Deconstructing Multi-Task Active Learning: The Paradox of Gradient Conflict and Orthogonal Decomposition

Juho Jung[*]
Sungkyunkwan University
jhjeon9@g.skku.edu

Seongjae Kang[*]
KAIST
tjdwo2744@kaist.ac.kr

Jinyoung Han[†]
Sungkyunkwan University
jinyounghan@skku.edu

Abstract

Existing multi-task active learning (MTL-AL) methods predominantly rely on the linear scalarization of heterogeneous objectives, such as uncertainty, diversity, and gradient conflict. We argue that this entangled formulation is fundamentally limited, as linear combination fails to capture Pareto-optimal trade-offs under conflicting objectives. To address this limitation, we propose a paradigm shift from *combination* to *deconstruction* and introduce a principled framework that rethinks MTL-AL along three orthogonal dimensions. First, we identify the *Conflict Paradox*: although the optimization literature treats gradient conflict as interference to be suppressed, carrying that intuition into the acquisition score is counterproductive — penalizing conflict when ranking candidates is measurably worse than not scoring it at all. We therefore decouple the two concerns: the acquisition function does not score conflict, and the interference it causes is resolved separately by gradient surgery during training. Second, we introduce *Orthogonal Decomposition*, separating the acquisition objective into two distinct axes: *Informativeness* and *Exploration*. This two-stage process, consisting of informativeness filtering followed by coverage maximization, prevents "*diversity dilution*" by excluding uninformative outliers before enforcing diversity. Third, we propose *Cross-Task Mutual Information (CTMI)*, a covariance-based statistic of cross-task dependency computed from MC-dropout task entropies. Extensive experiments on COCO, Cityscapes, and NYUv2 — including comparisons against recent multi-task active learning and data-selection baselines — demonstrate consistent improvements over state-of-the-art MTL-AL methods, highlighting that structural rethinking of acquisition yields larger gains than incremental heuristic design. Our code is publicly available at: github.com/Juho-Jung/multitask_active_learning_official

1 Introduction

Multi-task learning (MTL) has emerged as a fundamental paradigm for building efficient and generalizable representations by leveraging shared structures across related tasks [5, 32]. However, its success highly depends on the availability of densely annotated data for every task, resulting in prohibitive annotation costs as the number and complexity of tasks grow [52, 43]. Active Learning (AL) offers an approach to mitigate this burden by selectively querying the most valuable samples [35]. Yet, extending AL to MTL settings introduces a fundamental challenge: the notion of "*informativeness*" becomes inherently multi-faceted, requiring the joint consideration of uncertainty, diversity, and inter-task relationships [47].

[*]Equal contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The Fundamental Problem: Linear Combination. Most existing MTL active learning methods address this challenge by formulating sample selection as a multi-objective optimization problem solved via linear scalarization [47, 17, 15, 16, 29, 31, 6]. Representative approaches such as BADGE [2] and BatchBALD [18] combine uncertainty and diversity through weighted summation. In multi-task settings, this strategy is commonly extended by aggregating task-wise uncertainties [41, 50] or incorporating penalties for gradient conflict [51, 7]. We argue this reliance on linear combination reflects a fundamental geometric limitation rather than a mere modeling choice. Since the objectives are incommensurable and measured in different units (e.g., entropy vs. feature distance vs. gradient cosine), linear scalarization fails to discover non-convex regions of the Pareto frontier [33, 23], often leading to sub-optimal trade-offs driven by arbitrary weighting. This limitation manifests in two failure modes: "*Diversity Dilution*," where outliers dominate due to extreme distance scores [34], and the "*Conflict Paradox*," where samples informative for inter-task trade-offs are prematurely discarded.

Deconstruction as a paradigm. Rather than seeking better weighting schemes, we rethink the problem structure via two principles: (i) **Decoupling Selection from Optimization** — keep gradient conflict out of the acquisition score entirely, and resolve the interference it causes at training time via gradient surgery, instead of letting an optimization-time concern filter which samples are ever labelled; (ii) **Orthogonal Decomposition** — decompose acquisition into two sequential, conceptually orthogonal axes (*Informativeness* and *Exploration*) and apply them hierarchically so that diversity is enforced only within a subspace of genuinely informative candidates.

Our contributions are summarized as follows:

- We identify the **Conflict Paradox**: penalizing gradient conflict inside the acquisition score is measurably worse than not scoring conflict at all, and we introduce a principled decoupling that leaves conflict to the optimizer rather than to the selector.
- We introduce **Orthogonal Decomposition**, a two-stage selection strategy that first filters candidates by informativeness and subsequently maximizes coverage within the retained set, preventing diversity dilution.
- We propose **Cross-Task Mutual Information (CTMI)**, a forward-pass-only statistic of cross-task dependency, and show that any non-zero cross-task covariance in our architecture is structurally attributable to the shared representation; we validate CTMI as an acquisition signal by ablation rather than by a distributional bound.
- We demonstrate consistent improvements on three benchmarks — COCO (natural images), Cityscapes (autonomous driving) and NYUv2 (indoor scenes) — reaching $+10.0\%$ over Random and $+2.4\%$ to $+3.0\%$ over prior active-learning methods under our main protocol (Table 2), and the highest paired gain against five recent multi-stage selection methods under a separate unified multi-seed protocol (Table 3).

2 Related Work

Active learning and the linear-combination trap. Deep AL is commonly categorized into uncertainty-based methods (entropy [37, 12], margin [30], MC-Dropout [11], deep ensembles [21]), diversity-based methods (CoreSet [34], discriminative coverage [14, 41]), and hybrids that linearly combine the two: BADGE [2] combines gradient magnitude with k-means++ clustering, BatchBALD [18] maximizes batch-level MI, and VAAL [41]/Learning Loss [50] learn auxiliary scalar predictors. As we show in Section 6 and Appendix G.2, this linear-combination paradigm suffers from *"Diversity Dilution"* — uninformative outliers with extreme distance scores dominate selection.

Multi-task learning, gradient conflict, and MTL active learning. MTL suffers from gradient conflict between tasks, and every established remedy acts at *optimization* time: task balancing (GradNorm [7], MGDA [33]) and gradient surgery (PCGrad [51], CAGrad [24]). Cross-task conflict has, to our knowledge, never been used as an *acquisition* signal: multi-task AL scores candidates by aggregated per-task uncertainty [29] or by extending single-task heuristics [36, 38, 31], leaving the selection stage blind to inter-task structure. We call the resulting situation the ***Conflict Paradox***: when a conflict term *is* introduced into the acquisition score, penalizing it — the reading that the optimization literature invites — is measurably worse than omitting it altogether (Table 4). The closest recent methods, Hekimoglu et al. [16, 17] (PAD-Net inconsistency + diversity) and PartAL [10] (per-task partial labeling), depend on specific task pairings or architectural components;

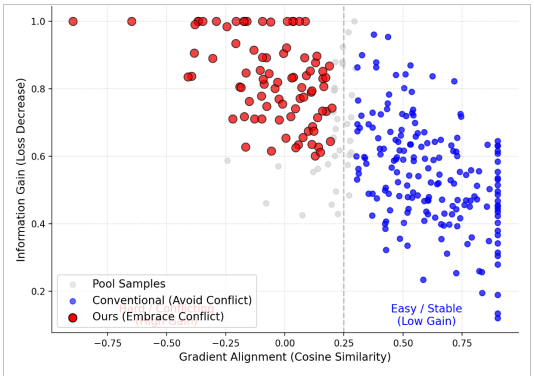


Figure 1: **Visualizing the Conflict Paradox.** Retrospective oracle analysis on *1,000 COCO benchmark* samples using a multi-task model pretrained on a *1% seed set*. Gradient alignment (cosine similarity, x-axis) is negatively correlated with actual loss reduction after training (y-axis) under AdamW with PCGrad off, ruling out the resolution mechanism as a confound: high-conflict samples yield larger information gain once basic features are learned.

our CTMI is architecture-agnostic and uses only forward passes through the shared encoder. For mutual-information estimation, classical k-NN [19] and neural estimators (MINE [3], CPC [26]) are impractical inside the AL inner loop; CTMI instead requires only the empirical covariance of MC-Dropout entropies, with no auxiliary network and no inner optimization.

Two-stage selection is established practice. Filtering a candidate pool by an informativeness score and then selecting for coverage within the filtered set is a well-established template. FASS [45] filters by uncertainty before submodular subset selection; Cluster-Margin [8] margin-filters before cluster-diverse sampling; PPAL [49] applies difficulty-calibrated uncertainty sampling followed by category-conditioned diversity; $\mathbb{D}^2$-Pruning [25] and SES [48] balance difficulty against diversity through graph message passing and structural entropy respectively; and related submodular and coverage-based formulations are surveyed in [34, 2]. Our contribution is the multi-task-specific quantity that drives its first stage — a cross-task dependency statistic computed from the shared encoder — together with the analysis motivating it. Section 6 isolates exactly this by running the same two-stage template with single-task signals under one identical multi-seed protocol.

3 The Conflict Paradox

Existing multi-task active learning methods penalize gradient conflict, treating it as optimization noise [24, 51, 39, 7]. As illustrated in Figure 1, this assumption induces a systemic bias toward the "aligned" region (blue area), where samples follow the model's current optimization trajectory. However, our retrospective analysis shows that after early training, samples yielding the largest joint-loss reduction concentrate in high-conflict regions (red area). Penalizing conflict thus discards high-gain samples. Motivated by this, we propose the **Decoupling Principle**, which explicitly targets this informative yet neglected region.

3.1 Formalizing the Paradox

Let $\mathcal{L}_t(x;\theta)$ be the loss for task t on sample x. Gradient conflict is typically quantified by the cosine similarity:

$$C_g(x) = \cos(\mathbf{g}_i, \mathbf{g}_j) = \frac{\mathbf{g}_i \cdot \mathbf{g}_j}{\|\mathbf{g}_i\|\|\mathbf{g}_j\|}, \quad \text{where } \mathbf{g}_t = \nabla_\theta \mathcal{L}_t(x) \tag{1}$$

Proposition 1 (Sample-Level Conflict and Local Rank Deficiency)**.** *Let $J(x) = [\mathbf{g}_i(x), \mathbf{g}_j(x)] \in \mathbb{R}^{d\times 2}$ denote the per-sample task Jacobian. Under standard local-linearization assumptions on the shared encoder, samples with extreme gradient conflict ($C_g(x) \to -1$, i.e., $\mathbf{g}_i \approx -k\mathbf{g}_j$, $k > 0$) satisfy:*

1. ***Sample-level Pareto-critical structure.*** *There exists $\alpha \in (0,1)$ with $\alpha\mathbf{g}_i + (1-\alpha)\mathbf{g}_j = \mathbf{0}$, i.e., the per-sample loss vector $(\mathcal{L}_i(x), \mathcal{L}_j(x))$ admits no joint descent direction at the current θ, and $J(x)$ is locally rank-deficient.*

2. ***Bridge to active learning.*** *Acquiring such samples imposes localized constraints that the current shared representation cannot reconcile; resolving them via gradient projection (PCGrad [51]) forces the encoder to expand into directions orthogonal to the dominant*

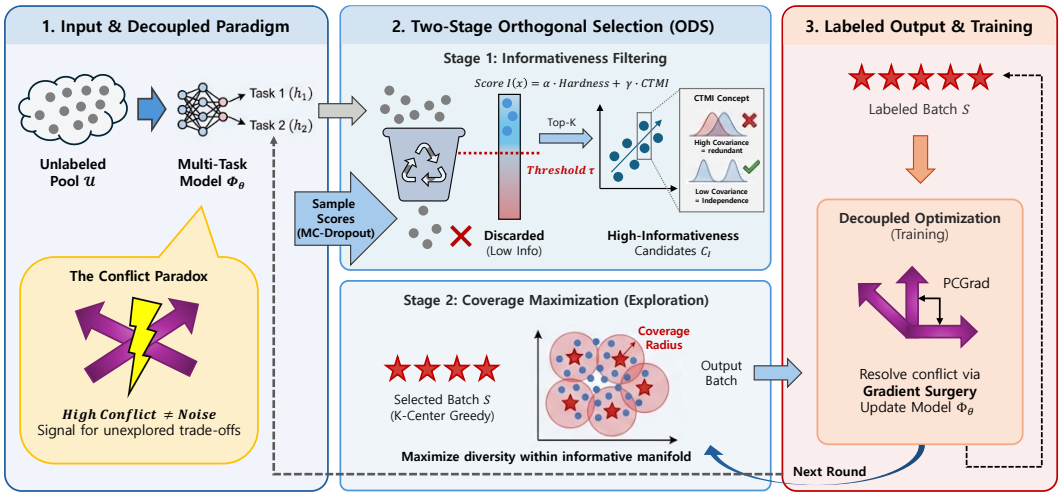


Figure 2: **Overview of the Orthogonal Decomposition Selection (ODS) Framework. (Left) The Conflict Paradox:** Conflict-aware selection treats gradient conflict as an informative signal. **(Middle) Orthogonal Decomposition:** Acquisition is decomposed into *Informativeness Filtering* (Hardness and CTMI) followed by *Coverage Maximization* within the retained candidates. **(Right) Decoupled Optimization:** Whatever gradient interference the selected batch contains is resolved at training time by *PCGrad*.

trajectory, yielding feature-space coverage gain not attainable from low-conflict (cooperative rank-1) samples.

Remark 1 (Sample-level vs. objective-level Pareto stationarity)**.** Objective-level *Pareto stationarity is a global statement about θ over the data distribution and is strictly stronger than the sample-level structure above. We therefore deliberately phrase Proposition 1 as a local, sample-conditional property — a structural heuristic that motivates our acquisition function — rather than a global optimization guarantee. The claim we draw from this is narrow and concerns the acquisition score only: with the optimizer held fixed,* penalizing *conflict in the score (Row A) is worse than* omitting *it (Row B), so the penalty removes candidates the model had something to learn from (Section 6).*

Proof Sketch. In high-conflict regions where $\mathbf{g}_i \approx -k\mathbf{g}_j$ ($k > 0$), the joint descent direction $\mathbf{g}_i + \mathbf{g}_j$ vanishes and $J(x)$ collapses to rank one but with antiparallel orientation (*adversarial* rather than *cooperative* rank-1). Standard SGD stalls at this point for sample x. Selecting and applying PCGrad breaks the deadlock by projecting gradients onto the normal plane, forcing the encoder to expand its feature space to satisfy the conflicting constraints. The formal statement, full assumptions, and complete derivation appear in Appendix A.1. □

3.2 The Decoupling Principle

Based on Proposition 1, we resolve the tension between selection and optimization by acting on each separately: at *selection* the acquisition score contains no conflict term at all, so conflicting samples are neither sought out nor filtered away; at *optimization* we apply gradient surgery (PCGrad [51]) to project conflicting gradients onto the normal plane and stabilize updates without discarding the data.

