SALC: STRUCTURED ACTIVE LEARNING UNDER CORRELATED OUTPUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-label active learning often scores examples by summing labelwise uncertainties. Even when a predictor uses conditionally independent output heads, averaging over uncertain model parameters can induce dependencies in its posterior-predictive distribution. The sum of marginal entropies then exceeds joint entropy by the total correlation. We use this identity to motivate Structured Active Learning under Correlated outputs (SALC), which subtracts a pairwise mutual-information correction from marginal predictive entropy and combines the result with labelwise Bayesian disagreement. With exact pairwise estimates and unit weights, the correction leaves only a higher-order interaction remainder. The practical rule estimates pairwise dependence from MC-dropout samples and uses empirical co-occurrence and frozen text-encoder similarities as reliability gates; these gates introduce an additional approximation error. On MS-COCO, Pascal VOC, and CheXpert, SALC improves the reported label-efficiency metrics over the tested uncertainty, diversity, and graph-based baselines. Controlled synthetic experiments and component ablations support the role of the pairwise correction, while identifying higher-order dependence and finite-sample estimation as limitations. The results motivate accounting for posterior-predictive output dependence when designing multi-label acquisition criteria.

1 INTRODUCTION

Active learning seeks to reduce annotation cost by querying informative examples. Common acquisition criteria use predictive entropy or Bayesian disagreement (Lewis & Gale, 1994; Settles, 2009; Houlsby et al., 2011; Gal et al., 2017). In multi-label problems, however, outputs often carry shared information: object categories co-occur in images, topics overlap in text, and findings are correlated in medical imaging. Summing labelwise uncertainties is computationally convenient (Brinker, 2003; Huang et al., 2017), but does not account for this shared information in the posterior-predictive distribution.

The relevant dependence can arise even with conditionally independent sigmoid heads. Although the outputs factorize for fixed model parameters, averaging over parameter uncertainty generally produces a joint distribution that does not factorize. Under this posterior predictive, the sum of marginal entropies exceeds joint entropy by the total correlation. For conditionally independent heads, the same quantity is the gap between summed labelwise and joint Bayesian information gain. This identity diagnoses an overcounting of uncertainty; it does not, by itself, quantify the annotation budget lost or guarantee that correcting the score will improve prediction.

Input-space diversity methods, including Core-Set (Sener & Savarese, 2018) and BADGE (Ash et al., 2020), address coverage among queried examples. Label-structure methods instead use relationships such as co-occurrence to guide acquisition (Mahapatra et al., 2023). Our approach focuses on a complementary quantity: dependence among the predicted labels for each candidate input. Following the information-theoretic view of acquisition (MacKay, 1992; Krause et al., 2008), we approximate total correlation by a sum of pairwise mutual informations. Subtracting this sum with unit coefficients leaves a higher-order interaction remainder. The resulting entropy approximation is useful when that remainder is small, rather than whenever the predictor has a pairwise representation.

We introduce *Structured Active Learning under Correlated outputs* (SALC), which estimates posterior-predictive pairwise dependence from MC-dropout samples. The practical rule attenuates these

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.

estimates using reliability gates that combine empirical label co-occurrence with semantic similarity from a frozen text encoder (Radford et al., 2021; Wang et al., 2022; Zhang et al., 2023). These gates address limited labeled data but introduce approximation error beyond the ideal uniform correction. SALC adds the corrected predictive term to labelwise BALD (Houlsby et al., 2011; Gal et al., 2017), yielding a composite acquisition criterion. Figure 1 summarizes the pipeline. Our contributions are:

(i) An entropy-level diagnosis of labelwise acquisition. We express the gap between marginal and joint uncertainty as posterior-predictive total correlation, and decompose it into a pairwise contribution and a higher-order remainder. This distinguishes dependence induced by parameter marginalization from the conditional independence of the output heads.

(ii) A practical pairwise correction with an explicit error decomposition. We establish the residual error of the ideal unit-weight correction and separate the additional errors introduced by reliability gates, global correction strength, and Monte Carlo estimation. SALC combines this practical correction with labelwise epistemic uncertainty.

(iii) Evaluation of label efficiency and approximation behavior. Experiments on MS-COCO, Pascal VOC, and CheXpert compare SALC with uncertainty, diversity, and graph-based baselines. Synthetic experiments directly evaluate the entropy approximation under controlled dependence, while ablations assess the contributions of pairwise correction, semantic information, and epistemic uncertainty.

Related work and reproducibility. Extended related work is deferred to Appendix B. Source code with all training and inference configurations will be released upon acceptance.

2 Problem Formulation

We consider pool-based active learning for multi-label prediction, with input space $\mathcal{X} \subset \mathbb{R}^d$ and output space $\mathcal{Y} = \{0,1\}^K$. At round t, the learner scores each $x \in \mathcal{D}_U^{(t)}$, queries the top-$b$ candidates, and updates $\mathcal{D}_L^{(t+1)} = \mathcal{D}_L^{(t)} \cup \mathcal{B}^{(t)}$. For fixed parameters θ, our predictor uses conditionally independent Bernoulli heads, $p_\theta(\mathbf{y} \mid x) = \prod_{k=1}^K p_\theta(y_k \mid x)$. However, the posterior-predictive distribution $q_t(\mathbf{y} \mid x) = \mathbb{E}_{\Theta \sim p(\theta \mid \mathcal{D}_L^{(t)})}[p_\Theta(\mathbf{y} \mid x)]$ generally does not factorize, because label probabilities can co-vary across posterior samples.

