

LABELING COPILOT: A Deep Research Agent for Automated Data Curation in Computer Vision

Debargha Ganguly^{*1,3}, Sumit Kumar^{*1,4}, Ishwar Balappanawar^{*1,4}, Weicong Chen^{*3}, Shashank Kambhatla^{1,5}
Srinivasan Iyengar², Shivkumar Kalyanaraman², Ponnurangam Kumaraguru⁴, Vipin Chaudhary³

¹Microsoft Research ²Microsoft Corporation ³Case Western Reserve University

⁴IIIT Hyderabad ⁵University of Pennsylvania

{debargha, weicong, vipin}@case.edu, {sriyengar, shkalya}@microsoft.com, sumit.k@research.iiit.ac.in,
ishwar.balappanawar@students.iiit.ac.in, pk.guru@iiit.ac.in, skamb@seas.upenn.edu

Abstract—Curating high-quality, domain-specific datasets is a major bottleneck for deploying robust vision systems, requiring complex trade-offs between data quality, diversity, and cost when researching vast, unlabeled data lakes. We introduce Labeling Copilot, the first data curation deep research agent for computer vision. A central orchestrator agent, powered by a large multimodal language model, uses multi-step reasoning to execute specialized tools across three core capabilities: (1) Calibrated Discovery sources relevant, in-distribution data from large repositories; (2) Controllable Synthesis generates novel data for rare scenarios with robust filtering; and (3) Consensus Annotation produces accurate labels by orchestrating multiple foundation models via a novel consensus mechanism incorporating non-maximum suppression and voting. Our large-scale validation proves the effectiveness of Labeling Copilot’s components. The Consensus Annotation module excels at object discovery: on the dense COCO dataset, it averages 14.2 candidate proposals per image—nearly double the 7.4 ground-truth objects—achieving a final annotation mAP of 37.1%. On the web-scale Open Images dataset, it navigated extreme class imbalance to discover 903 new bounding box categories, expanding its capability to over 1500 total. Concurrently, our Calibrated Discovery tool, tested at a 10-million sample scale, features an active learning strategy that is up to 40x more computationally efficient than alternatives with equivalent sample efficiency. These experiments validate that an agentic workflow with optimized, scalable tools provides a robust foundation for curating industrial-scale datasets.

I. INTRODUCTION

The remarkable progress in Computer Vision (CV) has been fundamentally enabled by large-scale, high-quality, domain-specific datasets such as ImageNet [1], COCO [2], and Open Images [3]. The curation of these datasets, however, remains a persistent bottleneck that limits the scalable deployment and real-world impact of CV models [4], [5]. Automating this complex workflow using foundation models [6] and agentic systems have become an active area of research [7], [8]. While popular models like GroundingDINO [9] and SAM [10] excel on benchmarks, they are yet to scale to the knowledge-intensive task of curating production-grade datasets. This

This work was supported in part by the NSF research grant #2320952, #2117439, #2112606, and #2117439.

^{*}These authors contributed equally to this work.

intricate process is not a linear pipeline but a dynamic challenge requiring intelligent decision-making, making it an ideal application for an agentic framework [11].

This data curation challenge presents several intertwined difficulties that require sophisticated reasoning and tool coordination. Firstly, data sources are often massive, heterogeneous data lakes like *LAION* [12] or *DataComp* [13], containing millions of irrelevant or low-quality images that must be intelligently filtered [14]. Secondly, real-world applications demand robustness to rare events (e.g., adverse weather, unusual object poses) [15], [16], which are inherently absent or underrepresented in training sets—a problem especially acute in data-scarce domains like medical imaging [17] or industrial inspection [18]. Finally, the annotation process itself is ambiguous; manual labeling is slow and expensive [19], while automated methods using different foundation models often produce conflicting labels [20], introducing significant noise if naively combined [21].