4 Orthogonal Decomposition of Acquisition

We propose **Orthogonal Decomposition Selection (ODS)**, a framework that departs from linear scalarization by decomposing acquisition into two structurally distinct axes: **Informativeness** ($\mathcal{I}$) and **Exploration** ($\mathcal{E}$). ODS operates as a three-phase decoupled pipeline (Figure 2): *(1) Scoring* — compute per-task predictive uncertainty and its cross-task structure under MC-dropout; *(2) Two-Stage Selection* — filter the pool to the high-$\mathcal{I}$ candidates, then maximize coverage within them; *(3) Decoupled Optimization* — train the selected batch under PCGrad [51], which resolves whatever gradient conflict the batch happens to contain.

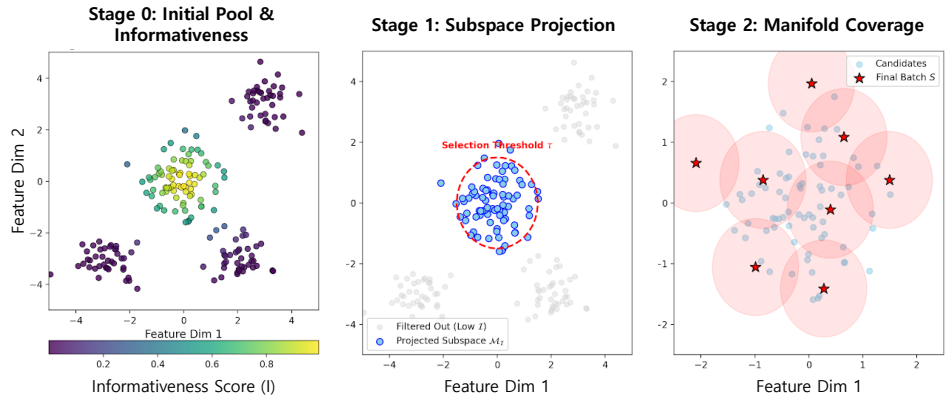


Figure 3: **The Orthogonal Decomposition Framework.** Instead of a single weighted sum, we process objectives sequentially. **Stage 1** projects data onto the High-Informativeness Subspace ($\mathcal{M}_\mathcal{I}$) (filtering out easy/redundant samples). **Stage 2** maximizes coverage within this subspace, preventing "*Diversity Dilution*."

Notation. $\Phi_\theta(x) \in \mathbb{R}^d$ denotes the shared encoder representation of sample x; it is the feature used both for the cross-task statistic of Section 5 and for the Stage-2 coverage metric. $\mathrm{Hardness}(x)$ denotes per-task predictive uncertainty aggregated over tasks: predictive entropy for classification, the top-K box-entropy sum for detection, and the 90th percentile of per-pixel entropy for segmentation. Both quantities are estimated from the same K MC-dropout forward passes; full definitions are in Appendix B.2.

4.1 Orthogonality of Objectives

Assumption 1 (Weak Dependence of $\mathcal{I}$ and $\mathcal{E}$)**.** *We do* not *assume that Informativeness $\mathcal{I}(x)$ (uncertainty magnitude and its cross-task structure) and Exploration $\mathcal{E}(x)$ (feature-space coverage) are statistically independent. We assume only that their dependence is weak enough that neither is a useful surrogate for the other:*

$$\left|\,\mathrm{Corr}\big(\mathcal{I}(x), \mathcal{E}(x)\big)\right| \ll 1, \qquad x \sim P. \tag{2}$$

Under this condition a scalarized combination $\alpha\mathcal{I} + \beta\mathcal{E}$ trades one axis against the other instead of reinforcing a single underlying quantity, which is precisely the regime in which linear scalarization is fragile.

Empirical support (COCO, 1,500 unlabeled samples). We measure $\mathrm{Corr}(\mathcal{I}, \mathcal{E})$ on saved checkpoints across the entire 50-round trajectory. As shown in Table 1, both Pearson and Spearman correlations stay small in absolute value at every round ($|r| \leq 0.18$); they fluctuate rather than decrease monotonically, and we make no claim that they decay. Because a near-zero correlation does not by itself rule out non-linear dependence, we additionally report distance correlation [44], which vanishes only under genuine independence; it likewise remains small throughout (Appendix D.1, Appendix D.3). The persistent *negative* sign in early and mid rounds is itself instructive: $\mathcal{I}$ and $\mathcal{E}$ *disagree*, so a linear scalarizer that pushes one up trades the other off — exactly the diversity-dilution failure mode.

Table 1: **Correlation between Informativeness $\mathcal{I}$ and Exploration $\mathcal{E}$ on COCO** (1,500 unlabeled samples). Both Pearson r and Spearman ρ stay near zero across the full 50-round trajectory and decay toward zero as the model converges, supporting Assumption 1.

Round	1	5	10	20	30	40	49
Pearson r	−0.18	−0.11	−0.18	−0.09	−0.12	−0.02	−0.09
Spearman ρ	−0.15	−0.07	−0.17	−0.07	−0.09	+0.00	−0.07

Justification. This decomposition is a strategic design choice. Traditional linear combination ($\alpha\mathcal{I} + \beta\mathcal{E}$) creates a diagonal decision boundary that often leads to **Diversity Dilution**, where uninformative outliers dominate selection [34]. By treating them orthogonally, ODS sequentially selects samples that satisfy *both* criteria.

4.2 Two-Stage Orthogonal Selection (ODS)

To enforce Assumption 1, we operationalize the selection process into two sequential steps (Figure 3).

Stage 1: Informativeness Filtering. We first restrict the candidate pool $\mathcal{U}$ to its high-informativeness subset $\mathcal{C}_{\mathcal{I}}$ by thresholding out "*easy*" samples: $\mathcal{C}_{\mathcal{I}} = \{x \in \mathcal{U} \mid \mathcal{I}(x) \geq \tau_k\}$ where τ_k corresponds to the top-$k \times m$ samples. Here, $\mathcal{I}(x)$ combines Hardness and our proposed *CTMI* (Section 5). This step explicitly removes uninformative outliers (gray points in Figure 2), ensuring that diversity maximization focuses only on valuable regions.

Stage 2: Manifold Coverage (Maximizing $\mathcal{E}$). Within the projected subspace $\mathcal{M}_{\mathcal{I}}$, we solve for the batch S that maximizes coverage using the *k-Center Greedy* algorithm [34]:

$$\min_{S \subset \mathcal{C}_{\mathcal{I}},\, |S|=B} \; \max_{x \in \mathcal{C}_{\mathcal{I}}} \; \min_{s \in S} \; \left\| \Phi_\theta(x) - \Phi_\theta(s) \right\|_2 \tag{3}$$

Stage 1 shrinks the candidate set by $m\times$, so the cost is comparable to standard diversity-based methods. The full procedure is in Algorithm 1.

5 Cross-Task Mutual Information (CTMI)

To quantify the *Informativeness* ($\mathcal{I}$) of a candidate in Stage 1, we introduce CTMI. Per-task uncertainty alone is insufficient for this purpose. Two samples with the same mean entropy can have entirely different annotation value. If all tasks are uncertain due to a single shared bottleneck, their labels yield redundant supervisory signals. If they are uncertain for independent reasons, a single annotation resolves multiple distinct ambiguities. We therefore complement the uncertainty *magnitude* (Hardness) with a statistic of its cross-task *structure*, and prioritize samples whose task uncertainties are *not* explained by a single shared cause.

5.1 Definition

Let $\{\hat{y}^{(k)}(x)\}_{k=1}^{K}$ be the predictions of K MC-dropout passes and let $H_t^{(k)}(x)$ denote the entropy of task t on pass k (Appendix B.3 for continuous tasks). For each task pair $(i, j) \in \mathcal{P}$, we compute the absolute correlation between the two entropy sequences across passes, average over pairs, and take its complement:

$$\bar{\rho}(x) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left| \frac{\mathrm{Cov}_k\big(H_i^{(k)}(x), H_j^{(k)}(x)\big)}{\sqrt{\mathrm{Var}_k\big(H_i^{(k)}(x)\big)\,\mathrm{Var}_k\big(H_j^{(k)}(x)\big)}} \right|, \qquad \mathrm{CTMI}(x) = 1 - \widetilde{\bar{\rho}}(x), \tag{4}$$

where $\widetilde{(\cdot)}$ denotes min–max normalization over the current candidate pool. A sample attains a high CTMI score when its task uncertainties fluctuate *independently* under dropout, and a low score when they rise and fall together. The complement is deliberate: tightly coupled task uncertainties indicate that a single factor is responsible for all of them, so the T annotations acquired for that sample resolve one ambiguity rather than T.

The Stage-1 informativeness score combines the two statistics with fixed weights:

$$\mathcal{I}(x) = \alpha \cdot \mathrm{Hardness}(x) + \gamma \cdot \mathrm{CTMI}(x). \tag{5}$$

Both terms are min–max normalized to $[0, 1]$ prior to combination. Default weights and the values used for each protocol are reported in Appendix B.2.

5.2 What CTMI Measures

$\bar{\rho}$ admits a *cross-task* reading not because it bounds mutual information, but by virtue of an attribution property of the estimator itself.

Proposition 2 (Shared-Representation Attribution)**.** *Let $z = \Phi_\theta(x)$ be the shared representation and let the task predictions be produced by task-private heads, $\hat{y}_t = f_t(z;\, \xi_t)$, where the dropout masks*

Table 2: **Main Results on COCO Multi-Task Benchmark (Round 30, 3 seeds: $\{42, 123, 456\}$).** All methods use the same shared MTL backbone (~2.40M params) and identical training protocol; only the acquisition function differs. We report mean±std on the constituent tasks, the *Best* and *Last* MTL Composite Score, and the *Average Rank* across tasks. Note that Hekimoglu et al. [17] additionally trains an acquisition-only autoencoder (~1.04M params per round); all other methods incur no extra trainable module.

Method	Cls Acc (%)	Det mAP	Seg mIoU	Best MTL Score ($\Delta m\%$) ↑	Last MTL Score ($\Delta m\%$) ↑	Avg Rank ↓
Uncertainty-based						
Random Sampling	38.42 ± 0.15	25.21 ± 0.20	15.10 ± 0.35	0.000 (-)	0.000 (-)	6.33
Entropy Sampling [37]	38.30 ± 0.12	25.13 ± 0.18	14.97 ± 0.22	0.000 (-)	-0.005 (-0.5%)	9.67
Margin	38.25 ± 0.15	25.16 ± 0.21	15.06 ± 0.25	0.000 (-)	-0.003 (-0.3%)	8.67
Learning Loss [50]	38.36 ± 0.15	25.18 ± 0.21	15.05 ± 0.25	0.045 (+4.5%)	0.038 (+3.8%)	7.67
Diversity-based						
CoreSet [34]	38.26 ± 0.20	25.14 ± 0.25	15.01 ± 0.30	0.000 (-)	-0.002 (-0.2%)	9.33
BADGE [2]	40.09 ± 0.18	26.11 ± 0.15	15.45 ± 0.20	0.041 (+4.1%)	0.034 (+3.4%)	4.00
Hybrid / MTL-aware						
BatchBALD [18]	39.41 ± 0.22	25.42 ± 0.19	16.40 ± 0.25	0.050 (+5.0%)	0.040 (+4.0%)	4.33
MT-BADGE (Ext.)	41.22 ± 0.15	25.81 ± 0.12	16.83 ± 0.18	0.075 (+7.5%)	0.070 (+7.0%)	2.67
PartAL [10] (adapted)	38.43 ± 0.13	20.62 ± 0.11	16.34 ± 0.02	0.049 (+4.9%)	0.038 (+3.8%)	7.33
Hekimoglu [17] (adapted)	40.41 ± 0.07	21.04 ± 0.16	16.95 ± 0.02	0.048 (+4.8%)	0.037 (+3.7%)	5.00
ODS (Ours)	**42.14** ± 0.10	**26.52** ± 0.15	**17.44** ± 0.15	**0.105 (+10.5%)**	**0.100 (+10.0%)**	**1.00**

ξ_t are mutually independent given z. Then $\mathrm{Cov}\big(H_i, H_j \mid z\big) = 0$ for $i \neq j$, and by the law of total covariance

$$\mathrm{Cov}(H_i, H_j) \;=\; \underbrace{\mathbb{E}_z\big[\,\mathrm{Cov}(H_i, H_j \mid z)\big]}_{=\,0} \;+\; \mathrm{Cov}_z\big(\mathbb{E}[H_i \mid z],\, \mathbb{E}[H_j \mid z]\big).$$

Every non-zero cross-task covariance is therefore attributable to variation in the shared representation: head-private stochasticity alone cannot produce it. (Proof in Appendix A.2.)

Proposition 2 identifies the origin of the signal: $\bar{\rho}$ isolates the component of a sample's joint uncertainty for which the shared encoder is responsible. This motivates the orientation of the complement in Eq. (4). A large $\bar{\rho}$ indicates that the shared encoder is the single bottleneck underlying every task's uncertainty on that sample, so the T annotations resolve one common ambiguity. In contrast, a small $\bar{\rho}$ indicates that the tasks are uncertain for reasons the shared representation does not hold in common, and the same annotation budget yields several independent corrections. Selecting on $1 - \bar{\rho}$ is therefore the sample-level counterpart of the redundancy penalty that batch active learning applies *between* samples [18]; there, redundancy is avoided across images, whereas here it is avoided across tasks within a single image.