A principled Bayesian acquisition rule maximizes the joint information gain, whereas a common tractable alternative sums labelwise information gains:

$$\begin{aligned} A_{\text{joint}}^{(t)}(x) &= I(\mathbf{Y}; \Theta \mid x, \mathcal{D}_L^{(t)}) = H_{q_t}(\mathbf{Y} \mid x) - \mathbb{E}_\Theta[H_{p_\Theta}(\mathbf{Y} \mid x)]\,, \\ A_{\text{marg}}^{(t)}(x) &= \sum_{k=1}^K I(Y_k; \Theta \mid x, \mathcal{D}_L^{(t)}). \end{aligned} \tag{1}$$

Computing $A_{\text{joint}}^{(t)}$ requires the entropy of a distribution over 2^K label configurations. The marginal surrogate is tractable, but it overcounts dependence induced by posterior marginalization. Since the fixed-parameter distribution factorizes, the discrepancy is exactly

$$\begin{aligned} \mathcal{B}_t(x) &= A_{\text{marg}}^{(t)}(x) - A_{\text{joint}}^{(t)}(x) \\ &= \sum_{k=1}^K H_{q_t}(Y_k \mid x) - H_{q_t}(\mathbf{Y} \mid x) \\ &= D_{\text{KL}}\left(q_t(\mathbf{y} \mid x) \,\middle\|\, \prod_{k=1}^K q_{t,k}(y_k \mid x) \right) \equiv \text{TC}_{q_t}(\mathbf{Y} \mid x) \geq 0, \end{aligned} \tag{2}$$

where $q_{t,k}$ is the kth posterior-predictive marginal and TC_{q_t} denotes posterior-predictive total correlation. This identity identifies the redundancy shared by marginal entropy and marginal information gain. We use a pairwise approximation to correct the predictive-entropy component of SALC; the final composite score in Eq. (14) also retains labelwise epistemic uncertainty. Figure 4 illustrates the entropy correction using two binary labels.

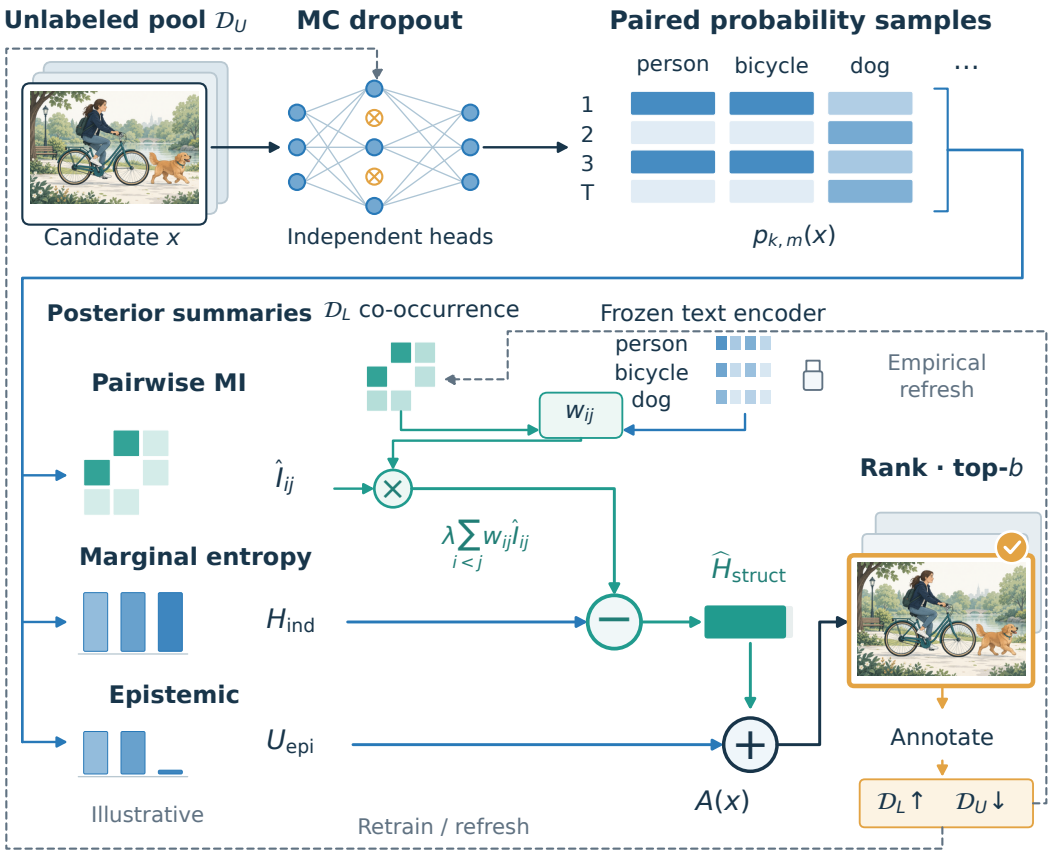


Figure 1: **SALC: from stochastic predictions to structured acquisition.** The illustrative scene and probability patterns show T dropout passes (crosses: inactive units). Paired predictions produce H_{ind}, $\widehat{I}_{ij}$, and U_{epi}. Empirical co-occurrence and frozen label embeddings gate only the MI correction. SALC subtracts $\lambda \sum_{i<j} w_{ij} \widehat{I}_{ij}$ and adds U_{epi} before top-b annotation. Dashed paths retrain the predictor and refresh empirical statistics; semantic embeddings remain fixed.