Overcoming these complexities, a daunting task even for human teams, requires a new type of agent – one that can research the data landscape, synthesize information, and reason about trade-offs. Inspired by these challenges, we propose **the first deep research agent for vision data curation, called LABELING COPILOT**. As shown in Figure 5, our agent operates in a continuous loop: (1) Discovery: Starting with a high-level query, the agent uses a calibrated retrieval tool to find relevant data using techniques from active learning [22] and out-of-distribution detection [23]. (2) Synthesis: If the dataset lacks diversity, it uses a controllable synthesis tool leveraging instruction-following diffusion models [24], [25] to generate images covering specific edge cases. (3) Annotation & Filtering: The agent orchestrates a suite of foundation models to generate candidate labels, builds a consensus for a single high-quality annotation using techniques inspired by ensemble learning [26], and filters out low-confidence samples. This curated data is then used to fine-tune a target model, allowing the agent to repeat the cycle and address model weaknesses iteratively. Overall, our primary contributions are as follows:

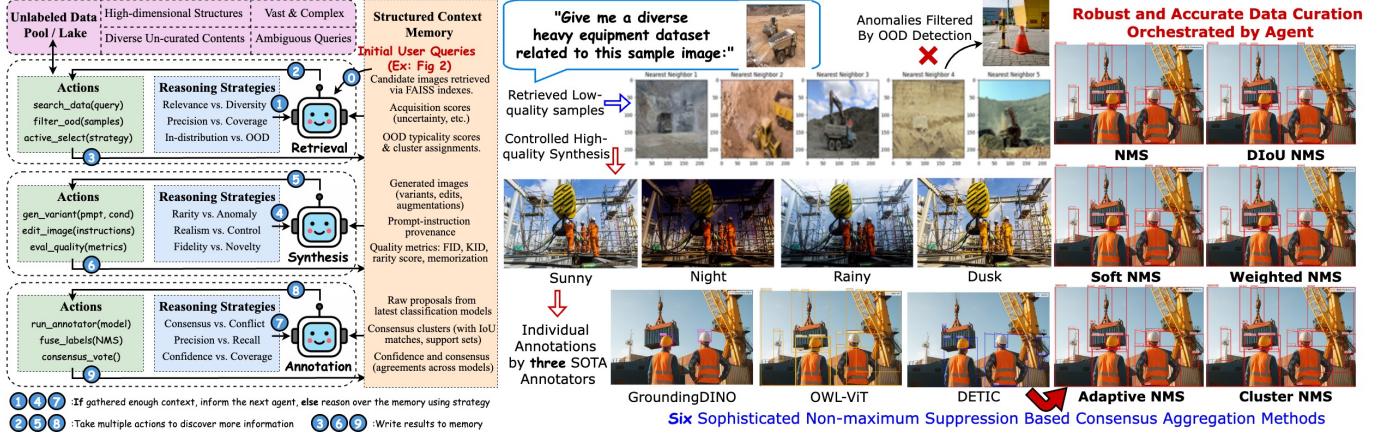


Fig. 1. Overview of LABELING COPILOT. **Left:** The system operates in three phases—Retrieval, Synthesis, and Annotation—with a Structured Context Memory storing candidates, synthetic variants, weak labels, and quality signals to guide reasoning. Retrieval balances relevance vs. diversity with active learning and OOD filtering; Synthesis adds rare but realistic data with fidelity and novelty checks; Annotation fuses outputs from multiple detectors via consensus and NMS. **Right:** an example trajectory shows how retrieved samples, controlled edits, and model proposals are combined into validated annotations, producing a curated dataset that feeds back into the loop.

1 An Agentic Framework for Data Curation: We design and implement a novel agentic system that unifies data retrieval, synthesis, and annotation into a cohesive, goal-driven workflow. Unlike rigid pipelines, our agent intelligently orchestrates these tools, utilizing quality signals passed between modules to determine whether to dynamically find more real data, synthesize novel examples for rare cases, or fuse annotations to enhance label quality.