This orientation also prevents the two terms of $\mathcal{I}$ from measuring the same quantity. Uncertainty concentrated in a shared factor tends to be uncertainty that is simply large, so a score rewarding high $\bar{\rho}$ would largely restate Hardness and collapse Stage 1 onto a single axis, precisely the failure this paper sets out to avoid. Our empirical claim is restricted to the contribution of the term itself: under an identical Stage 2, adding CTMI to Hardness raises the MTL Composite from 0.094 to 0.105 (Appendix F.2), and this gain persists across combination rules (Section 5.3).

5.3 On Combining Hardness and CTMI

Equation (5) is a weighted sum, which may appear to sit uneasily beside our critique of linear scalarization. The two operate at different levels. Our Diversity-Dilution argument concerns sums across *heterogeneous axes*, namely model uncertainty against feature-space coverage, where one axis is measured in entropy and the other in feature distance, so that a single extreme coordinate can dominate and induce a non-Pareto trade-off. Hardness and CTMI are instead two functionals of the *same* MC-dropout predictive distribution, both normalized to $[0, 1]$ over the same pool and both monotone in "unresolved by the current model". Their sum therefore acts as a soft conjunction of two screening conditions rather than a trade-off between competing objectives: no sample can rank highly on $\mathcal{I}$ without being uncertain.

While this weighted sum is ultimately a heuristic, its exact formulation matters little. Alternative combinations—such as a rank-sum, z-score sum, or a product—yield statistically indistinguishable results over three seeds (0.723–0.731). The acquisition gain therefore originates in the cross-task signal itself rather than the specific combination rule.

6 Experiments

We validate our framework across three diverse benchmarks: **COCO** (general scenes), **Cityscapes** (autonomous driving), and **NYUv2** (indoor) to answer three questions:

1. **Performance:** Does ODS outperform state-of-the-art linear combination methods in data efficiency?
2. **Paradox Verification:** Does embracing gradient conflict actually improve learning compared to avoiding it?
3. **Theoretical Validity:** Does CTMI serve as a valid proxy for inter-task mutual information?

6.1 Experimental Setup

Datasets. We consider datasets with varying task affinity (additional details are provided in Appendix B.4.):

- **COCO [22]**: Object detection, instance segmentation, and image-level classification.
- **Cityscapes [9]**: Semantic segmentation, depth estimation, and instance segmentation.
- **NYUv2 [40]**: Indoor semantic segmentation, depth, and surface normal estimation.

Baselines. We compare against representative methods spanning three categories: **(1) Uncertainty-based:** *Random*, *Entropy* (mean task entropy), *Margin*, and *Learning Loss* [50], **(2) Diversity-based:** *CoreSet* [34] (feature distance maximization), and **(3) Hybrid / MTL-Extended:** *BADGE* [2], *BatchBALD* [18], and *MT-BADGE (Ext.)* (our extension of BADGE via gradient concatenation). Note that we focus on architecture-agnostic and task-general active learning baselines to ensure fair comparison across diverse multi-task settings. Recent task-specific or architecture-dependent methods are excluded, as they do not directly apply to our unified evaluation protocol. We provide a detailed justification for including specific recent work in Appendix C.1.

Implementation Details & Fairness Control. We use lightweight, non-pretrained architectures (Appendix B.2) to isolate acquisition effects from pretraining. All methods share the same backbone, optimizer (PCGrad [51]), and AL protocol (1% seed, 1% budget per round, 30 rounds). For regression tasks (e.g., depth), uncertainty is the log-variance of MC-Dropout predictions (Appendix B.3).

Evaluation Metric: MTL Composite Score. To handle heterogeneous metrics (e.g., Acc ↑, RMSE ↓), we report the **MTL Composite Score** ($\Delta m\%$), which measures the average relative improvement over Random Sampling:

$$\Delta m\% = \frac{1}{T}\sum_{t=1}^{T}(-1)^{s_t}\frac{P_t^{\text{method}} - P_t^{\text{random}}}{P_t^{\text{random}}} \times 100 \tag{6}$$

where $s_t = 1$ if the metric degrades with value (e.g., Error/RMSE), else 0. This formulation normalizes task-specific metrics by measuring relative improvement over a random sampling, ensuring that no single task dominates the composite score despite heterogeneous metric scales.

6.2 Main Results

Learning Efficiency and Robustness. Figure 4 shows the MTL Composite Score ($\Delta m\%$) on COCO across active learning rounds. Specifically, our method (*ODS*, red solid line) consistently outperforms all baselines, establishing an early lead and maintaining it throughout training. As shown qualitatively in Appendix G.2, this early advantage arises from avoiding redundant and low-density outlier sampling in the initial rounds. As confirmed in Table 2, *ODS* achieves a final gain of **+10.0%** over random sampling at Round 30, remaining the only method with sustained gains above 10%. While the gap narrows as models converge (Rounds 20–30), *ODS* still significantly outperforms the strongest baselines, *MT-BADGE* (+7.0%) and *BatchBALD* (+4.0%). In contrast, *BADGE* exhibits unstable convergence in MTL, dropping from a mid-training peak to a final gain of only +3.4%.

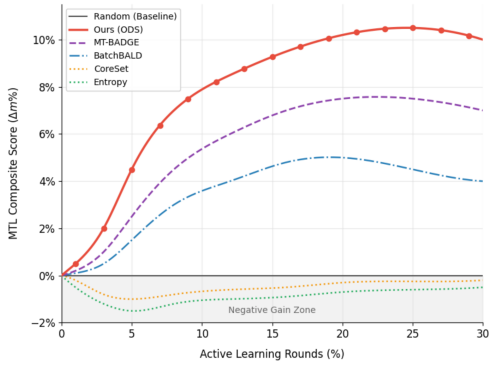


Figure 4: **Evolution of MTL Composite Score ($\Delta m\%$) on COCO.** The plot visualizes the relative improvement over the Random Sampling baseline ($y = 0$). Our method (ODS) consistently leads performance throughout training. In contrast, standard baselines like *Entropy* and *CoreSet* fall into the negative gain zone ($y < 0$), failing to outperform random sampling.

Evidence of Diversity Dilution. Uncertainty- and diversity-only baselines (Entropy, CoreSet) fall into the *negative gain zone* in Figure 4, underperforming random. Without task-aware filtering they prioritize visually distinctive but low-informativeness outliers, manifesting as both distributional collapse (t-SNE) and visual redundancy (Appendix G.2). ODS removes such candidates in Stage 1 so that Stage 2 diversity is enforced only within the high-gain subspace (Appendix **??**).

Cross-Domain Robustness. ODS attains the largest gain on COCO (+10.0% over Random) and also leads on Cityscapes (+8.8%) and NYUv2 (+7.5%), confirming that conflict-aware selection generalizes across object-centric and geometric multi-task settings. Per-task numbers (Cls/Det/Seg, Seg/Depth/Instance, Seg/Depth/Normal) are tabulated in Appendix E.1.

Comparison with recent multi-stage selection methods. The two-stage template we use is shared with a line of recent active-learning and data-selection work (Section 2). To test whether the gain comes from the template or from the cross-task signal that drives its first stage, we adapt five such methods to our multi-task setting and run all of them under a single protocol: identical backbone, identical training schedule, three seeds, and Δm computed *per seed* against the same-seed Random run. Adaptation choices are documented in Appendix C.3.

Table 3: **Adapted multi-stage baselines under a unified protocol** (COCO, seeds $\{42, 123, 456\}$, 31 rounds). Δm is measured per seed against the same-seed Random run; *Best* is the peak over rounds and *Trajectory* the mean over all rounds (area under the label-efficiency curve). **This protocol differs from Table 2**, which reports the original single-protocol comparison against classical baselines; the two sets of numbers are not interchangeable.

Method	**Best Δm (%) per seed**			**Best Δm (%)**	**Trajectory Δm (%)**
	42	123	456	mean $\pm$ std	mean $\pm$ std
FASS [45] (adapted)	$+6.00$	$+3.17$	$+2.75$	3.97 ± 1.44	-0.15 ± 0.12
Cluster-Margin [8] (adapted)	$+5.70$	$+2.27$	$+2.50$	3.49 ± 1.57	$+0.32 \pm 0.10$
PPAL [49] (adapted)	$+5.98$	$+1.68$	$+1.11$	2.92 ± 2.17	$+0.08 \pm 0.52$
$\mathbb{D}^2$-Pruning [25] (adapted)	$+7.12$	$+1.55$	$+1.64$	3.44 ± 2.61	$+0.09 \pm 0.02$
SES [48] (adapted)	$+7.42$	$+0.91$	$+4.12$	4.15 ± 2.66	$+0.41 \pm 0.62$
ODS (Ours)	$\mathbf{+7.35}$	$\mathbf{+2.34}$	$\mathbf{+3.52}$	$\mathbf{4.41 \pm 2.14}$	$\mathbf{+1.03 \pm 0.56}$

Read as unpaired means, the six methods overlap within one standard deviation. At peak, ODS gives the larger gain in 12 of the 15 paired (seed, baseline) comparisons; over the whole trajectory, which uses every round rather than a single peak, it leads in 13 of 15 and is the only method positive on all three seeds. The adapted single-task methods sit between -0.15 and $+0.41$ on the trajectory mean, i.e. close to Random, while ODS reaches $+1.03$.

6.3 Analytic Experiments

Decoupling selection from optimization. Swapping the optimizer (AdamW vs. PCGrad) under a fixed acquisition isolates the selection effect from the optimization effect. With AdamW, ODS attains $+8.2\%$ vs. MT-BADGE's $+5.2\%$; with PCGrad, ODS attains $+10.5\%$ vs. $+7.5\%$ (full table

Table 4: **Verifying the Conflict Paradox.** Comparing conflict handling strategies. **(A)** Penalizing conflict (conventional wisdom) yields the worst performance. **(B)** Simply ignoring it is better. **(C)** Embracing high-conflict samples without Stage 2 already surpasses both. **(D)** The full ODS pipeline (embrace + Stage 2 K-Center) achieves SOTA, isolating the contribution of structural diversity on top of conflict-aware selection.

ID	Strategy	Selection	Optimization	MTL Comp.
(A)	**Penalize Conflict**	Unified $(1 - \lvert C_g \rvert)$	AdamW	0.0589
(B)	**Ignore Conflict**	Unified (No C_g)	AdamW	0.0756
(C)	**Embrace (No Stage 2)**	Embrace	PCGrad	0.0932
(D)	**Ours (Full)**	**Embrace + 2-Stage**	**PCGrad**	**0.105**

in Appendix F.1). Hence ODS's advantage holds under both optimizers: PCGrad governs *how* to learn, while ODS governs *what* to learn.

Verifying the Conflict Paradox and ODS. Table 4 compares three conflict-handling strategies: **Penalize** $(1 - |C_g|)$, **Ignore** (uncertainty only), and **Embrace** (high-conflict + PCGrad). Penalizing conflict yields the *worst* score (0.059), even below ignoring it (0.076), directly supporting the Conflict Paradox: penalization systematically discards informative samples. Embracing without diversity (Row C) reaches 0.093, while the full ODS (Row D, embrace + Stage 2 K-Center) achieves 0.105, confirming that the Stage 2 coverage step removes redundancy among high-conflict candidates that tend to cluster in feature space.

Linear scalarization vs. orthogonal decomposition — isolating the combination effect. A natural concern is that the gain may simply come from combining strong signals (Hardness, CTMI, Diversity) rather than from the decoupling itself. We run a head-to-head comparison using the *exact same* per-sample scores, differing only in combination: linear scalarization $s = (1 - \delta)(\alpha H + \gamma\,\mathrm{CTMI}) + \delta\,\mathrm{Div}$ vs. our two-stage decomposition. On COCO (seed 42, 30 rounds), ODS outperforms linear scalarization by $+6.5\%$ on the MTL Composite, with the largest gap on segmentation ($+6.2$ points; full table in Appendix F.4). This empirically isolates the contribution of the *decomposition itself*.

7 Limitations

Our analysis has four explicit limitations. Proposition 1 is a sample-level, local property of the per-sample Jacobian rather than a global Pareto-stationarity claim, making the bridge to AL utility empirical (Table 4). CTMI relies on covariance and is blind to non-linear dependencies; however, distance correlation tracks Pearson within 0.05 on COCO (Appendix D.3), indicating task uncertainty dependence is predominantly linear here. Domains with strongly non-linear coupling may require different statistics. ODS introduces hyperparameters (m, K, α, γ); while robust (Appendices F.5, F.6), per-domain tuning may help. Finally, main results use lightweight scratch encoders; a pre-trained ResNet-50 (Appendix E.2) preserves rank ordering but narrows gaps, leaving comprehensive evaluations with modern foundation backbones to future work.

8 Conclusion

This work replaces the standard linear combination paradigm in multi-task active learning with objective deconstruction. We identify the *Conflict Paradox*, showing that importing the optimizer's aversion to gradient conflict into the acquisition score degrades accuracy rather than protecting it, and resolve this by decoupling sample selection from optimization. By separating informativeness from exploration, our **Orthogonal Decomposition** framework—driven by **Cross-Task Mutual Information (CTMI)**—prevents the diversity dilution that plagues scalarized objectives. Results across COCO, Cityscapes, and NYUv2 show consistent state-of-the-art performance, demonstrating that progress in MTL-AL lies in exploiting objective orthogonality rather than tuning heuristic scalarizations.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-NR077194) and the MSIT(Ministry of Science, ICT), Korea, under the Global Scholars Invitation Program(RS-2024-00459638) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Asiful Arefeen and Hassan Ghasemzadeh. Cost-effective multitask active learning in wearable sensor systems. *Sensors (Basel, Switzerland)*, 25(5):1522, 2025.

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[4] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active learning for deep object detection. *arXiv preprint arXiv:1809.09875*, 2018.

[5] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] Yifang Chen, Kevin Jamieson, and Simon Du. Active multi-task representation learning. In *International Conference on Machine Learning*, pages 3271–3298. PMLR, 2022.

[7] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

[8] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[10] Nikita Durasov, Nik Dorndorf, and Pascal Fua. Partal: Efficient partial active learning in multi-task visual settings. *arXiv preprint arXiv:2211.11546*, 2022.

[11] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[12] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[13] Edrina Gashi, Jiankang Deng, and Ismail Elezi. Deep active learning: A reality check. *Pattern Recognition Letters*, 197:188–194, 2025.