3 Structured Uncertainty under Correlated Outputs

3.1 Method Overview

Figure 1 shows one acquisition round. From T paired MC-dropout passes per candidate, SALC computes marginal entropy, pairwise MI, and labelwise BALD. Motivated by Eq. (2), it subtracts weighted pairwise MI from marginal entropy and adds BALD separately. Empirical co-occurrence and semantic similarity provide finite-sample reliability gates; these are practical heuristics, distinct from the ideal unit-weight correction. After querying the top-b candidates, the learner updates both pools, retrains the predictor, and refreshes empirical statistics while keeping semantic embeddings fixed.

3.2 Posterior-Predictive Pairwise Decomposition

For theoretical analysis, we approximate the posterior-predictive distribution $q_t(\mathbf{y} \mid x)$ by a pairwise exponential-family model

$$q_t(\mathbf{y} \mid x) = \frac{1}{Z_t(x)} \exp\left(\sum_{k=1}^{K} \phi_k^{(t)}(x) y_k + \sum_{i<j} \psi_{ij}^{(t)}(x) y_i y_j \right). \tag{3}$$

This is an analytical model of the posterior-predictive mixture, not an additional correlated output head: the implemented predictor remains a set of conditionally independent sigmoid heads for each fixed parameter realization. Equation (3) motivates a local weak-coupling analysis; pairwise potentials alone do not eliminate higher-order interaction information. Appendix A.2 states the scope and assumptions of this approximation.

Suppressing the conditioning on $\mathcal{D}_L^{(t)}$ for readability, the multivariate entropy expansion gives

$$H_{q_t}(\mathbf{Y} \mid x) = H_{\mathrm{ind}}^{(t)}(x) - \sum_{i<j} I_{q_t}(Y_i; Y_j \mid x) + R_3^{(t)}(x), \tag{4}$$

where $H_{\mathrm{ind}}^{(t)}(x) = \sum_k H_{q_t}(Y_k \mid x)$, $I_{q_t}(Y_i; Y_j \mid x)$ is pairwise mutual information under the posterior predictive, and $R_3^{(t)}(x)$ collects interaction-information terms involving three or more labels. The identity itself holds for any q_t when $R_3^{(t)}$ is defined by Eq. (4); the approximation is useful when the higher-order remainder is small relative to the pairwise contribution.

3.3 Redundancy Bias and Structured Correction

Combining Eqs. (2) and (4) yields the pairwise characterization of posterior-predictive redundancy.

Theorem 1 (Posterior-Predictive Redundancy Bias)**.** *For any posterior-predictive distribution* $q_t(\mathbf{y} \mid x)$*,*

$$\mathcal{B}_t(x) = \sum_{i<j} I_{q_t}(Y_i; Y_j \mid x) - R_3^{(t)}(x). \tag{5}$$

The ideal pairwise correction subtracts every pairwise term with coefficient one. The practical SALC surrogate is

$$\widehat{H}_{\mathrm{struct}}^{(t)}(x) = H_{\mathrm{ind}}^{(t)}(x) - \lambda \sum_{i<j} w_{ij}^{(t)} \widehat{I}_{ij}^{(t)}(x), \tag{6}$$

where $\widehat{I}_{ij}^{(t)}$ estimates posterior-predictive pairwise MI, $\lambda \geq 0$ controls the global correction magnitude, and $w_{ij}^{(t)} \in [0, 1]$ is a finite-sample reliability gate.

Theorem 2 (Uniform Pairwise Correction)**.** *If* $\widehat{I}_{ij}^{(t)}(x) = I_{q_t}(Y_i; Y_j \mid x)$*,* $w_{ij}^{(t)} = 1$*, and* $\lambda = 1$*, then*

$$\widehat{H}_{\mathrm{struct}}^{(t)}(x) = H_{q_t}(\mathbf{Y} \mid x) - R_3^{(t)}(x). \tag{7}$$

Theorem 2 guarantees the ideal unweighted surrogate. The deployed pair-dependent weighting is a theory-motivated finite-sample extension: $w_{ij}^{(t)}$ does not estimate an algebraic coefficient different from one, but attenuates pairwise estimates that are weakly supported by the available labeled data.

We estimate the reliability gates as

$$w_{ij}^{(t)} = \alpha \widetilde{\mathrm{NPMI}}_{ij}^{(t)} + (1 - \alpha) s_{ij}, \tag{8}$$

where $\widetilde{\mathrm{NPMI}}_{ij}^{(t)} \in [0, 1]$ is a rectified empirical co-occurrence score from $\mathcal{D}_L^{(t)}$, and $s_{ij} \in [0, 1]$ is the rescaled cosine similarity between frozen text embeddings of the label names. The empirical signal is task specific but noisy in early rounds; the semantic signal is immediately available but may be task misaligned. We use $\lambda = 1$ and $\alpha = 0.5$ in all headline experiments.

