTeach	0.8932	0.9041	0.6789	0.9696	0.3515	0.5706	0.5301	0.1018	0.6732	0.3398
BML	0.8932	0.9041	0.6789	0.9696	0.3515	0.5706	0.5301	0.1018	0.6732	0.3398
MIMLSVM	0.5264	0.8611	0.5051	0.8934	0.2167	0.3321	0.3629	0.0823	0.3712	0.1008
MIMLkNN	0.8923	0.8903	0.6075	0.9610	0.0835	0.4248	0.3616	0.0304	0.5053	0.1606
MIMLfast	0.7338	0.8112	0.5165	0.9209	0.2751	0.3500	0.3569	0.1396	0.4361	0.1269
MIMLRBF	0.4634	0.4436	0.0000	0.7043	0.0043	0.0526	0.0391	0.0000	0.0653	0.0000
MIMLNN	0.7944	0.7785	0.5379	0.8552	0.2387	0.3400	0.3795	0.1180	0.4429	0.1405
KISAR	0.6107	0.3968	0.1265	0.7100	0.1627	0.3134	0.3287	0.1431	0.0088	0.0596
MLDN	0.8375	0.8821	0.7308	0.9695	0.1475	0.3310	0.3268	0.1004	0.3209	0.1502
MIMLVAE	0.8997	0.9041	0.6773	0.9696	0.1633	0.3051	0.2243	0.0576	0.1719	0.0702
Ours	0.8975	**0.9044**	0.7307	0.9696	**0.3969**	**0.5987**	**0.6111**	**0.1644**	**0.7168**	0.3601