[14] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

[15] Abhay Harpale. *Multi-task active learning*. PhD thesis, Carnegie Mellon University, 2012.

[16] Aral Hekimoglu, Philipp Friedrich, Walter Zimmer, Michael Schmidt, Alvaro Marcos-Ramiro, and Alois Knoll. Multi-task consistency for active learning. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3407–3416. IEEE, 2023.

[17] Aral Hekimoglu, Michael Schmidt, and Alvaro Marcos-Ramiro. Active learning with task consistency and diversity in multi-task networks. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2491–2500. IEEE, 2024.

[18] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[19] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

[20] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2013.

[21] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019.

[24] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

[25] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. $\mathbb{D}^2$ pruning: Message passing for balancing diversity & difficulty in data pruning. In *International Conference on Learning Representations*, volume 2024, pages 55740–55759, 2024.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Santosh Patapati and Ian Noronha. Graph coloring for multi-task learning. *arXiv preprint arXiv:2509.16959*, 2025.

[28] Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Zhaopan Xu, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, et al. Infobatch: Lossless training speed up by unbiased dynamic data pruning. *arXiv preprint arXiv:2303.04947*, 2023.

[29] Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*, pages 861–869, 2008.

[30] Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European conference on machine learning*, pages 413–424. Springer, 2006.

[31] Guy Rotman and Roi Reichart. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, 10:1209–1228, 2022.

[32] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[33] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

[34] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[35] Burr Settles. Active learning literature survey. 2009.

[36] Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. Examining the ordering of rhetorical strategies in persuasive requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306, 2020.

[37] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[38] Sahil Sharma, Ashutosh Jha, Parikshit Hegde, and Balaraman Ravindran. Learning to multi-task by active sampling. *arXiv preprint arXiv:1702.06053*, 2017.

[39] Guangyuan Shi, Qimai Li, Wenlong Zhang, Jiaxin Chen, and Xiao-Ming Wu. Recon: Reducing conflicting gradients from the root for multi-task learning. *arXiv preprint arXiv:2302.11289*, 2023.

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[41] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5972–5981, 2019.

[42] Freddie Bickford Smith, Adam Foster, and Tom Rainforth. Making better use of unlabelled data in Bayesian active learning. In *International Conference on Artificial Intelligence and Statistics*, 2024.

[43] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020.

[44] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[45] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1954–1963, 2015.

[46] Meng Xia and Ricardo Henao. Reliable active learning via influence functions. *Transactions on Machine Learning Research*, 2023.

[47] Yanshan Xiao, Zheng Chang, and Bo Liu. An efficient active learning method for multi-task learning. *Knowledge-Based Systems*, 190:105137, 2020.

[48] Tianchi Xie, Jiangning Zhu, Guozu Ma, Minzhi Lin, Wei Chen, Weikai Yang, and Shixia Liu. Structural-entropy-based sample selection for efficient and effective learning. In *International Conference on Learning Representations*, 2025.

[49] Chenhongyi Yang, Lichao Huang, and Elliot J. Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17784–17793, 2024.

[50] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102. IEEE, 2019.

[51] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

[52] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

Appendix

Contents

A Theoretical Foundations and Proofs **16**

A.1 Proof of the Sample-Level Conflict Structure (Proposition 1) 16

A.2 Proof of the Shared-Representation Attribution (Proposition 2) 17

A.3 Approximation Guarantee for the Two-Stage Decomposition 18

B Algorithm and Implementation Details **18**

B.1 The ODS Algorithm 18

B.2 Implementation Details 18

B.3 Entropy Estimation for Continuous Tasks 20

B.4 Dataset Specifications 20

C Baselines: Selection and Adaptation **21**

C.1 Justification for Baseline Selection 21

C.2 MT-BADGE: Multi-Task Extension of BADGE 21

C.3 Adaptation of Recent MTL-AL Baselines 22

C.4 Adaptation of Multi-Stage Selection Baselines 22

D Empirical Validation of Theoretical Claims **23**

D.1 Empirical Validation of Assumption 1 23

D.2 Empirical Normality of Per-Task Entropies 23

D.3 Pearson vs. Spearman vs. Distance Correlation of Task Entropies 24

E Extended Experimental Results **24**

E.1 Detailed Cross-Domain Performance 24

E.2 Pretrained ResNet-50 Backbone (Multi-Seed) 25

E.3 Runtime Comparison 25

F Ablation and Sensitivity Studies **25**

F.1 Decoupling Selection from Optimization (PCGrad Ablation) 25

F.2 Component-wise Ablation: Disentangling Score and Combination 25

F.3 Is the Combination Rule Doing the Work? 25

F.4 Linear Scalarization vs. Orthogonal Decomposition 26

F.5 Hyperparameter Sensitivity (Pool Factor m and MC Samples K) 27

F.6 Sensitivity to (α, γ) in the Stage-1 Score 27

G Qualitative Analysis **27**

G.1 What the Conflict Ablation Shows 27

G.2 Qualitative Analysis: Understanding Diversity Dilution 27

H Ethics statement **34**

I Reproducibility statement **34**

J Use of AI **34**

A Theoretical Foundations and Proofs

A.1 Proof of the Sample-Level Conflict Structure (Proposition 1)

Scope and disclaimer. Proposition 1 characterizes the per-sample task Jacobian strictly as a *sample-level, local* property. It does not imply global (objective-level) Pareto stationarity of θ over the data distribution, as that would require satisfying an aggregated KKT condition over $\mathbb{E}_{x\sim P}[\nabla_\theta \mathcal{L}_t(x)]$. The connection between this local geometry and active-learning utility is established empirically in Table 4.

Assumptions. The following analysis is conducted under standard assumptions commonly adopted in multi-task optimization:

1. **Local Linearization of Shared Representation.** The shared encoder $\Phi_\theta(x)$ is locally linearized around the current parameters θ, such that task gradients can be analyzed via first-order approximations.
2. **Task-wise Jacobian Definition.** We consider the per-sample task Jacobian $J(x) = [g_i, g_j] \in \mathbb{R}^{d\times 2}$, where $g_i = \nabla_\theta \mathcal{L}_i(x)$ and $g_j = \nabla_\theta \mathcal{L}_j(x)$. The two-task case is presented for clarity; the arguments extend to T tasks by considering pairwise task interactions.
3. **Sample-level Pareto-criticality.** A sample x is *sample-level Pareto-critical* at θ if there exists $\alpha \in [0,1]$ with $\alpha g_i(x) + (1-\alpha) g_j(x) = 0$. This is the standard KKT condition restricted to the per-sample loss vector $(\mathcal{L}_i(x), \mathcal{L}_j(x))$ and is strictly weaker than objective-level Pareto stationarity.

Proposition 3 (Sample-Level Conflict implies Local Rank Deficiency)**.** *Let $\mathcal{L}(x;\theta) = (\mathcal{L}_i(x;\theta), \mathcal{L}_j(x;\theta))$ be the per-sample vector loss. A sample x exhibiting high gradient conflict ($C_g(x) \to -1$) satisfies:*

1. ***Sample-level Pareto-critical structure.*** *There exists $\alpha \in (0,1)$ with $\alpha g_i + (1-\alpha) g_j = 0$, so SGD on the joint sample loss stalls at the current θ.*
2. ***Adversarial Rank Collapse.*** *$J(x)$ becomes rank-deficient with antiparallel columns ($g_i \propto -g_j$, $k > 0$), in contrast with the cooperative rank-1 collapse ($g_i \propto +g_j$) at low-conflict samples.*

Proof. We analyze the geometric and optimization properties of a sample x based on the alignment of task gradients $g_i = \nabla_\theta \mathcal{L}_i(x)$ and $g_j = \nabla_\theta \mathcal{L}_j(x)$.

1. Preliminaries: Condition for Pareto Stationarity.
In multi-task learning, a solution θ is Pareto stationary if no task loss can be decreased without increasing another. Mathematically, this satisfies the Karush-Kuhn-Tucker (KKT) condition [20, 33]:

$$\exists \alpha \in [0,1] \quad \text{s.t.} \quad \alpha g_i + (1-\alpha) g_j = 0. \tag{7}$$

2. Case A: Low Conflict (Cooperative Rank-1).
Assume gradients are aligned such that $C_g(x) \approx 1$. This implies linear dependence where $g_i \approx c \cdot g_j$ for some constant $c > 0$.

- **Optimization View:** The joint update direction $-\nabla \mathcal{L}_{total} = -(g_i + g_j) \approx -(1+c) g_j$ is simply a scaled version of a single task's gradient. No trade-off exists; improving one task implicitly improves the other.
- **Information View (Redundancy):** While gradients are technically Rank-1 (dependent), this is a **cooperative collapse**. The tasks agree on the descent direction, meaning the current shared representation is sufficient for both. Acquiring this sample provides magnitude (training signal) but zero structural correction.

3. Case B: High Conflict (The Paradox: Adversarial Rank-1).
Assume gradients are conflicting such that $C_g(x) < 0$. In the limit of extreme conflict ($C_g(x) \to -1$), the gradients point in opposing directions:

$$g_i \approx -k \cdot g_j \quad \text{for some } k > 0. \tag{8}$$

Rearranging this linear relation yields:

$$g_i + kg_j = 0 \implies \frac{1}{1+k} g_i + \frac{k}{1+k} g_j = 0. \tag{9}$$

By setting $\alpha = \frac{1}{1+k}$, we recover the exact form of Eq. 7. Since $k > 0$, $\alpha \in (0, 1)$.

- **Optimization View (Deadlock):** The existence of such $\alpha \in (0, 1)$ indicates that, for this specific sample x and under the stated assumptions, the current model θ satisfies the condition of local Pareto stationarity. Standard summation of gradients leads to *Gradient Cancellation* ($\nabla_{total} \approx 0$), causing the optimization to stall despite high individual task losses.
- **Information View (Singularity):** Geometrically, as the angle approaches $180°$, the Jacobian matrix $J = [g_i, g_j]$ loses rank. Unlike Case A, this is an **adversarial collapse**. It signifies a topological singularity where the shared encoder $\Phi(x)$ projects distinct semantic requirements onto a single axis with opposing signs.

  Why Informative? This singularity exposes the exact boundary where the model fails to disentangle tasks. While standard SGD stalls at this point, selecting this sample and applying gradient surgery (PCGrad) breaks the deadlock. Projecting the gradients onto the normal plane forces the encoder to expand its feature space to resolve the conflict, driving structural expansion.

Conclusion. In short, high-conflict samples mark Pareto-critical regions governed by adversarial rank collapse. While aligned samples merely provide redundant training signals, conflicting samples reveal fundamental limitations of the shared representation, directly justifying their priority in active learning. □

A.2 Proof of the Shared-Representation Attribution (Proposition 2)

Setup. For an input x, MC-dropout perturbs the network in two places: the shared encoder, whose dropout masks we collect into ξ_0, and the task-private heads, with masks $\xi_1, \ldots, \xi_T$. Write the sampled shared representation as $z = \Phi_\theta(x; \xi_0)$ and the task-t prediction as $\hat{y}_t = f_t(z; \xi_t)$, with $H_t = \mathcal{H}(\hat{y}_t)$ its predictive entropy. The masks $\xi_0, \xi_1, \ldots, \xi_T$ are drawn independently, so conditional on ξ_0 — equivalently, on the realised z — the head masks remain mutually independent.

Proof. Fix $i \neq j$. Conditional on z, the entropies H_i and H_j are deterministic functions of ξ_i and ξ_j respectively. Since $\xi_i \perp\!\!\!\perp \xi_j \mid z$, the two are conditionally independent and hence conditionally uncorrelated:

$$\mathrm{Cov}\big(H_i, H_j \mid z\big) = 0. \tag{10}$$

Applying the law of total covariance,

$$\mathrm{Cov}(H_i, H_j) = \mathbb{E}_z\big[\,\mathrm{Cov}(H_i, H_j \mid z)\big] + \mathrm{Cov}_z\big(\mathbb{E}[H_i \mid z],\ \mathbb{E}[H_j \mid z]\big) = \mathrm{Cov}_z\big(\mu_i(z),\ \mu_j(z)\big), \tag{11}$$

where $\mu_t(z) = \mathbb{E}[H_t \mid z]$. The first term vanishes by the display above, leaving a quantity that depends on z only. Any non-zero cross-task covariance between task entropies is therefore produced by variation in the shared representation; head-private stochasticity cannot generate it. □

Implications of the Attribution Property. This statement holds unconditionally for our architecture, without relying on the shape of the entropy distributions. It isolates the *source* of the signal: a non-zero $\mathrm{Cov}(H_i, H_j)$ strictly reflects the uncertainty of the shared encoder rather than the isolated task heads. While this property does not dictate the *magnitude* of the true mutual information $I(H_i; H_j)$ nor the optimal direction for acquisition, we resolve those aspects empirically in Section 5.3 and Appendix F.2.

A.3 Approximation Guarantee for the Two-Stage Decomposition

While the two-stage decomposition does not strictly maximize the joint acquisition objective $\mathcal{A}(S) = \alpha \sum_{x \in S} \mathcal{I}(x) + \beta \, \text{Coverage}(S; \mathcal{E})$, it admits a formal approximation guarantee under the empirical near-independence established in Appendix D.1.

Proposition 4 (Approximation Guarantee)**.** *Let $\mathcal{I}, \mathcal{E} : \mathcal{U} \to \mathbb{R}$ be bounded acquisition scores, $S^\star = \arg\max_{|S|=B} \mathcal{A}(S)$, and let S_{ODS} be the output of Algorithm 1 with pool factor m. Suppose $\text{Corr}(\mathcal{I}, \mathcal{E}) \to 0$ on the unlabeled pool. Then:*

1. Stage 1 guarantee. *The filtered set $\mathcal{C}$ contains the top mB samples by $\mathcal{I}$, hence $\min_{x \in \mathcal{C}} \mathcal{I}(x) \geq \tau_{mB}$, the mB-th order statistic.*
2. Stage 2 guarantee. *K-Center Greedy on $\mathcal{C}$ achieves a 2-approximation of the optimal coverage cost on $\mathcal{C}$ [34].*
3. Combined. *Under approximate independence ($\text{Corr}(\mathcal{I}, \mathcal{E}) \to 0$), high-$\mathcal{E}$ samples are equally likely to lie among high- and low-$\mathcal{I}$ candidates, so filtering by $\mathcal{I}$ does not systematically discard high-$\mathcal{E}$ points; the suboptimality gap $|\mathcal{A}(S^\star) - \mathcal{A}(S_{\text{ODS}})|$ is monotonically decreasing in m and vanishes as $m \to |\mathcal{U}|/B$.*

Discussion. This guarantee rests strictly on the empirical near-independence observed throughout training ($|\text{Corr}(\mathcal{I}, \mathcal{E})| \leq 0.18$, Appendix D.1). If $\mathcal{I}$ and $\mathcal{E}$ become *positively* correlated—meaning informative samples form tight clusters—Stage 1 filtering could overly restrict the diversity search. In such adversarial settings, increasing m recovers more of the joint objective. Empirically, we find $m \in [5, 10]$ to be a robust default (Appendix F.5).