To separate the approximation errors, let $\varepsilon_{ij}^w = w_{ij}^{(t)} - 1$ and $\varepsilon_{ij}^I = \widehat{I}_{ij}^{(t)} - I_{q_t}(Y_i; Y_j \mid x)$. Then

$$\begin{aligned} \widehat{H}_{\mathrm{struct}}^{(t)}(x) - H_{q_t}(\mathbf{Y} \mid x) = & -R_3^{(t)}(x) - \lambda \sum_{i<j} \varepsilon_{ij}^w I_{q_t}(Y_i; Y_j \mid x) \\ & + (1 - \lambda) \sum_{i<j} I_{q_t}(Y_i; Y_j \mid x) - \lambda \sum_{i<j} w_{ij}^{(t)} \varepsilon_{ij}^I. \end{aligned} \tag{9}$$

Thus, the practical error comprises the higher-order remainder, reliability-gate shrinkage, global under- or over-correction, and Monte Carlo estimation error. Full proofs and estimator details are deferred to Appendices A.3 and A.4.

3.4 Posterior-Predictive Estimation and Final Acquisition

For MC-dropout pass m, let $p_{k,m}(x) = p_{\theta^{(m)}}(Y_k = 1 \mid x)$ and

$$\pi_{k,m}(a \mid x) = p_{k,m}(x)^a \left[1 - p_{k,m}(x)\right]^{1-a}, \qquad a \in \{0, 1\}. \tag{10}$$

The posterior-predictive marginal and pairwise joint used in the implementation are

$$\begin{aligned} \widehat{q}_{t,k}(a \mid x) &= \frac{1}{T} \sum_{m=1}^{T} \pi_{k,m}(a \mid x), \\ \widehat{q}_{t,ij}(a, b \mid x) &= \frac{1}{T} \sum_{m=1}^{T} \pi_{i,m}(a \mid x) \pi_{j,m}(b \mid x). \end{aligned} \tag{11}$$

Although every fixed-parameter component factorizes, the mixture in Eq. (11) generally does not. Figure 6 provides a worked visual example of why averaging paired products differs from multiplying averaged marginals. We therefore compute

$$\widehat{I}_{ij}^{(t)}(x) = \sum_{a,b \in \{0,1\}} \widehat{q}_{t,ij}(a, b \mid x) \log \frac{\widehat{q}_{t,ij}(a, b \mid x)}{\widehat{q}_{t,i}(a \mid x) \widehat{q}_{t,j}(b \mid x)}. \tag{12}$$

The labelwise epistemic term is

$$U_{\mathrm{epi}}^{(t)}(x) = \sum_{k=1}^{K} \left[H(\widehat{q}_{t,k}(\cdot \mid x)) - \frac{1}{T} \sum_{m=1}^{T} H(\pi_{k,m}(\cdot \mid x)) \right]. \tag{13}$$

It contains only labelwise parameter mutual information; the cross-label pairwise MI enters once, through Eq. (6). The implemented acquisition score is

$$A^{(t)}(x) = U_{\mathrm{epi}}^{(t)}(x) + \widehat{H}_{\mathrm{struct}}^{(t)}(x). \tag{14}$$

Equation (14) is a composite criterion combining dependency-corrected posterior-predictive uncertainty with labelwise epistemic exploration. It is not claimed to equal exact joint predictive entropy plus exact joint BALD. This distinction preserves the interpretation of the implemented score and makes the two contributions separately available for ablation.

Dense pairwise estimation requires $\mathcal{O}(TK^2)$ arithmetic per candidate and is implemented through batched tensor operations. The full pool-level acquisition cost is therefore $\mathcal{O}(|\mathcal{D}_U^{(t)}|TK^2)$; retaining only a sparse edge set $\mathcal{E}$ reduces the pairwise component to $\mathcal{O}(|\mathcal{D}_U^{(t)}|T|\mathcal{E}|)$, in addition to $\mathcal{O}(|\mathcal{D}_U^{(t)}|TK)$ for the labelwise terms. The complete algorithm and implementation details are given in Appendix A.6.

4 Experimental Evaluation

We evaluate three questions arising from Section 3: (i) whether additive labelwise uncertainty overcounts posterior-predictive dependence, as characterized by Theorem 1; (ii) whether the practical reliability-gated pairwise correction improves label efficiency; and (iii) whether its benefit is associated with the strength of output dependence. Theorem 2 applies directly to the ideal uniform correction, while the empirical–semantic weighting used by SALC is evaluated as a finite-sample extension.

Experiments use MS-COCO (80 labels), Pascal VOC 2007 (20 labels), CheXpert (14 labels), and a controlled synthetic benchmark with $K = 20$ binary labels and correlation parameter $\rho \in [0, 0.8]$. The synthetic setting permits exact entropy computation, whereas the real benchmarks evaluate posterior-predictive estimates obtained from the implemented sigmoid models.