2 A Scalable, Calibrated Retrieval Tool: We design an agent-controlled retrieval module that combines active learning strategies [27], [28] with Out-of-Distribution (OOD) detection [29]. By reformulating classical active learning algorithms within an FAISS-based approximate nearest neighbor framework, the tool delivers reliable quality signals at scale, handling datasets with over 10 million samples. This enables the agent to efficiently surface the most informative data for downstream model training.

3 A Language-Controllable Synthesis Tool: We empower the agent with a unified synthesis tool that generates targeted, realistic data from natural language instructions. By combining instruction-following diffusion models [30], [31] with multimodal large language models, the agent can create fine-grained semantic variations (e.g., “a car at night in the rain”) to systematically plan and address data gaps while preserving object identity.

4 A Multi-Tool Consensus Annotation Strategy: We propose a consensus-based annotation strategy where the agent orchestrates multiple foundation models (e.g., DETIC [32], GroundingDINO [9]) as independent “expert” tools. The agent then employs sophisticated aggregation mechanisms, such as Soft-NMS [33] and Weighted-NMS [34], to synthesize their diverse outputs, transforming noisy weak labels into high-quality pseudo-labels that are more robust than any single model could produce.

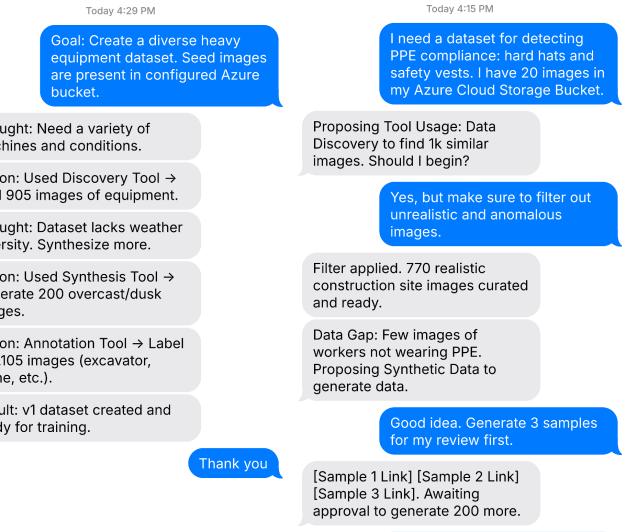


Fig. 2. Illustrative demonstration two trajectories for dataset curation: a fully autonomous agent performing discovery, synthesis, and annotation (left), and a human-in-the-loop session where a user guides the agent’s actions (right).

II. RELATED WORK

LABELING COPILOT addresses the critical challenge of automated dataset curation by integrating advances in data acquisition, data labeling, and data quality improvement. While prior efforts tackle these aspects in isolation, our agentic framework uniquely orchestrates them into a unified system that can reason about trade-offs and adapt strategies based on data characteristics and quality signals.

Data Acquisition and Discovery: Effective machine learning relies on acquiring sufficient, relevant data. Traditional approaches span *discovery*, *augmentation*, and *generation* [35]. Discovery efforts focus on indexing and sharing large repos-

itories: early platforms such as Google Fusion Tables and Kaggle democratized dataset access, while enterprise systems like *Datahub* provide versioning and metadata management. *GOODS* scales further by cataloging billions of datasets for efficient search and provenance. Despite these advances, identifying task-relevant visual data remains difficult due to the complexity of semantic similarity. Data augmentation has evolved from simple transformations to advanced mixing strategies, including *Mixup* [36], *CutMix* [37]. Automated policy search methods like *AutoAugment* [38] and *RandAugment* [39]. Recent transformer-based approaches such as *TransMix* [40] and *MixPro* [41] exploit attention maps for more intelligent mixing. Yet, these methods cannot produce semantic, photorealistic modifications (e.g., altering weather or lighting). Generative approaches address this gap. While platforms like Amazon Mechanical Turk support manual data creation, modern GANs [42] and diffusion models [43], [44] enable high-quality synthetic data for images, text, and tabular domains [45]. However, diffusion-based data often underperforms real samples [46], underscoring the need to integrate retrieval and generation intelligently.