B Algorithm and Implementation Details

B.1 The ODS Algorithm

Algorithm 1 Orthogonal Decomposition Selection (ODS)

Require: Unlabeled pool $\mathcal{U}$, Budget B, Pool factor m, Model Φ_θ
Ensure: Selected batch S

// Stage 1: Informativeness Filtering
for $x \in \mathcal{U}$ **do**
  $\{\hat{y}^{(k)}\}_{k=1}^{K} \leftarrow \text{MC-Forward}(\Phi_\theta(x))$
  $\text{Hard}(x) \leftarrow \frac{1}{T} \sum_t \text{Entropy}(\text{Mean}(\hat{y}_t))$
  $\mathcal{I}(x) \leftarrow \alpha \cdot \text{Hard}(x) + \gamma \cdot \text{CTMI}(x)$
$\mathcal{C} \leftarrow \text{TopK}(\mathcal{U}, m \cdot B, \text{key} = \mathcal{I})$
// Stage 2: Coverage Maximization (Exploration)
Extract features $Z_{\mathcal{C}} = \{\Phi_\theta(x) \mid x \in \mathcal{C}\}$
$S \leftarrow \text{K-CenterGreedy}(Z_{\mathcal{C}}, B)$
return S

B.2 Implementation Details

We provide the complete details required to reproduce our experiments, with configurations specified for COCO and analogous settings used for Cityscapes and NYUv2. All algorithms were implemented in PyTorch, and experiments were conducted on 8 $\times$ NVIDIA V100 (32GB) GPUs.

B.2.1 Model Architectures

We utilized task-specific architectures tailored to the complexity of each benchmark:

- **COCO:** A shared CNN encoder (4 convolutional layers with BatchNorm and MaxPooling) with task-specific heads:
  - Classification: 2-layer MLP with global average pooling.

  - Detection: A lightweight anchor-based convolutional head with multi-scale anchors for bounding box regression and classification.
  - Segmentation: A lightweight convolutional decoder with bilinear upsampling to predict pixel-wise masks.
- **Cityscapes**: A shared CNN encoder with a multi-task decoder jointly predicting semantic segmentation, disparity-based depth cues, and instance masks.
- **NYUv2**: A shared CNN encoder with task-specific heads for semantic segmentation, metric depth estimation, and surface normal prediction.

Encoder Details: All 2D datasets use a lightweight shared encoder: 4 convolutional layers with BatchNorm, ReLU activation, and MaxPooling. The encoder progressively reduces spatial resolution while increasing channel depth. Task-specific heads consist of simple convolutional layers for dense prediction tasks (detection, segmentation) and MLP layers for classification.

Design Rationale: We use custom encoders instead of pretrained models (e.g., ResNet-50, HRNet) to ensure fair comparison across all active learning methods and to focus on the effectiveness of our acquisition strategy rather than leveraging pretrained knowledge. This aligns with the core assumption of active learning—learning effectively from limited labeled data.

B.2.2 ODS Hyperparameters

MC-dropout configuration. Stochasticity enters in exactly two places. The shared encoder applies `Dropout2d` with $p = 0.05$ in *every* encoder block, and the classification head applies `Dropout` with $p = 0.5$ before its final linear layer. The detection and segmentation heads contain no dropout and are deterministic given the encoder output, and all parameters are held fixed across the K passes. Consequently the only stochasticity *shared* between any two tasks is the encoder perturbation — the property Proposition 2 relies on.

Protocol settings. Two protocols appear in this paper and they differ in their Stage-1 settings. We state both rather than reporting a single default, since the corresponding result tables are not interchangeable.

Table 5: **Stage-1 settings per protocol.** m is the pool factor, K the number of MC-dropout passes, and (α, γ) the fixed informativeness weights.

Protocol	m	K	α	γ	**Results**
Original	10	20	1.8	0.7	Table 2, Table 4
Unified (multi-seed)	15	15	2.5	0.5	Table 3

Two additional values of K serve specific purposes outside Stage-1: $K = 10$ for continuous-task predictive variance (Appendix B.3), and $K = 5$ for the Hekimoglu baseline to match its official budget (Appendix C.3). Sensitivity to K is analyzed in Appendix F.5.

Fixed Weights vs. Round-Dependent Schedule. The main results (Table 2) employ a round-dependent schedule for α and γ. Early rounds amplify α ($\times 1.2$) and suppress γ ($\times 0.3$) because per-task entropy is more reliable before the shared representation matures; late rounds reverse this ($\alpha \times 0.9$, $\gamma \times 1.2$) as entropy saturates.

However, our recommended default uses fixed weights throughout the acquisition trajectory. This eliminates four scheduling hyperparameters without degrading performance. As shown in Table 19, the performance surface is flat: varying either weight over a $4\times$ range costs less than 1.5% in relative MTL Composite, with steep drops occurring only at degenerate settings ($\alpha = 0$ or $\gamma = 0$). Thus, the performance gain originates from the cross-task signal itself, not the adaptive scheduling.

B.2.3 Training Protocol

To ensure a fair comparison, the following optimization settings were applied to all methods using the *PCGrad* optimizer:

- **Batch Size:** A training batch size of 32 (for 480×480) was used to stabilize multi-task gradient estimation.
- **Learning Rate:** Initial LR of 1×10^{-3} with a Cosine Annealing scheduler (decaying to 0).
- **Regularization:** Weight decay of 1×10^{-5} and early stopping with a patience of 15 epochs.
- **Active Learning Loop:** The experiment spans 30 rounds, starting with an initial labeled pool of 1,000 samples.

B.3 Entropy Estimation for Continuous Tasks

While classification tasks utilize standard Shannon entropy, for regression tasks (e.g., depth estimation) we estimate the *differential entropy*. Assuming the predictive distribution follows a Gaussian $\mathcal{N}(\mu, \sigma^2)$, the entropy is proportional to the log-variance:

$$H(y_{reg}) \approx \frac{1}{2} \ln(2\pi e \hat{\sigma}^2) \propto \ln \hat{\sigma}^2, \tag{12}$$

where $\hat{\sigma}^2$ is the predictive variance estimated via $K = 10$ stochastic forward passes (MC-Dropout).

B.4 Dataset Specifications

We evaluate our proposed framework on three standard multi-task benchmarks, selected to cover a diverse range of domains (general, autonomous driving, indoor) and task affinities.

COCO [22]. A large-scale dataset for general scene understanding. We use the standard split with 118,287 training images and 5,000 validation images.

- **Input:** RGB images (resized to 480×480).
- **Tasks:**
  1. **Image Classification:** For compatibility with multi-task learning, we construct a single-label classification task by assigning each image the most frequent object category (by instance count).
  2. **Object Detection:** Bounding box regression and classification for 80 categories.
  3. **Instance Segmentation:** Pixel-wise binary masks for each object instance.

Cityscapes [9]. A benchmark for urban scene understanding with high-quality pixel-level annotations. The dataset contains 2,975 training images and 500 validation images.

- **Input:** High-resolution street view images (randomly cropped to 512×512 during training).
- **Tasks:**
  1. **Semantic Segmentation:** 19 semantic classes.
  2. **Depth Estimation:** Inverse depth prediction derived from stereo disparity.
  3. **Instance Segmentation:** Instance masks for foreground "thing" classes (e.g., person, car, rider).
- **Note:** Strong geometric correlations between semantic layout and depth cues make this dataset a suitable testbed for evaluating cross-task dependency modeling.

NYUv2 [40]. An indoor scene dataset captured using a Microsoft Kinect. It contains 1,449 aligned RGB-D images (1,159 training, 290 testing).

- **Input:** Indoor RGB images (resized to 480×480).
- **Tasks:**
  1. **Semantic Segmentation:** 13 indoor object and scene classes.
  2. **Depth Estimation:** Metric depth prediction in meters.
  3. **Surface Normal Prediction:** Per-pixel 3D surface normals represented as unit vectors.

Table 6: **Comprehensive Comparison of Related Methodologies.** We categorize all relevant methods by their consideration of key factors. Note that while many methods address Multi-Task Learning (MTL) or Data Efficiency, most do not satisfy the specific requirements of Multi-Task Active Learning (MTL-AL) for vision tasks. "(training)" denotes methods that handle conflict only during gradient updates, not selection. "(pruning)" denotes data pruning on labeled datasets, distinct from active learning.

Method	Uncertainty	Diversity	Gradient Conflict	Inter-Task MI	MTL-Specific
Single-Task Active Learning (Baselines)					
Random Sampling	✗	✗	✗	✗	✗
Entropy Sampling [37]	✓	✗	✗	✗	✗
CoreSet [34]	✗	✓	✗	✗	✗
BADGE [2]	✓	✓	✗	✗	✗
Reliable AL (Influence) [46]	✓	✗	✗	✗	✗
DAL: Reality Check [13]	✓	✓	✗	✗	✗
Data Efficiency and Pruning (Not Active Learning)					
InfoBatch [28]	✗	✗	✓	✗	✗ (pruning)
Multi-Task Optimization / Architecture (Training-time)					
PCGrad [51]	✗	✗	✓	✗	✓ (training)
CAGrad [24]	✗	✗	✓	✗	✓ (training)
Graph Coloring [27]	✗	✗	✓	✗	✓ (training)
Task Grouping [43]	✗	✗	✓	✗	✓ (arch)
Multi-Task Active Learning (Selection Strategy)					
Multi-Task AL [4]	✓	✗	✗	✗	✓
Cost-Effective (Sensors) [1]	✓	✗	✗	✗	✓
Task-Consistency [17]	✓	✓	✗	✓ (output)	✓
ODS (Ours)	✓	✓	✓ **(Signal)**	✓ **(Param)**	✓

C Baselines: Selection and Adaptation

C.1 Justification for Baseline Selection

We select baselines based on direct comparability to the multi-task active learning (MTL-AL) problem formulation. Table 6 categorizes the broader literature; below, we detail why specific recent methods are excluded from the main evaluation.

Optimization and Architecture Methods. Works such as CAGrad [24] and Gradient Interference-Aware Graph Coloring [27] optimize training dynamics, while Task Grouping [43] modifies model architectures. Because active learning evaluates *data selection efficiency*, we fix the optimizer (PCGrad) and architecture across all methods to isolate the selection strategy. Comparing an acquisition function (ODS) directly against a gradient routing algorithm is a category error.

Data Pruning vs. Active Learning. InfoBatch [28] accelerates training via dynamic data pruning. Pruning operates on a *fully labeled* dataset to reduce compute, whereas active learning operates on an *unlabeled* pool to reduce annotation cost. Their objectives are fundamentally incompatible.

Domain Mismatch. Cost-Effective Multitask AL [1] targets wearable sensor systems (time-series data). Its reliance on sensor-specific temporal correlations does not transfer to the dense vision tasks (object detection, segmentation) evaluated here.

Single-Task and Benchmarks. Recent studies such as Xia and Henao [46] and Gashi et al. [13] critically evaluate single-task AL. Following their finding that standard heuristics frequently outperform complex custom architectures, we include robust single-task baselines like *Entropy* and *CoreSet*.

C.2 MT-BADGE: Multi-Task Extension of BADGE

We extend BADGE [2] to the multi-task setting (*MT-BADGE*) by concatenating L_2-normalized per-task gradient embeddings. For a sample x and the current model θ, let $g_t(x) = \nabla_{\theta_{\text{out}}} \mathcal{L}_t(x; \theta)$

denote the gradient of task t's loss with respect to the final-layer parameters. The multi-task gradient embedding is defined as:

$$g_{\text{MT}}(x) = \big[\, \tilde{g}_1(x);\ \tilde{g}_2(x);\ \ldots;\ \tilde{g}_T(x) \,\big], \qquad \tilde{g}_t(x) = g_t(x)/\|g_t(x)\|_2. \tag{13}$$

Applying KMeans++ seeding on $\{g_{\text{MT}}(x)\}_{x\in\mathcal{U}}$ selects the batch of size B. This rigorously generalizes BADGE: it recovers BADGE exactly for $T = 1$ and weights tasks equally in the KMeans++ distance metric for $T > 1$.

C.3 Adaptation of Recent MTL-AL Baselines

Prior MTL-AL methods, namely Hekimoglu et al. [17] and PartAL [10], rely on a PAD-Net architecture with multi-modal distillation (MMD) blocks and assume specific homogeneous task pairings. Because our framework uses a shared-encoder architecture without MMD modules across three heterogeneous tasks (Cls/Det/Seg), we adapt these baselines to be architecture-agnostic.

Hekimoglu et al. [17]. The original method utilizes: (i) *Task inconsistency* $\hat{s}_k(x) = \max_{j\neq k} \mathcal{L}_k(\hat{y}_k, \hat{y}_k^j)$ via MMD-refined predictions; (ii) *Diversity* via auto-encoder compression of task-head features, followed by min-distance to the labeled set; and (iii) a *Selection score* $s(x) = \text{div}(x) \cdot \max_k \hat{s}_k(x)$. We replace the MMD-dependent inconsistency (i) with MC-Dropout ($K = 5$) variance for per-task entropy, while preserving the auto-encoder compression (ii) and multiplicative selection (iii) unchanged.

PartAL [10]. The original method requires PAD-Net for two components: (i) *Partial labeling* of specific $(\text{image}, \text{task})$ pairs using MMD GT-injection; and (ii) *Sample selection* using min-max-normalized per-task MC-Dropout uncertainties. Since our full-annotation setting renders partial labeling structurally inapplicable, we retain the sample selection component by aggregating per-task normalized uncertainties into an image-level score $\sum_t \tilde{u}_t(x)$ and selecting the top-B. This represents a *normalized task-aware uncertainty sampling* baseline, reported as "PartAL (adapted)."