All methods use the same five seed-matched initial 5% labeled sets, acquire 1% of the training pool per round until 50% is labeled, and retrain from scratch after every acquisition. SALC and BALD use $T = 20$ MC-dropout predictions; SALC uses $\lambda = 1$ and $\alpha = 0.5$ on every dataset without per-dataset

Table 1: **Label efficiency on real benchmarks.** mAP for MS-COCO and Pascal VOC and macro-AUC for CheXpert. Values are mean ± standard deviation over five seed-matched runs. The main text reports paired confidence intervals for the prespecified COCO and VOC comparisons. Underlined entries denote the strongest non-SALC result.

	MS-COCO		Pascal VOC		CheXpert	
Method	10%	AULC	10%	AULC	10%	AULC
Entropy	63.2 ± 0.31	71.1 ± 0.35	78.4 ± 0.41	84.7 ± 0.38	81.3 ± 0.38	86.8 ± 0.42
BALD	64.7 ± 0.35	72.6 ± 0.46	79.3 ± 0.37	85.6 ± 0.42	82.5 ± 0.33	88.1 ± 0.41
Core-Set	65.1 ± 0.44	72.9 ± 0.41	79.8 ± 0.31	85.9 ± 0.35	83.0 ± 0.48	88.3 ± 0.42
Co-occurrence Graph AL	66.8 ± 0.35	73.8 ± 0.40	80.2 ± 0.37	86.3 ± 0.35	84.2 ± 0.44	89.0 ± 0.43
Class-balanced entropy	$67.5 \pm 0.XX$	$74.3 \pm 0.XX$	$\underline{81.8 \pm 0.54}$	$\underline{87.3 \pm 0.51}$	$\underline{84.9 \pm 0.47}$	$\underline{89.8 \pm 0.38}$
Calibrated entropy	$\underline{67.9 \pm 0.47}$	$\underline{74.8 \pm 0.31}$	81.4 ± 0.44	86.9 ± 0.51	84.3 ± 0.37	89.1 ± 0.45
SALC	$\mathbf{69.4} \pm 0.23$	$\mathbf{76.9 \pm 0.37}$	$\mathbf{82.5} \pm 0.36$	$\mathbf{88.9 \pm 0.39}$	$\mathbf{87.1} \pm 0.38$	$\mathbf{91.2} \pm 0.42$

retuning. We compare against Random, Entropy, BALD, Core-Set, and Co-occurrence Graph AL, a custom GESTALT-inspired graph baseline. Table 1 additionally reports class-balanced and calibrated entropy on all three benchmarks; the available learning-curve figure shows the other comparison methods.

We report mAP for MS-COCO and Pascal VOC, macro-AUC for CheXpert, and AULC computed separately for each seed-level learning curve. Tables report five-run means, with standard deviations where available. The reported COCO and VOC headline comparisons use paired t-tests over matched seeds. Warm-start, initial-split, batch-size, higher-order, semantic-encoder, and runtime analyses are provided in Appendix C.

4.1 Label Efficiency on Real Benchmarks

Table 1 reports performance at the 10% labeled budget and AULC over the complete acquisition trajectory. SALC achieves the highest mean performance on all three benchmarks. On MS-COCO, it outperforms the strongest tested alternative, calibrated entropy, by 1.5 mAP at the 10% budget and 2.1 AULC. On Pascal VOC, SALC improves over class-balanced entropy by 0.7 mAP and 1.6 AULC, respectively. On CheXpert, the corresponding improvements over class-balanced entropy are 2.2 macro-AUC and 1.4 AULC.

For MS-COCO, the paired 95% confidence intervals for the improvements over calibrated entropy are $[1.23, 1.77]$ for mAP and $[1.80, 2.40]$ for AULC. For Pascal VOC, the paired 95% confidence intervals relative to class-balanced entropy are $[0.62, 0.78]$ and $[1.42, 1.78]$, respectively. Paired tests across the five seed-matched runs give $p < 0.01$ for each of these four prespecified comparisons. The CheXpert results are reported descriptively, and we make no corresponding significance claim. These strongest-baseline comparisons provide the most relevant estimates of SALC's incremental benefit.

Figure 2 shows the complete acquisition trajectories, with the largest absolute gains occurring in the label-scarce regime. Appendix C.6 provides additional per-budget and per-class results and examines the association between output dependence and the observed performance gains. This association is descriptive and should not be interpreted causally.

4.2 Synthetic Validation of the Pairwise Approximation

The synthetic benchmark permits exact enumeration of all 2^K label configurations and controlled variation of the dependence parameter ρ. We use it to evaluate the ideal uniform correction analyzed in Theorem 2, separately from the finite-sample reliability gates used on real data. Define $H_{\text{pair}}(x) = H_{\text{ind}}(x) - \sum_{i<j} I_q(Y_i; Y_j \mid x)$, together with the absolute approximation errors $\mathcal{E}_{\text{ind}}(x) = |H_{\text{ind}}(x) - H_q(\mathbf{Y} \mid x)|$ and $\mathcal{E}_{\text{pair}}(x) = |H_{\text{pair}}(x) - H_q(\mathbf{Y} \mid x)|$.