Data Labeling and Annotation: Acquisition must be coupled with effective annotation strategies. Approaches range from fully supervised to weakly supervised, each with accuracy–scalability trade-offs. Semi-supervised learning leverages small labeled sets to expand larger unlabeled corpora, with classical methods including self-training [47], co-training [48], and tri-training [49], which underpin modern pseudo-labeling. Active learning prioritizes informative samples via uncertainty sampling [50], query-by-committee [51], and geometric strategies such as K-Center Greedy [27]. Recent work explores adaptive multi-armed bandit frameworks [52] to combine strategies dynamically. Weak supervision exploits noisy sources at scale. Data programming (e.g., *Snorkel* [53]) aggregates labeling functions probabilistically, while multi-instance learning (MIL) [54] infers instance-level labels from coarser supervision. Although powerful, these methods introduce noise that requires careful denoising.

Foundation Models for Vision Tasks: Large vision-language foundation models now provide strong automated labeling capabilities. *CLIP* [55] supports zero-shot classification, *DEiTIC* [32] and *GroundingDINO* [9] enable open-vocabulary detection, and *SAM* [10] delivers universal segmentation. However, their outputs often conflict, motivating consensus-based fusion. Recent multitask models such as *Florence-2* [6], trained on FLD-5B, highlight the promise of unified architectures, yet curating such datasets remains a major bottleneck requiring intelligent filtering. Beyond acquisition, data quality improvement is crucial when new data is limited. Systems such as *HoloClean* [56], *ActiveClean* [57], and *BoostClean* [58] demonstrate structured-data cleaning at scale. In computer vision, filtering typically relies on non-maximum suppression and its variants to reduce redundant detections, though these are usually applied per model rather than orchestrated across multiple detectors.

Positioning of Our Work: Existing approaches address individual steps in data curation but lack intelligent orchestration. LABELING COPILOT uniquely integrates retrieval, synthesis, and consensus annotation into an agentic framework that reasons about data quality, adapts strategies iteratively, and makes principled trade-offs. To our knowledge, this is the first work to frame dataset curation as an agentic research process, transforming disconnected tools into a cohesive, goal-driven system.

III. METHODOLOGY: AN AGENT-DRIVEN TOOLKIT

The LABELING COPILOT framework is built around a central orchestrator agent that intelligently manages a suite of specialized tools, qualifying as a deep research agent under the formal definition established by recent literature [59] through its high search intensity (processing millions of information units via scalable active learning algorithms), reasoning intensity across all three complexity dimensions (finding information units via sophisticated retrieval strategies, processing them through multi-model consensus mechanisms, and combining conflicting outputs via advanced NMS variants), and dynamic agent orchestration where the central coordinator intelligently selects and sequences specialized tools based on data characteristics and quality signals, with the capability to iteratively adapt strategies (e.g., invoking synthesis when discovery yields insufficient diversity) throughout the curation workflow. This section details the core tools the agent uses for data discovery, synthesis, and annotation.

A. The Data Discovery Tool

The agent’s first task is to source relevant data from a large, unlabeled data pool \mathcal{U} . It employs the data discovery, which is designed to select an informative batch \mathcal{B} for labeling. To enable efficient similarity search, all data points in \mathcal{U} are pre-indexed using FAISS [60]. The workflow consists of two main stages: selecting a batch of candidate samples using a scalable active learning (AL) strategy, and then filtering out-of-distribution (OOD) samples from that batch.