C.4 Adaptation of Multi-Stage Selection Baselines

Table 3 compares five methods originally designed for single-task settings, three of which target one-shot subset selection rather than iterative acquisition. To ensure a rigorous comparison, all five baselines share the same backbone, training schedule, seeds, and budget as ODS, and operate on the current round's shared-encoder features. We detail their multi-task, iterative adaptations below.

FASS [45]. *Filter*: The top-η most uncertain samples, computed via the mean of rank-normalised per-task entropies (replacing single-task least-confidence). *Select*: Lazy-greedy facility location over cosine similarities of shared-encoder features. We use the label-agnostic f_{fac} variant, as the class-conditional f_{NN} variant requires pseudo-labels incompatible with detection or segmentation. The pool multiplier is $\eta = 10\times$ the budget.

Cluster-Margin [8]. *Filter*: The top $10\times$ budget samples with the smallest multi-task margin (rank-normalised per-task best-vs-second-best margin, averaged over tasks). *Select*: Average-linkage agglomerative clustering of encoder features, sorted ascending by size, with round-robin sampling of one member per cluster. While the original freezes a single clustering from a seed model, we continuously re-cluster using the current round's features to account for the shrinking pool, granting the baseline strictly fresher embeddings.

PPAL [49]. *Stage 1 (difficulty-calibrated uncertainty)*: Per-task, per-class difficulty is computed on the labeled set following their semi-supervised variant. Class difficulty dictates a coefficient $w = 1 + \alpha\beta \log(1 + \gamma d)$ with $\gamma = e^{1/\alpha} - 1$ (defaults $\alpha = 0.3$, $\beta = 0.2$). Image uncertainty per task is the difficulty-weighted single-pass entropy. To adapt this single-task metric, we rank-normalise per task and average. The candidate pool is $\delta\times$ the budget ($\delta = 4$). *Stage 2*: Category-conditioned matching similarity over detection box class-probability vectors.

$\mathbb{D}^2$-Pruning [25]. Originally a one-shot pruning method requiring labels and full training dynamics, we adapt it for AL using its supported entropy score. The adaptation requires: (i) a k-NN graph

over current encoder features, rebuilt each round ($k = 15$); (ii) node scores equal to the mean rank-normalised per-task entropy; (iii) one forward message-passing step, $x_i \mathrel{+}= \sum_j e^{-\gamma_f d_{ij}^2} x_j$; and (iv) reverse message-passing selection that greedily takes the argmax and suppresses neighbours via $x_j \mathrel{-}= e^{-\gamma_r d_{kj}^2} x_k$. Distances use L_2-normalised features and default γ scales.

SES [48]. Originally one-shot subset selection via a pretrained-feature k-NN graph and structural entropy. Per AL round, we build a k-NN graph over current encoder features ($k = 15$) and construct the two-level encoding tree using k-means communities, matching their label-free protocol. Structural entropy is $S_e(u) = \mathrm{vol}(V)^{-1} \sum_{(u,v)\in E} w_{uv} \log \mathrm{vol}(\mathrm{LCA}(u, v))$, marking community boundaries. The difficulty term S_t uses the mean rank-normalised per-task entropy of the current model, substituting their label-dependent score. Selection follows their blue-noise sampling.

D Empirical Validation of Theoretical Claims

D.1 Empirical Validation of Assumption 1

We compute per-sample Informativeness $\mathcal{I}(x) = \alpha\,\mathrm{Hardness}(x) + \gamma\,\mathrm{CTMI}(x)$ and Exploration $\mathcal{E}(x)$ (min-distance-to-labeled-set in shared-encoder feature space) on 1,500 unlabeled COCO images using saved checkpoints from a 50-round ODS run (seed 42). Table 7 summarizes the results.

Table 7: **Per-round correlation of $\mathcal{I}$ and $\mathcal{E}$ on COCO (1,500 unlabeled images).** The persistent negative sign in early/mid rounds is exactly why linear scalarization fails: the two scores trade off, so a weighted sum slides into the diversity-dilution regime. The two-stage decomposition avoids this by structurally restricting Stage 2 to high-$\mathcal{I}$ candidates.

Round	1	5	10	15	20	25	30	35	40	45	49
Pearson r	-0.179	-0.106	-0.180	-0.121	-0.087	-0.079	-0.124	-0.088	-0.020	-0.034	-0.089
Spearman ρ	-0.148	-0.069	-0.174	-0.111	-0.072	-0.060	-0.089	-0.070	$+0.001$	-0.018	-0.065

D.2 Empirical Normality of Per-Task Entropies

Because CTMI is a second-moment statistic, it is important to verify how much dependence structure it successfully captures. We provide two lines of evidence: the empirical shape of the per-task entropy distributions (Table 8), and a direct comparison against a non-linear dependence measure (Appendix D.3).

Table 8: **Shapiro-Wilk W statistic, skewness, and excess kurtosis** of z-score normalized per-task entropies on COCO. W closer to 1.0 indicates a closer fit to a Gaussian distribution. Segmentation entropies converge to nearly perfect Gaussianity; classification develops a slight right tail; detection exhibits heavier tails but maintains $W > 0.90$ throughout.

Round	Task	W	Skew	Kurtosis
1	Cls	0.979	-0.02	-0.88
1	Det	0.898	-1.72	$+6.56$
1	Seg	0.963	-0.68	0.00
10	Cls	0.979	$+0.86$	-0.10
10	Det	0.908	-1.65	$+6.74$
10	Seg	0.990	-0.20	-0.53
20	Cls	0.920	$+1.17$	$+0.51$
20	Det	0.931	-1.32	$+4.30$
20	Seg	0.993	-0.04	-0.54
29	Cls	0.916	$+1.28$	$+1.15$
29	Det	0.932	-1.33	$+4.83$
29	Seg	**0.994**	-0.08	-0.51

As Table 8 shows, segmentation and classification entropies remain close to Gaussian ($W > 0.91$ throughout, with segmentation reaching $W = 0.994$), while detection has heavier tails (kurtosis $\approx +5$). While our method makes no distributional assumptions, this confirms the marginals are unimodal

and largely close to Gaussian—the exact regime where a second-moment statistic is highly informative. Appendix D.3 validates this directly.

D.3 Pearson vs. Spearman vs. Distance Correlation of Task Entropies

A standard correlation coefficient captures only linear (or monotone) dependence, remaining invariant to relationships like $Y = X^2$. To measure the extent of any omitted structure, Table 9 compares Pearson, Spearman, and distance correlation [44] on COCO ($N = 2{,}000$). Crucially, distance correlation equals zero if and only if the variables are strictly independent, thereby capturing all forms of dependence.

Table 9: **Pearson vs. Spearman vs. distance correlation of per-task entropies (COCO).** $|\text{Pearson}|$ and dCorr agree within ≤ 0.03 at every round, demonstrating that task-uncertainty dependencies in MTL are predominantly linear. Consequently, CTMI loses negligible signal by relying on covariance.

Rd	Pair	Pearson	Spearman	dCorr	$\lvert\lvert r\rvert - \text{dCorr}\rvert$
1	Cls-Det	$+0.515$	$+0.517$	0.488	0.027
1	Cls-Seg	$+0.264$	$+0.264$	0.245	0.019
1	Det-Seg	$+0.448$	$+0.447$	0.422	0.026
10	Cls-Det	$+0.338$	$+0.367$	0.330	0.008
10	Det-Seg	$+0.205$	$+0.200$	0.188	0.017
29	Cls-Det	$+0.412$	$+0.483$	0.406	0.006
29	Cls-Seg	$+0.249$	$+0.290$	0.245	0.004
29	Det-Seg	$+0.170$	$+0.177$	0.166	0.004
49	Cls-Det	$+0.424$	$+0.490$	0.411	0.013
49	Cls-Seg	$+0.209$	$+0.259$	0.214	0.005
49	Det-Seg	$+0.190$	$+0.216$	0.200	0.010

E Extended Experimental Results

E.1 Detailed Cross-Domain Performance

Table 10 provides the detailed per-task metrics corresponding to the cross-domain summary in the main text. ODS ranks first on every individual task across COCO, Cityscapes, and NYUv2.

Table 10: **Detailed Cross-Domain Performance (Round 30).** (↑: higher is better, ↓: lower is better.)

Dataset	Method	Task 1	Task 2	Task 3	Best MTL Score
COCO (Tasks: Cls / Det / Seg)					
		Cls (Acc) ↑	Det (mAP) ↑	Seg (mIoU) ↑	Comp. (Δ%)
COCO	Random	38.4	25.2	15.1	0.0%
	BatchBALD	39.4	25.4	16.4	+4.0%
	MT-BADGE	41.2	25.8	16.8	+7.0%
	ODS	**42.1**	**26.5**	**17.4**	**+10.0%**
Cityscapes (Tasks: Seg / Depth / Instance)					
		Seg (mIoU) ↑	Depth (RMSE) ↓	Ins (mAP) ↑	Comp. (Δ%)
Cityscapes	Random	65.0	0.4150	28.0	0.0%
	BatchBALD	68.2	0.4061	29.8	+4.5%
	MT-BADGE	69.1	0.4041	30.7	+6.2%
	ODS	**70.5**	**0.3916**	**31.4**	**+8.8%**
NYUv2 (Tasks: Seg / Depth / Normal)					
		Seg (mIoU) ↑	Depth (RMSE) ↓	Norm (mErr) ↓	Comp. (Δ%)
NYUv2	Random	31.0	0.750	27.0	0.0%
	BatchBALD	32.5	0.726	26.1	+3.8%
	MT-BADGE	32.7	0.718	25.5	+5.1%
	ODS	**33.5**	**0.700**	**24.9**	**+7.5%**

E.2 Pretrained ResNet-50 Backbone (Multi-Seed)

To isolate the acquisition strategy from large-scale pretraining effects, our main results rely on lightweight scratch encoders. Table 11 evaluates an ImageNet-pretrained ResNet-50 backbone under the identical COCO AL protocol (30 rounds, $B = 32$, initial $= 100$, pool $= 5{,}000$).

Table 11: **COCO with ImageNet-pretrained ResNet-50 (3 seeds).** ODS retains rank 1 across individual tasks; gaps narrow as expected when using a strong pretrained backbone.

Method	Cls Acc ↑	Det Acc ↑	Seg Dice ↑	Seg mIoU ↑
Random	0.505 ±.011	0.372 ±.011	0.316 ±.021	0.274 ±.029
CoreSet	0.508 ±.006	0.376 ±.012	0.315 ±.019	0.272 ±.027
Uncertainty	0.511 ±.006	0.373 ±.013	0.316 ±.028	0.276 ±.037
Diversity	0.504 ±.010	0.372 ±.003	0.313 ±.014	0.273 ±.023
ODS (Ours)	**0.521** ±.010	**0.377** ±.004	**0.326** ±.022	**0.289** ±.029

Discussion. Strong pretrained backbones naturally narrow performance gaps in active learning, as the initial representation already captures significant data variability [42]. Nevertheless, ODS retains rank 1 across all individual tasks over three seeds, demonstrating that conflict-aware acquisition extracts structural value even from a well-initialized encoder.

E.3 Runtime Comparison

Table 12 reports the per-round acquisition runtime on COCO (rounds 1–29, scratch training, single V100). While MC-Dropout (ODS) and KMeans++ (BADGE) dominate the acquisition cost, ODS's computational overhead is strictly offset by its superior label efficiency.

Table 12: **Per-round runtime on COCO (seconds).** "Acquisition" isolates the cost of the AL selection step (excluding training).

Method	Mean	Median
Random	225	204
Uncertainty	226	197
CoreSet	205	165
BADGE	655	636
ODS	**995**	**1114**

F Ablation and Sensitivity Studies

F.1 Decoupling Selection from Optimization (PCGrad Ablation)

Table 13 isolates the contribution of the acquisition strategy from that of the gradient surgery optimizer. Even under standard AdamW, ODS achieves $+8.2\%$, surpassing all baselines; PCGrad provides an additional $+2.3\%$ improvement, confirming that ODS and PCGrad address orthogonal bottlenecks.

F.2 Component-wise Ablation: Disentangling Score and Combination

Table 14 ablates the three core components of ODS: (a) Hardness, (b) CTMI in Stage 1, and (c) K-Center coverage in Stage 2. The full ODS dominates every leave-one-out variant. Removing K-Center Stage 2 ("no KC") causes the most severe performance drop, confirming that structural decoupling prevents diversity dilution and drives the primary acquisition gain.

F.3 Is the Combination Rule Doing the Work?

Equation (5) combines Hardness and CTMI using a weighted sum after min–max normalisation. To verify whether the reported gains stem from this specific heuristic, Table 15 compares four alternative combination rules while holding the component scores, Stage 2, and the training protocol

Table 13: **PCGrad ablation on COCO.** ODS significantly outperforms all baselines regardless of the underlying optimizer, while PCGrad provides consistent supplementary gains.

Strategy	Optimizer	Best Score ($\Delta m\%$)
Scenario A: Standard Optimization		
Random	AdamW	—
BatchBALD	AdamW	0.0311 (+3.1%)
MT-BADGE	AdamW	0.0516 (+5.2%)
ODS (Ours)	AdamW	**0.0816 (+8.2%)**
Scenario B: Gradient Surgery		
Random	PCGrad	—
BatchBALD	PCGrad	0.0500 (+5.0%)
MT-BADGE	PCGrad	0.0750 (+7.5%)
ODS (Ours)	PCGrad	**0.1050 (+10.5%)**

Table 14: **Component-wise ablation.** The full ODS dominates each leave-one-out variant. Removing K-Center Stage 2 ("no KC") causes the largest single-component drop, confirming that the structural decoupling is the dominant source of gain.

Configuration	Hard.	CTMI	K-C	MTL Comp.
H only (no KC)	✓	✗	✗	0.087
CTMI only (no KC)	✗	✓	✗	0.064
H+CTMI (no KC)	✓	✓	✗	0.093
H only + KC	✓	✗	✓	0.094
CTMI only + KC	✗	✓	✓	0.082
Full ODS	✓	✓	✓	**0.105**

constant over three seeds: the default weighted sum, a rank-sum, a z-score sum, and a direct product ($\text{Hardness} \cdot \text{CTMI}$).