Figure 3a shows that $\mathcal{E}_{\text{ind}}$ increases as output dependence strengthens, whereas $\mathcal{E}_{\text{pair}}$ remains substantially smaller. Under the uniform setting $w_{ij} = 1$ and $\lambda = 1$, the remaining error corresponds to the higher-order term $R_3(x)$ in Theorem 2. Figure 3b shows that the downstream mAP advantage over independent entropy is negligible at $\rho = 0$ and increases with ρ. These controlled results support

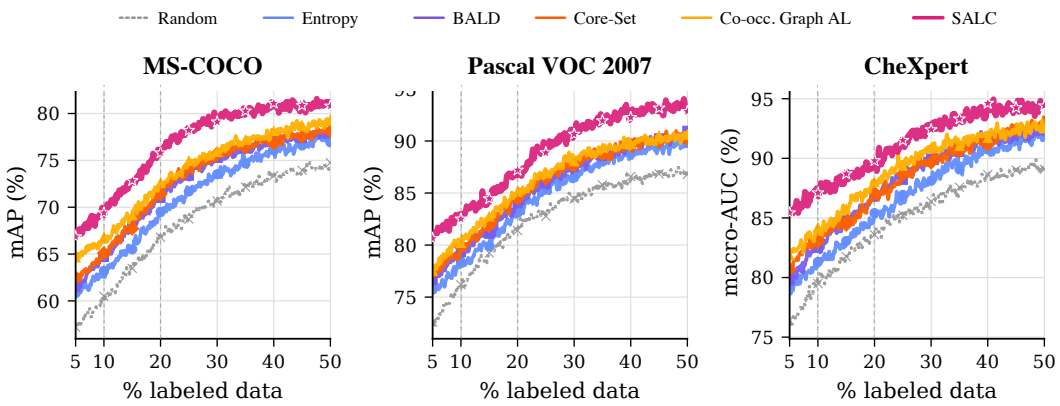
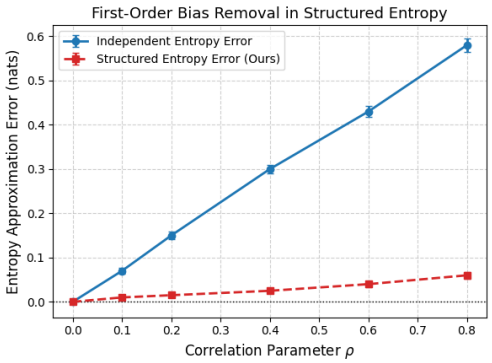
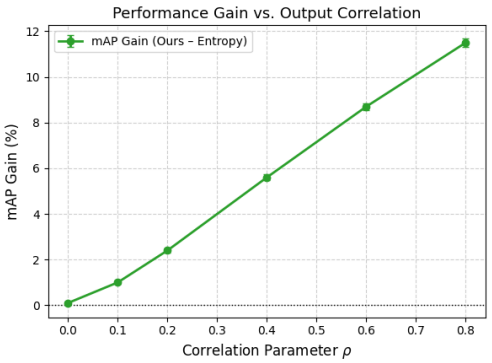


Figure 2: **Per-round learning curves on the three real benchmarks.** SALC is compared with Random, Entropy, BALD, Core-Set, and Co-occurrence Graph AL. Class-balanced and calibrated entropy are reported in Table 1 but are not shown in these curves. Shaded regions denote ± 1 standard deviation over five seed-matched runs; vertical dashed lines mark the 10% and 20% budgets.



(a) **Entropy approximation error.** Independent marginal entropy increasingly overestimates the joint entropy, whereas the uniform pairwise correction leaves a substantially smaller higher-order residual.



(b) **Downstream performance.** The mAP gain of SALC over independent entropy increases with output dependence and is negligible when $\rho = 0$.

Figure 3: **Synthetic validation under controlled output dependence.** Exact enumeration shows that the ideal coefficient-one pairwise correction substantially reduces the entropy approximation error, while the downstream advantage of structured acquisition increases with correlation strength. Panel (a) evaluates $w_{ij} = 1$ and $\lambda = 1$; Panel (b) reports the active-learning performance gain labelled SALC in the supplied plot.

the proposed pairwise-redundancy mechanism under the synthetic data-generating process; they do not by themselves establish a causal explanation for the real-benchmark gains. Synthetic-data construction, sensitivity to K, and higher-order stress tests are provided in Appendix C.7.

4.3 Mechanism Ablation

Table 2 separates the ideal coefficient-one correction of Theorem 2 from the finite-sample reliability gates used by SALC. Redundancy is evaluated using the same fixed posterior-predictive pairwise-MI diagnostic on the batches selected by every configuration, independently of that configuration's λ or w_{ij}.

Removing the pairwise correction causes the largest degradation: -3.3 COCO mAP, -3.3 COCO AULC, and -2.3 CheXpert macro-AUC, while redundancy rises from 0.27 to 0.41. Uniform correction recovers 79%, 73%, and 65% of these respective gains. The empirical–semantic gates then add $+0.7$ COCO mAP, $+0.9$ AULC, and $+0.8$ CheXpert macro-AUC while further reducing

Table 2: **Mechanism ablation.** COCO mAP and CheXpert macro-AUC are measured at the 10% labeled budget; COCO AULC covers the complete acquisition trajectory. Redundancy is the fixed queried-batch pairwise-MI diagnostic. Values are means over five seeds.