Stage 1: Scalable Active Learning for Candidate Selection
 Canonical active learning (AL) algorithms are fundamentally incompatible with the scale of modern datasets. Their reliance on operations that scan the entire unlabeled data pool—such as full distance matrix computations or exhaustive model inference—introduces a computational bottleneck that is untenable in a big data context. Our methodology systematically dismantles this bottleneck by reformulating core AL strategies around the sub-linear time complexities of Approximate Nearest Neighbor (ANN) search. We utilize the FAISS library not only for search but also as a computational engine to replace brute-force operations with efficient, index-based approximations.

We anchor our methodology in FAISS, a library purpose-built for vector similarity search on massive-scale datasets. Rather than using it as a simple search tool, we employ its indexing structures as a computational substrate to overcome

specific scalability challenges in active learning. Our strategic choices include:

- *Tackling Search Complexity with Inverted File Systems (IVF)*: To avoid linear scans, we partition the vector space using IndexIVF structures. The dataset is divided into n_{list} Voronoi cells, and at query time, the search is constrained to a small subset, n_{probe} , of these cells. This reduces search complexity from $\mathcal{O}(N)$ to approximately $\mathcal{O}(\frac{n_{probe}}{n_{list}} \cdot N)$, forming the basis of our localized sampling strategies.
- *Managing Memory Footprint with Product Quantization (PQ)*: For datasets exceeding main memory capacity, we employ IndexIVFPQ to compress vectors. By decomposing vectors into sub-vectors and quantizing each independently, PQ dramatically reduces the per-vector memory cost from 4d bytes to as few as 8 or 16 bytes, making billion-scale in-memory active learning feasible.
- *Ensuring High-Recall Search with HNSW Graphs*: In regimes where high search accuracy is paramount, we utilize Hierarchical Navigable Small World (IndexHNSW) graph indexes. HNSW provides superior speed-recall trade-offs compared to IVF-based methods and can serve either as a primary index or as a powerful coarse quantizer for a hybrid IVF system (e.g., IVF65536_HNSW32_Flat).

The efficacy of our entire framework hinges on the synergistic use of two fundamental FAISS operations: `search()` for efficient candidate localization and `reconstruct_batch()` for retrieving the full-precision vectors required for model training. This combination transforms the FAISS index from a passive search structure into an active engine for scalable data selection.

Scalable Reformulation of Active Learning Algorithms:

With this tool, our primary contribution is the redesign of classic AL algorithms to operate on small, intelligently-selected data subsets rather than the entire unlabeled pool.

K-Center Greedy via Candidate Subsampling: The K-Center Greedy algorithm seeks to select points that maximally cover the feature space. Its brute-force implementation, however, requires iteratively computing distances from every unlabeled point to the growing set of labeled centers, a process with complexity that scales prohibitively with the dataset size $|\mathcal{U}|$.

To overcome this, we implement an approximate version that operates on a fixed-size candidate pool $\mathcal{U}_c \subset \mathcal{U}$ of size N_c . This pool is randomly sampled from the unlabeled set at the beginning of each selection round. The greedy selection logic then proceeds as normal, but is confined to this computationally manageable subset, by using `faiss.pairwise_distances`, we leverage optimized BLAS routines for the core distance calculations. This reformulation bounds the complexity of selecting a batch of size B to $\mathcal{O}(B \cdot N_c \cdot |\mathcal{L}| \cdot d)$, making the algorithm's runtime independent of the total dataset size $|\mathcal{U}|$ and dependent only on the configurable pool size N_c .