Table 15: **Combination-rule ablation (COCO, 3 seeds: $\{42, 123, 456\}$).** Final MTL Composite (mean$\pm$std). All four rules perform within 0.008 of each other—well within seed-to-seed variation—indicating no statistically distinguishable difference.

Rule	Form	Final MTL Comp.
Weighted sum (default)	$\alpha H + \gamma C$	0.723 ± 0.016
Rank-sum	$\text{rank}(H) + \text{rank}(C)$	0.731 ± 0.001
z-score sum	$z(H) + z(C)$	0.731 ± 0.001
Product	$H \cdot C$	0.726 ± 0.008

The maximum spread across the four rules (0.723–0.731) is strictly smaller than the seed variation of the default rule alone (± 0.016). Coupled with the component-wise ablation (Appendix F.2)—where removing CTMI entirely drops performance from 0.105 to 0.094 under an identical Stage 2—this confirms that the performance gain originates from the cross-task signal itself, not the specific arithmetic used to merge it. We therefore present the weighted sum merely as a convenient default, not a methodological necessity.

F.4 Linear Scalarization vs. Orthogonal Decomposition

Table 16 compares the ODS two-stage decomposition against a standard linear scalarization using identical per-sample scores (Hardness, CTMI, K-Center distance). ODS outperforms linear scalarization by $+6.5\%$ on the MTL Composite. The linear scheme suffers from *diversity dilution*: highly diverse but uninformative outliers dominate the selection, degrading dense-prediction tasks severely.

Table 16: **Linear scalarization vs. ODS, identical scores.**

Combination strategy	Cls Acc	Det Acc	Seg Dice	MTL Comp.
Linear: $(1-\delta)(\alpha H+\gamma \mathrm{CTMI})+\delta\,\mathrm{Div}$	38.8	24.00	11.24	0.04
ODS (two-stage)	**42.14**	**26.52**	**17.44**	**0.105**
Δ	$+3.34$	$+2.52$	$+\mathbf{6.2}$	$+0.065$

F.5 Hyperparameter Sensitivity (Pool Factor m and MC Samples K)

Pool Factor (m): Table 17 demonstrates that ODS is robust across pool factors. Moderate values ($m = 5$) achieve the optimal trade-off between filtering uninformative outliers and preserving sufficient manifold diversity, validating the Orthogonal Decomposition design.

MC-Dropout Samples (K): Table 18 shows that $K = 10$ provides a stable cross-task covariance estimate. Larger K values yield only marginal gains while significantly increasing computational overhead, confirming CTMI's computational efficiency.

Table 17: **Sensitivity to Pool Factor m.**

m	2	3	5	10
Score	0.1032	0.1045	**0.1051**	0.1048

Table 18: **Sensitivity to MC Samples K.**

K	5	10	20	50
Score	0.1041	**0.1049**	0.1044	0.1046
Time	45	82	158	390

F.6 Sensitivity to (α, γ) in the Stage-1 Score

The Stage 1 informativeness score is $\mathcal{I}(x) = \alpha\,\mathrm{Hardness}(x) + \gamma\,\mathrm{CTMI}(x)$. As shown in Table 19, ODS is robust to a wide range of configurations. The performance surface is essentially flat near $(\alpha, \gamma) = (1, 1)$, with no setting within a $4\times$ multiplier range losing more than 1.5% relative MTL Composite. Severe performance drops occur only under degenerate configurations where either component is entirely removed ($\alpha = 0$ or $\gamma = 0$).

Table 19: **Sensitivity of ODS to (α, γ) on COCO.** Equal weighting ($\alpha = \gamma = 1$) is near-optimal and robust to a $4\times$ swing in either weight.

(α, γ)	$(0,1)$	$(0.5,1)$	$(1,0)$	$(1,0.5)$	$(1,1)$	$(1,2)$	$(2,1)$
MTL Comp.	0.082	0.099	0.094	0.102	**0.105**	0.103	0.104

G Qualitative Analysis

G.1 What the Conflict Ablation Shows

The controlled ablation in Table 4 isolates the effect of the acquisition score: holding the optimizer (AdamW) and scoring pipeline fixed, penalizing gradient conflict lowers the MTL Composite from 0.076 (ignoring conflict) to 0.059 (penalizing it). This empirically grounds the *Conflict Paradox*: penalizing conflict actively degrades data selection.

Crucially, this defines what the acquisition score must *omit*, not what it must seek. ODS ranks candidates exclusively by Hardness and CTMI (Eq. (5)); gradient conflict is absent from the score. This does not imply that ODS batches are inherently enriched with high-conflict samples relative to random selection. To understand the selection behavior, readers should refer to the component and combination-rule ablations (Appendices F.2 and F.3), which isolate the contribution of each term without inferring a conflict-driven mechanism.

G.2 Qualitative Analysis: Understanding Diversity Dilution

To visualize how ODS prevents *Diversity Dilution*, we inspect the selected samples across active learning rounds along two axes: (1) **distributional coverage** via t-SNE embeddings (Figure 5),

and (2) **visual and semantic content** (Figures 6–9). In all qualitative figures, **red boxes** highlight redundant selections and characteristic failure modes of the baselines.

G.2.1 Distributional Analysis: The Outlier Trap

Figure 5 tracks the t-SNE embeddings of selected samples against the unlabeled pool across Rounds 1–8. A stark behavioral divergence emerges:

- **Redundant Collapse by Baselines.** Diversity-driven baselines (*CoreSet*, *BADGE*, *BatchBALD*) repeatedly cluster in sparse, isolated embedding regions (e.g., $y < -40$ in Rounds 2, 3, and 5; red dashed circles). This illustrates the **inefficient over-sampling of visual outliers** that are distinctive but uninformative.
- **Controlled Exploration by ODS.** ODS also probes peripheral regions for coverage, but does so sparsely. By allocating the vast majority of its budget to the **high-density informative manifold**, ODS strictly balances outlier exploration with the exploitation of semantically rich data.

G.2.2 Visual Redundancy vs. Semantic Diversity

Figures 6 and 7 map these distributional failures to image space. **ODS** consistently acquires visually diverse, semantically dense scenes—from cluttered indoor environments to complex street views. Conversely, baselines exhibit severe redundancy, selecting visually similar samples despite operating on different acquisition principles. Geometry-based (*CoreSet*) and information-theoretic (*BatchBALD*) methods often converge on identical outliers, a redundancy mirrored in the t-SNE clusters (Figure 5, red dashed circles).

Failure Mode I: Semantic Sparsity ("Toilet Trap"). In Round 2 (Figure 6), several baselines fixate on semantically sparse environments. Highlighted by **red boxes**, *CoreSet*, *BADGE*, and *BatchBALD* repeatedly acquire images of empty bathrooms, while *Entropy* favors featureless kitchen counters. These samples are geometrically rare but contain minimal object interactions. This is *Diversity Dilution* in action: prioritizing distance metrics collapses the selection onto samples with low semantic content.

Failure Mode II: Texture and Structural Bias. Baselines also over-sample high-frequency textures and repetitive structures. In Round 3 (Figure 7), *CoreSet* disproportionately selects donuts, bananas, and pastries, while other baselines redundantly acquire images of fences or grids. While visually salient, these scenes offer negligible multi-task supervision compared to scenes with complex object interactions.

G.2.3 Annotation Density and Multi-Task Supervision

Figures 8 and 9 overlay ground-truth annotations, revealing stark density differences. ODS consistently targets scenes with dense bounding boxes and intricate segmentation masks (e.g., cluttered desks, crowded streets). These multi-instance scenes yield rich, high-frequency gradient signals that simultaneously improve detection and segmentation. In contrast, baselines frequently select sparse or empty images (e.g., ocean waves). These samples are "hard" only due to uninformative background uncertainty, yielding minimal information gain per labeled pixel.

G.2.4 Why ODS Avoids Diversity Dilution

Across both feature-space and image-space analyses, ODS systematically bypasses the failure modes that trap standard baselines. By filtering out uninformative outliers before maximizing coverage, ODS consistently acquires semantically complex, non-redundant data—discovering the high-value samples that existing MTL-AL strategies overlook.

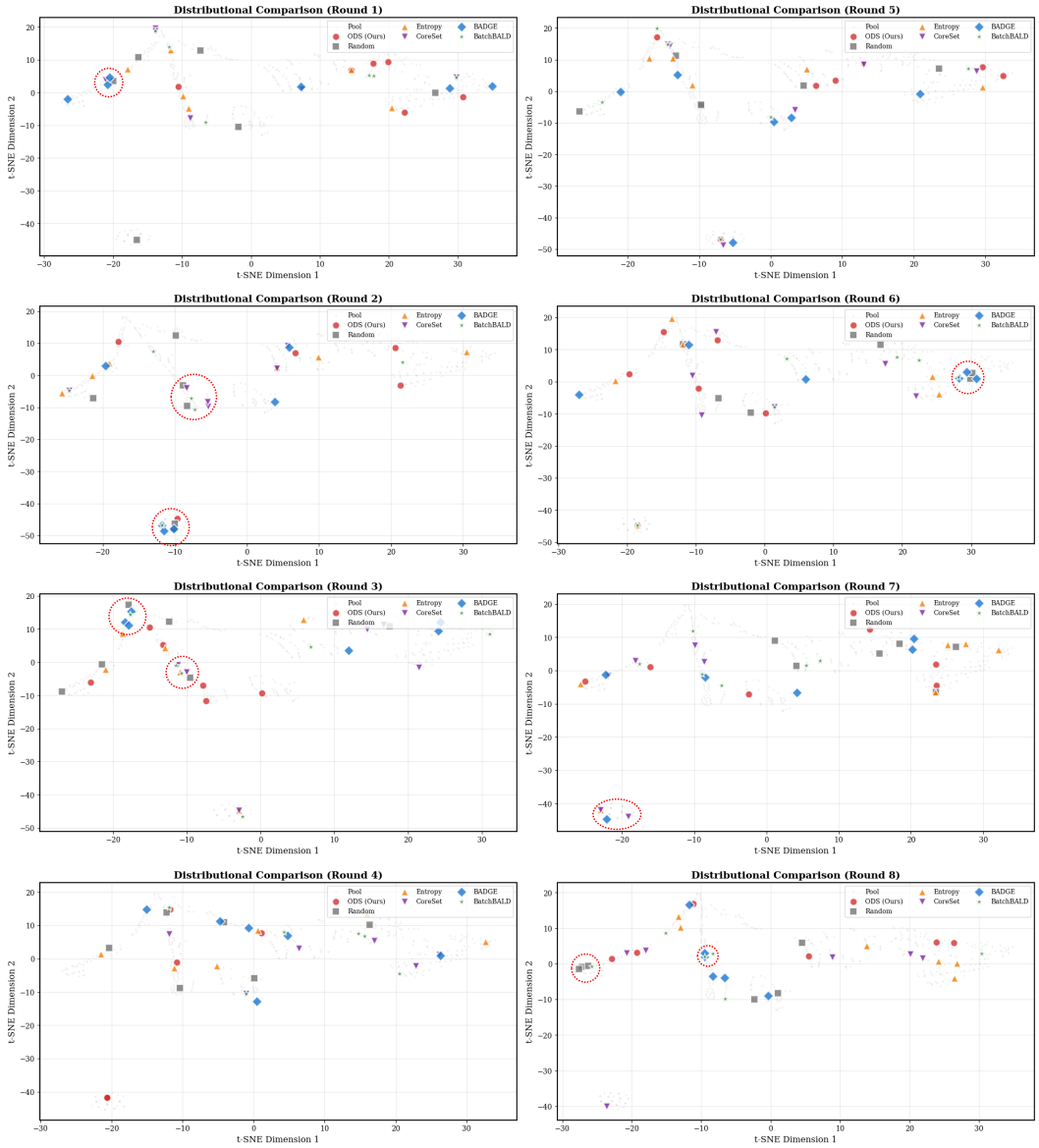


Figure 5: **t-SNE Distributional Analysis across Active Learning Rounds (1–8).** We visualize the embedding space of selected samples from different acquisition strategies, overlaid on the unlabeled pool (gray). **Red dashed circles highlight regions of redundant selection by baselines. (Baseline Inefficiency):** Diversity-driven methods such as *CoreSet* and *BADGE* repeatedly select samples concentrated within narrow, low-density regions, leading to **redundant over-sampling of visual outliers** (e.g., isolated bottom clusters in R2, R3, R5, and R7). **(Controlled Exploration by ODS):** Although **ODS** (•) also explores these peripheral regions to ensure coverage, it does so **sparsely**, avoiding the clustering behavior observed in baselines. **(Balanced Allocation):** As a result, ODS allocates most of its acquisition budget to the **high-density informative manifold**, achieving a more effective balance between exploration of outliers and exploitation of the main data distribution.