Configuration	COCO mAP	COCO AULC	CheXpert AUC	Redundancy
Full SALC	**69.4**	**76.9**	**87.1**	**0.27**
Uniform $w_{ij} = 1$	68.7	76.0	86.3	0.30
Without correction	66.1	73.6	84.8	0.41
Without semantic prior	67.2	74.8	85.6	0.31
Without epistemic term	68.0	75.5	86.2	0.29

Table 3: **Robustness to model-update protocol.** AULC is mean$\pm$standard deviation over five seed-matched runs.

Dataset	Protocol	Graph AL	SALC	Gain
MS-COCO	From scratch	73.8 ± 0.40	$\mathbf{76.9 \pm 0.37}$	$+3.1$
MS-COCO	Warm start	74.9 ± 0.40	$\mathbf{78.9 \pm 0.42}$	$+4.0$
Pascal VOC	From scratch	86.3 ± 0.35	$\mathbf{88.9 \pm 0.37}$	$+2.6$
Pascal VOC	Warm start	87.4 ± 0.40	$\mathbf{90.2 \pm 0.40}$	$+2.8$

redundancy from 0.30 to 0.27. Thus, coefficient-one pairwise subtraction is the primary mechanism, whereas the empirical–semantic gates provide an additional benefit in these experiments. The smaller degradation without U_{epi} indicates that labelwise epistemic exploration is complementary but not the dominant source of improvement.

4.4 Protocol Robustness, Approximation Quality, and Cost

The submitted experiments retrain every method from scratch after each acquisition. To test sensitivity to this choice, we repeated MS-COCO and Pascal VOC under a matched warm-start protocol in which each method initializes from its own preceding-round checkpoint.

The ordering is unchanged under warm-start training. Across 20 independently drawn COCO initial labeled sets, the AULC standard deviation is approximately 0.4, substantially below the default SALC advantage. SALC also remains ahead for acquisition batches of 0.5%, 1%, and 5%, with AULC gains of $+2.2$, $+3.1$, and $+3.4$, respectively. These results support robustness to initialization and acquisition granularity; they do not indicate a decreasing batch-size trend within the tested range.

We additionally evaluate a triplet-based diagnostic of interactions beyond pairs on the real benchmarks. At the 10% labeled budget, the reported normalized estimate is $\widehat{r}_3 = |\widehat{R}_3| / \sum_{i<j} \widehat{I}_{ij}$, where $\widehat{R}_3$ is the triplet-based proxy described in Appendix D; it does not estimate every term of order three and higher in R_3.

The median triplet-based proxy is 8%–13% of the pairwise mass, and an explicit triplet correction changes real-benchmark performance by at most 0.2 points. In contrast, triplet correction becomes beneficial in the deliberately strong higher-order synthetic stress test reported in Appendix D. These findings support the practical adequacy of the pairwise score in the reported comparisons, while leaving interactions of order four and above untested by this diagnostic.

All experiments use a single NVIDIA A100 80 GB GPU. Acquisition-only time per COCO round is 21 s for calibrated entropy, 24 s for class-balanced entropy, 31.5 s for BALD, 33 s for Co-occurrence Graph AL, and 41 s for SALC. SALC therefore costs 1.30 times BALD, 1.24 times the graph baseline, 1.71 times class-balanced entropy, and 1.95 times calibrated entropy; model retraining remains the dominant end-to-end cost.

5 Discussion and Limitations

Interpretation. The synthetic results in Figure 3 show that the unit-weight correction of Theorem 2 reduces entropy approximation error under controlled dependence. On real data, uniform pairwise

Table 4: **Triplet-based diagnostic at the 10% budget.** The last two columns report mAP for COCO/VOC and macro-AUC for CheXpert.

Dataset	Median $\widehat{r}_3$	90th percentile	Pairwise	Pairwise+triplet
MS-COCO	0.13	0.27	69.4	69.5
Pascal VOC	0.08	0.18	82.5	82.3
CheXpert	0.10	0.21	87.1	87.2

subtraction recovers most of the improvement in Table 2, while empirical–semantic reliability gates provide a further benefit. Tables 3 and 4 show stable performance ordering across the evaluated update protocols and limited benefit from explicit triplet correction. These findings support pairwise redundancy correction as a contributor to the observed gains, while leaving other influences on acquisition performance unresolved.

Comparison and robustness. In Table 1, calibrated entropy is the strongest tested alternative on MS-COCO, and class-balanced entropy is strongest on Pascal VOC and CheXpert. Co-occurrence Graph AL is a separate, custom GESTALT-inspired baseline that propagates uncertainty through label co-occurrence. Core-Set addresses input-space coverage, while SALC targets dependence among posterior-predictive label uncertainties; combining these mechanisms remains a possible extension. We fix $\lambda = 1$ and $\alpha = 0.5$ across all datasets. The sensitivity, warm-start, initialization, batch-size, semantic-encoder, and higher-order analyses in Appendices E and D report stable method ordering under the tested variations. These results support the chosen settings empirically, but do not establish that a fixed α is optimal at every acquisition round.