Localized Acquisition for Uncertainty and Representative Sampling: A broad class of powerful AL strategies, includ-

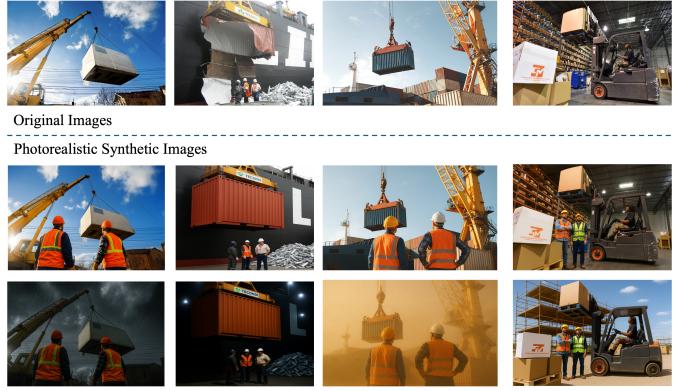


Fig. 3. The LABELING COPILOT’s synthesis generates diverse, photorealistic variants (middle and bottom rows) from original industrial images (top row). This capability is critical for creating training data for workplace hazards that are unethical or impossible to capture in reality, such as workers standing under a suspended load (a serious OSHA violation). The agent can also alter environmental conditions (e.g., adding rain or dust) and context (e.g., moving an indoor scene outdoors) to create data for rare scenarios, thereby systematically improving the robustness of safety detection computer vision models.

ing margin sampling, entropy sampling, and representative sampling, is predicated on first identifying points of high uncertainty. The common bottleneck across all these methods is the need to perform model inference over every point in \mathcal{U} to compute these uncertainty scores.

Our solution is to replace this exhaustive inference with an efficient, ANN-driven localization. We hypothesize that the most informative samples for a model are located in the feature space neighborhood of the data it has already been trained on. We formalize this as a generalized localized acquisition strategy, detailed in Algorithm 1.

This strategy creates a compact “micro-pool” \mathcal{N}_u of promising candidates, drastically reducing the scope of expensive downstream computations. For *Margin Sampling*, we compute decision boundary distances for points within \mathcal{N}_u . For more complex *Representative Sampling* strategies, we first identify the most uncertain points within \mathcal{N}_u and then apply a fast clustering algorithm (e.g., MiniBatchKMeans) to this much smaller set to select a diverse batch of medoids. In all cases, the cost of inference is reduced from $\mathcal{O}(|\mathcal{U}|)$ to $\mathcal{O}(K_s)$, where the neighborhood size K_s is a tunable hyperparameter orders of magnitude smaller than $|\mathcal{U}|$.

Stage 2: Filtering Spurious Samples with OOD Detection:

To enforce distributional consistency and ensure that the active learning process focuses on semantically relevant samples, we introduce a probabilistic filtering stage after candidate retrieval. This step is designed to reject statistical outliers that may be returned by the ANN search, particularly from sparse regions of the embedding space. We implement this filter by adapting the Forte typicality estimation framework [29], which operates in a learned self-supervised representation space. The process consists of two primary stages: manifold modeling and probabilistic filtering.

First, we model the high-dimensional manifold of the known

Algorithm 1 Generalized Localized Active Learning Framework

Require: FAISS index \mathcal{I} , Labeled set IDs $\text{ids}(\mathcal{L})$, Batch size B , Neighborhood size K_s .

- 1: $\mathbf{V}_{\mathcal{L}} \leftarrow \mathcal{I}.\text{reconstruct_batch}(\text{ids}(\mathcal{L}))$. \triangleright Retrieve labeled vectors
- 2: $\mathbf{q} \leftarrow \text{ComputeCentroid}(\mathbf{V}_{\mathcal{L}})$. \triangleright Define search query (e.g., centroid)
- 3: $\text{ids}(\mathcal{N}) \leftarrow \mathcal{I}.\text{search}(\mathbf{q}, K_s)$. \triangleright Find local neighborhood via ANN search
- 4: $\text{ids}(\mathcal{N}_u) \leftarrow \text{ids}(\mathcal{N}) \setminus \text{ids}(\mathcal{L})$. \triangleright Isolate unlabeled candidates
- 5: $\mathbf{V}_{\mathcal{N}_u} \leftarrow \mathcal{I}.\text{reconstruct_batch}(\text{ids}(\mathcal{N}_u))$. \triangleright Retrieve candidate vectors
- 6: Train model \mathcal{M} on $(\mathbf{V}_{\mathcal{L}}, \mathbf{y}_{\mathcal{L}})$.
- 7: Compute acquisition scores for all $\mathbf{v}_i \in \mathbf{V}_{\mathcal{N}_u}$ using \mathcal{M} .
- 8: Select batch \mathcal{B} of size B from $\text{ids}(\mathcal{N}_u)$ based on scores.
- 9: **return** \mathcal{B} .