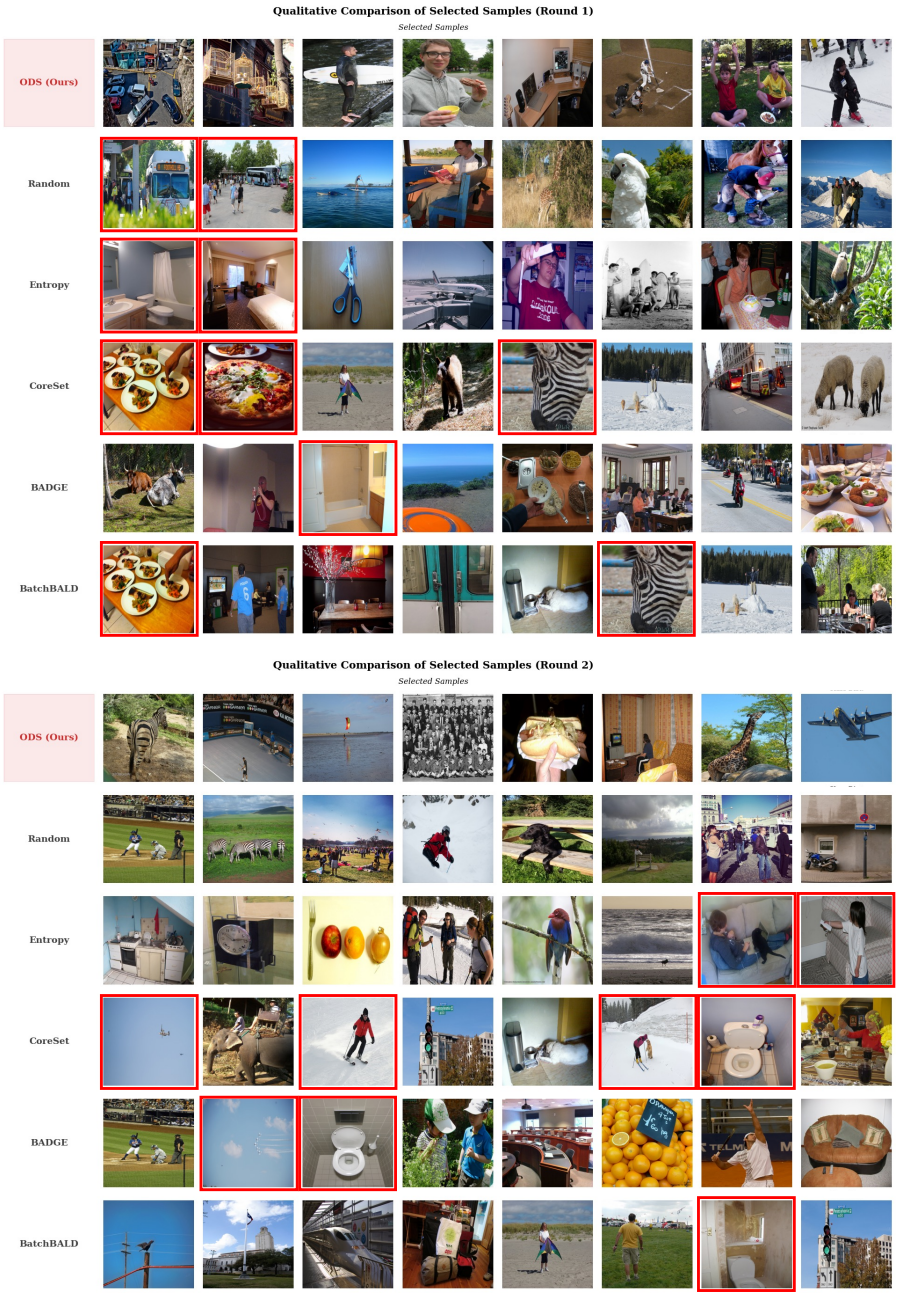


Figure 6: **Qualitative Comparison of Selected Samples (COCO, Rounds 1–2).** Each row shows images selected by a different acquisition strategy. **Red boxes mark redundant selections by baselines,** where visually or semantically similar samples are repeatedly acquired within the same round. In contrast, **ODS avoids such early-stage redundancy**, allocating its budget to more diverse and complementary visual content.



Figure 7: **Qualitative Comparison of Selected Samples (COCO, Rounds 3–4). Red boxes highlight structural and textural redundancies** among samples selected by baseline methods, such as repeated acquisition of near-identical scenes or objects. Despite operating at later rounds with a more informative pool, baselines continue to exhibit inefficient budget usage, whereas **ODS maintains diverse coverage without collapsing onto narrow visual patterns**.



Figure 8: **Annotation Density Analysis (COCO, Rounds 1 & 2).** We visualize the ground-truth bounding boxes and segmentation masks for selected samples. **ODS (Top Row):** consistently selects images with **high instance density** (e.g., the cluttered desk in R1, the crowd in R2), providing rich supervision for multi-task learning. **Baselines (Bottom Rows):** frequent failure modes include selecting "empty" images with **sparse or no annotations** (e.g., *Entropy* selecting less informative rooms in R2, *Random* selecting simple single-object images). This confirms that ODS maximizes information gain per pixel.

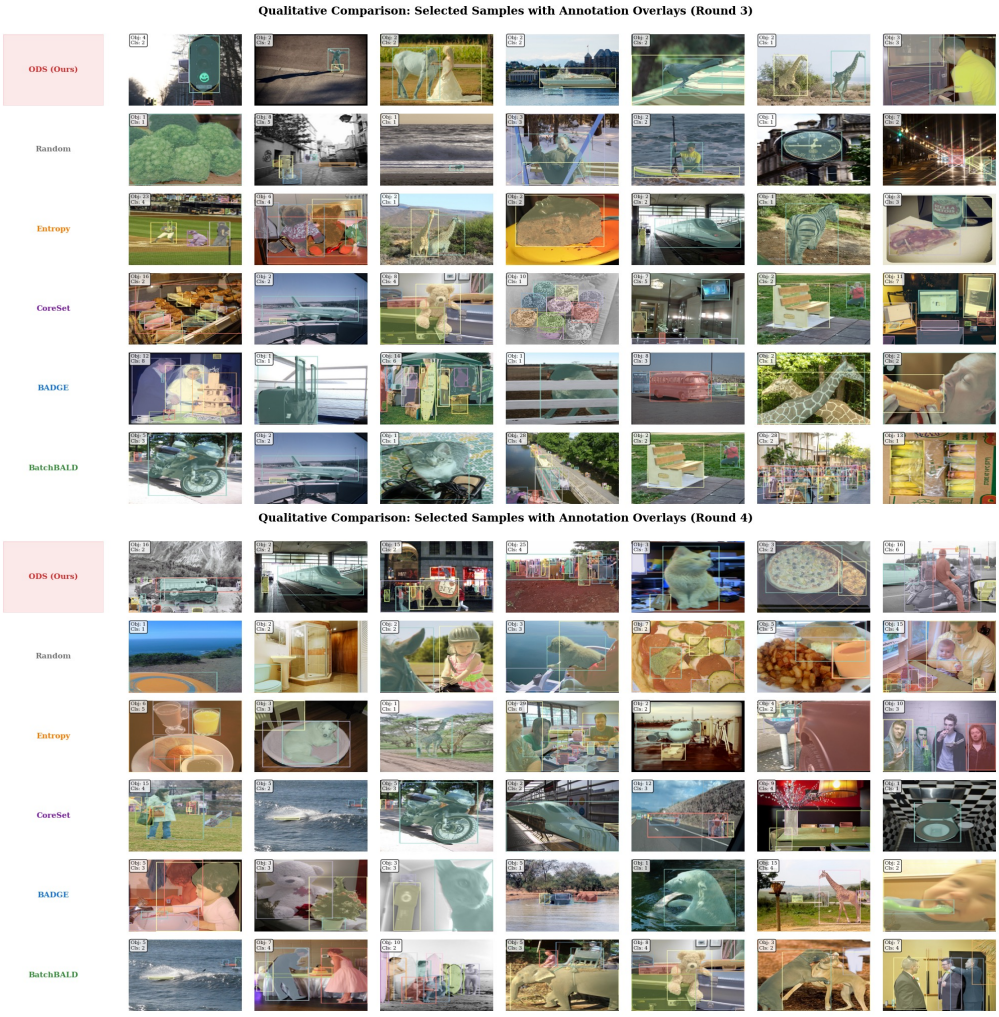


Figure 9: **Annotation Density Analysis (COCO, Rounds 3 & 4).** Overlays reveal the semantic complexity of selected samples. **ODS (Top Rows):** selects structurally complex scenes with overlapping objects, forcing the model to resolve difficult occlusions. **Baselines:** *CoreSet* and *BatchBALD* often select images with large, simple textures or single large objects, which offer limited geometric challenges compared to the dense scenes selected by ODS.

H Ethics statement

This work investigates acquisition strategies for multi-task learning and collects no new data. All experiments rely on the public COCO, Cityscapes, and NYUv2 benchmarks under their respective licenses. No human-subject research was conducted, and we handle no personally identifiable information beyond the anonymized distributions provided by the dataset creators (e.g., blurred faces and license plates in Cityscapes). Like any active learning method, our acquisition function concentrates the annotation budget on a specific data subset, which may amplify existing pool imbalances. Because our score relies strictly on predictive uncertainty and cross-task dependency without explicit fairness constraints, deployment in safety-critical settings requires auditing the selected batches. We anticipate no risks of misuse beyond those inherent to general label-efficient learning.

I Reproducibility statement

All code, configuration files, and analysis scripts required to reproduce our results are publicly available at `github.com/Juho-Jung/multitask_active_learning_official`. Appendix B.2 specifies the architectures, optimization settings, and MC-dropout configurations. Table 5 lists the exact (m, K, α, γ) hyperparameters for all experimental protocols. Appendix B.4 documents the dataset splits and seed sets, and Appendix B.1 details the complete selection procedure. Unless noted otherwise, results are reported across three random seeds $(42, 123, 456)$ as mean $\pm$ standard deviation. While non-deterministic GPU operations may cause minor per-round numerical variances, the performance gaps and rank orderings central to our claims are strictly robust and fully reproducible.

J Use of AI

Large language models served exclusively as writing assistants for grammar, LaTeX formatting, and literature searches. They contributed no original research ideas, methodological designs, experimental executions, or data analyses. All technical claims, mathematical proofs, and empirical results were produced directly by the human authors, with LLMs limited to secondary proofreading. The authors take full responsibility for the content of this paper. We provide this declaration for transparency, though it falls under the permitted-use category of the NeurIPS LLM policy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) state our three contributions — the Conflict Paradox, Orthogonal Decomposition, and CTMI — which match the formal results (Proposition 1, Assumption 1, Proposition 2, Algorithm 1) and the empirical evidence (Tables 2, 4, 16, 10, 13). Empirical scope (COCO, Cityscapes, NYUv2; lightweight scratch encoders by default) is stated up front.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a dedicated Limitations section (Section 7) covering: (L1) the sample-level / objective-level distinction in Proposition 1; (L2) covariance-based CTMI being blind to non-linear dependence (mitigated by maximum-entropy lower-bound semantics and validated by distance correlation in Appendix D.3); (L3) the (m, K, α, γ) hyperparameters (with sensitivity studies); (L4) the lightweight scratch backbone scope (with a pretrained ResNet-50 study in Appendix E.2).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results state their assumptions explicitly and reference complete proofs. Proposition 1 (sample-level conflict structure) lists three assumptions and the full proof appears in Appendix A.1. Proposition 2 (shared-representation attribution for CTMI) is proved in Appendix A.2; the submitted version's maximum-entropy lower-bound claim was withdrawn during discussion and the appendix states explicitly what replaced it and why. The two-stage approximation guarantee (Proposition in Appendix A.3) states the empirical-independence assumption and its 2-approximation component (Sener & Savarese). Proof sketches in the main text are paired with formal proofs in the appendix; we are explicit about heuristic vs. formal claims (see Remark on sample-level vs. objective-level Pareto stationarity).

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix B.2 reports the model architectures, the active-learning protocol (initial labeled, budget per round, number of rounds), the optimizer (PCGrad with AdamW, learning rate 10^{-3}, cosine schedule), batch size, regularization, and dropout rate. ODS-specific hyperparameters (m, K, α, γ) are listed per protocol in Table 5, and Algorithm 1 is fully specified. Datasets are public and described in Appendix B.4. The code repository is linked in the abstract.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the full implementation at the anonymous repository linked in the abstract, including training/AL loops, all baselines, and configuration files for COCO, Cityscapes, and NYUv2. Datasets are public; access procedures for COCO [22], Cityscapes [9], and NYUv2 [40] follow the standard splits and licensing terms documented in Appendix B.4.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 provides the high-level setup (datasets, baselines, fairness controls, the MTL Composite Score and Average Rank metrics). Appendix B.2 adds full architectures, training hyperparameters (LR, batch size, optimizer, scheduler), the AL protocol (initial labeled $=$ 1000, 30 rounds, batch $=$ 32), MC-Dropout settings ($K = 20$, $p = 0.05$),

and how (α, γ) are chosen with adaptive scheduling. Sensitivity studies are reported in Appendices F.5 and F.6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Main Table 2 reports mean$\pm$std over 3 seeds $\{42, 123, 456\}$ for all methods on COCO under identical protocols. The pretrained-backbone study (Appendix E.2) is also reported with 3-seed mean$\pm$std. We discuss the statistical-significance limitations of 3 seeds in Section 7 and rely on cross-task / cross-dataset rank consistency as complementary evidence.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix B.2 states that experiments were conducted on $8\times$ NVIDIA V100 (32GB) GPUs. Appendix E.3 reports per-round wall-clock acquisition time (mean and median in seconds) for ODS and all baselines, isolating selection cost from training cost.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses only publicly released, standard computer-vision benchmarks (COCO, Cityscapes, NYUv2), proposes an active-learning algorithm, and involves no human subjects, sensitive personal data, or deployment scenarios. We have reviewed the NeurIPS Code of Ethics and the work conforms in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The proposed method reduces the labeled-data requirement for multi-task vision systems, with a positive impact on settings where annotation is expensive (e.g., medical, autonomous driving prototyping). We do not foresee a direct negative societal impact specific to this work; the underlying tasks (classification, detection, segmentation, depth estimation) are general-purpose and not tied to a high-risk deployment. Like any data-efficient learning method, ODS could in principle reduce the cost of training models for downstream applications with negative impact, but the algorithmic contribution is upstream of any specific deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a generative model, scraped dataset, or any artifact with a high risk of misuse. We release only an active-learning algorithm and lightweight task models trained on standard public datasets (COCO, Cityscapes, NYUv2).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets used in the paper are publicly available and are cited at first use: COCO [22] (CC BY 4.0 / Flickr terms for images), Cityscapes [9] (research-only license), NYUv2 [40] (research-only license). Baseline implementations follow the publicly released reference implementations of BADGE [2], BatchBALD [18], CoreSet [34], PCGrad [51], Hekimoglu et al. [17], and PartAL [10], all under their original licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the source code for ODS, released at the anonymous repository linked from the abstract. The repository includes a README, configuration files, training/AL loops, and scripts to reproduce the headline results on each dataset. No new datasets or pre-trained models are released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments are conducted on existing publicly released image datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so no IRB approval is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the core methodology, scientific results, or experimental design. They may have been used for incidental writing/editing assistance, which per NeurIPS policy does not require declaration.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.