Limitations and broader impact. Theorem 2 characterizes the ideal uniform correction with exact pairwise estimates. The practical gates shrink that correction and can introduce bias; their finite-sample benefit is assessed empirically. The pairwise approximation also deteriorates when higher-order posterior-predictive interactions are substantial, and its estimates depend on how well MC-dropout represents parameter uncertainty. Empirical co-occurrence may be unreliable early in acquisition, while semantic similarity can be poorly aligned with the task. Dense estimation costs $\mathcal{O}(TK^2)$ per candidate: on MS-COCO, SALC requires 41 seconds per acquisition round, compared with 21–24 seconds for the deterministic calibrated and class-balanced entropy baselines, although retraining remains the dominant cost. Sparse pair selection and adaptive reliability schedules may reduce these costs and estimation errors. The medical-imaging evaluation concerns annotation prioritization; it does not establish clinical utility or freedom from biases inherited from the predictor, labeled set, or semantic encoder.

6 Conclusion

We introduced SALC, a multi-label acquisition criterion motivated by the total-correlation gap between marginal and joint posterior-predictive uncertainty. An ideal unit-weight pairwise correction leaves only a higher-order remainder; the practical rule adds reliability gates and combines the corrected predictive term with labelwise epistemic uncertainty. SALC improves the reported label-efficiency metrics over the tested alternatives on MS-COCO, Pascal VOC, and CheXpert. Synthetic experiments and component ablations support the contribution of pairwise correction, while the analysis distinguishes its ideal behavior from finite-sample estimation. Future work includes adaptive reliability gates, sparse corrections for large label spaces, and richer models of higher-order dependence.

Reproducibility Statement

Appendix A provides the multivariate-entropy decomposition underlying SALC, the validity regime of the pairwise approximation, complete proofs of Theorems 1 and 2, the reliability-weight error decomposition, algorithmic details, and computational-complexity analysis. Appendix C describes the preparation of MS-COCO, Pascal VOC 2007, and CheXpert; the synthetic data-generating process; predictor architectures; optimization settings; baseline implementations; evaluation metrics;

and statistical procedures. All methods use the same official dataset splits, five seed-matched initial labeled sets, predictor architectures, and training schedules. The default protocol begins with 5% labeled data, acquires 1% per round until 50% is labeled, and retrains the predictor from scratch after each acquisition. Appendix C also specifies the Monte Carlo dropout configuration, semantic encoder, construction and round-wise updating of the dependency weights, and the fixed default values $\lambda = 1$ and $\alpha = 0.5$. Appendices D and E report per-class and per-baseline analyses, warm-start and initialization robustness, acquisition-batch sensitivity, higher-order stress tests, practical diagnostics, negative results, hyperparameter sensitivity, and extended component ablations. Hardware and acquisition-time measurements are reported for all principal methods. Upon acceptance, we will release the source code, data-processing and split-generation procedures, random seeds, training and acquisition configurations, and scripts required to reproduce the reported tables and figures, subject to the licenses and access requirements of the original datasets.

USE OF LARGE LANGUAGE MODELS

Generative AI tools were used for prose editing and manuscript organization. They were not used to execute experiments, access raw data or model outputs, or generate, modify, estimate, or interpolate any reported numerical result. The authors independently verified the mathematical derivations, implementation, citations, experiments, and analysis, and claims, and take full responsibility for the entire content of the paper.

ETHICS STATEMENT

This work develops an annotation-efficient active-learning method and evaluates it on MS-COCO, Pascal VOC 2007, CheXpert, and synthetic data. All benchmark datasets are used in accordance with their licenses and access requirements, and the study introduces no new human-subject data collection. For CheXpert, we use the released de-identified data and make no attempt to re-identify individuals. SALC may inherit biases from the training data, clinical labels, pretrained visual representations, and the frozen text encoder used to construct semantic reliability gates. Its acquisition decisions may also amplify class imbalance or institutional sampling biases by preferentially selecting particular correlated label patterns. Moreover, empirical co-occurrence and semantic similarity encode statistical associations rather than causal or clinically validated relationships. Consequently, improved label efficiency or predictive performance does not establish demographic fairness, clinical validity, or robustness across healthcare institutions. The medical-imaging results are intended solely for methodological research and do not support autonomous diagnosis or clinical deployment without prospective external validation, subgroup analysis, appropriate regulatory review, and oversight by qualified clinicians. We report computational requirements to support transparent assessment of the method's resource costs.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pp. 31–40, Tübingen, Germany, 2009.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 59–66, 2003.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1183–1192, 2017.

Andrew Guillory and Jeff A. Bilmes. Active semi-supervised learning using submodular functions. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pp. 274–282, 2011.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011.

Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. Cost-effective active learning from diverse labelers. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1879–1885, 2017.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

Dwarikanath Mahapatra, Alexander Poellinger, and Mauricio Reyes. Graph node based interpretability guided sample selection for active learning. *IEEE Transactions on Medical Imaging*, 42(3): 661–673, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1070–1079, 2008.

Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March 2002.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.

Sheng Zhang et al. BioMedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *CoRR*, abs/2303.00915, 2023.