in-distribution data, represented by the feature set $\mathbf{V}_{\mathcal{L}}$ from the labeled pool \mathcal{L} . We employ a Gaussian Mixture Model (GMM) for this task, chosen for its ability to capture complex, multimodal data structures inherent in real-world visual classes. The GMM parameters, $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^K$, representing the mixture weights, means, and covariances, are estimated from $\mathbf{V}_{\mathcal{L}}$ via the Expectation-Maximization (EM) algorithm. This yields a probabilistic model of the in-distribution data density: $p(\mathbf{x}|\Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$

Second, for each retrieved candidate vector \mathbf{x}' , we use the fitted GMM to perform probabilistic filtering. Our innovation here is to derive a continuous typicality score rather than a discrete classification. We calculate the posterior probability, or responsibility, of each mixture component k for generating the sample \mathbf{x}' : $\gamma_k(\mathbf{x}') = P(k|\mathbf{x}', \Theta) = \frac{\pi_k \mathcal{N}(\mathbf{x}'|\mu_k, \Sigma_k)}{p(\mathbf{x}'|\Theta)}$. The final typicality score, $S(\mathbf{x}')$, is the maximum responsibility across all components, $S(\mathbf{x}') = \max_k \gamma_k(\mathbf{x}')$. This score measures how well \mathbf{x}' aligns with the densest regions of the learned in-distribution manifold. A sample is accepted into the final batch for labeling only if its score surpasses a predefined threshold, $S(\mathbf{x}') \geq \tau$. This mechanism acts as a robust semantic guardrail, preventing the labeling budget from being expended on noisy or irrelevant outliers and thereby promoting a more efficient and stable active learning loop.

B. The Synthetic Data Tool

When the agent determines that the existing data is insufficient (e.g., lacks diversity or rare cases), it employs the synthetic data to generate new, high-quality images.

Step 1: Choosing a Generation Technique: The agent selects the best synthesis method based on the complexity of the desired image.

- For simple scenes, the agent uses state-of-the-art captioning models (GPT-4, BLIP, Florence 2) to generate rich text prompts, which are then fed to Stable Diffusion, DALL-E.

- For modifying natural scenes, the agent can use the original image as a strong conditioning reference for an image-to-image Stable Diffusion model, setting a low strength value to generate realistic variations.
- For complex or niche domains, the agent uses GPT-4o to generate a large set of specific editing instructions for a single image (e.g., “add rain,” “make it nighttime”). It then executes these instructions using a prompt-based image editing model, such as InstructPix2Pix. We achieve the highest quality synthetic images using natively multimodal large language models like GPT-Image-1.

Step 2: Evaluating Synthetic Data Quality: Before adding synthetic data to the dataset, the agent rigorously evaluates its quality using a comprehensive suite of metrics to assess four key criteria: fidelity, diversity, rarity, and memorization. [61]

- *Fidelity & Diversity:* The agent computes metrics like Fréchet Inception Distance (FID), Kernel Inception Distance (KID), precision, recall, density, and coverage in the Di-noV2 representation space.
- *Memorization:* When using fine-tuned models, the agent checks for overfitting and memorization using metrics such as Percentage of authentic samples (AuthPCT) and Feature Likelihood Score (FLS).
- *Automated Selection:* The agent treats these normalized scores as logits, allowing it to use top-k or top-p sampling to automatically select the best-performing models and prompts for generation, with human experts validating the final hyperparameters.

C. The Annotation Tool

Once a batch of data is curated, the agent employs the annotation tool to generate precise and reliable labels. We present a high-throughput weak supervision framework designed for HPC environments, which transforms vast, unlabeled image corpora into high-quality annotated datasets. The methodology is architected as a two-phase, massively parallel pipeline: (1) weak label generation from a heterogeneous ensemble of foundation models, and (2) consensus-based fusion to distill a robust ground truth from these noisy, multi-source predictions.

Phase 1: Weak Supervision via a Heterogeneous Ensemble: The foundational principle of our framework is to leverage the complementary strengths of diverse, open-vocabulary object detection models to generate a rich set of weak labels. The heterogeneity of the model architectures is a critical design choice, as it promotes uncorrelated error modes, which are essential for effective downstream aggregation and error correction. Our programmatic annotator ensemble comprises models representing distinct architectural paradigms:

- *DETIC:* A CLIP-driven detector that excels at zero-shot generalization by dynamically embedding a user-defined text vocabulary into its classification head.
- *GroundingDINO:* A sophisticated encoder-decoder transformer that performs language-to-vision grounding, enabling it to detect objects specified by complex, free-form text prompts.

- **OWL-ViT**: A Vision Transformer (ViT) architecture adapted for open-vocabulary detection, providing a distinct feature extraction and localization mechanism.

For a given image $I \in \mathcal{D}$ from a dataset corpus and a shared target vocabulary \mathcal{V} , each model M_i in the ensemble $\mathcal{M} = \{M_1, \dots, M_N\}$ generates an independent set of proposals P_i . Each proposal $p_{ij} \in P_i$ is a tuple (b_{ij}, c_{ij}, s_{ij}) , consisting of a bounding box $b_{ij} \in \mathbb{R}^4$, a class label $c_{ij} \in \mathcal{V}$, and a model-specific confidence score $s_{ij} \in [0, 1]$.

HPC Implementation Strategy: The annotation of each image by each model is a stateless, independent task. This allows for near-linear scalability on a distributed computing cluster, where each image-model pair can be scheduled as a separate job. To facilitate this and ensure modularity, the output for each image is a collection of standardized PASCAL VOC XML files—one for each model. This use of a file-based intermediate representation decouples the generation and fusion stages, a critical feature for managing complex, large-scale workflows and enhancing fault tolerance.

Phase 2: Consensus-Based Annotation Fusion: The weak labels produced in Phase 1 are numerous, often conflicting, and contain significant noise. The fusion phase distills these raw proposals into a single, high-confidence annotation set for each image through a principled, voting-based algorithm.

Consensus Set Generation via Support-Based Matching: Unlike traditional NMS, which operates on proposals from a single model, our fusion logic must reconcile predictions from multiple, disparate sources. For each class $c \in \mathcal{V}$, we aggregate all proposals $\{p_j\}$ from all models. For each proposal p_j , we compute its *support set* by identifying the best-matching proposal (if any) from each of the other models M_i based on an Intersection over Union (IoU) threshold, τ_{iou} . The collection of a proposal and its supporting matches forms a consensus cluster.

From each cluster, we derive a fused bounding box b^* by averaging the coordinates of all constituent proposals. A consensus confidence score, \mathcal{S} , is then computed. This score is not a simple average of model scores but rather a measure of inter-model agreement, defined as the proportion of unique models in the ensemble that support the cluster:

$$\mathcal{S}(C_k) = \frac{|\{M_i \in \mathcal{M} \mid \exists p \in P_i \text{ s.t. } p \in C_k\}|}{N}$$

where C_k is the set of proposals in the k -th cluster. This mechanism intrinsically up-weights detections that are consistently identified across different model architectures. The output of this step is a refined set of candidate annotations, $A^* = \{(b_k^*, c_k, \mathcal{S}_k)\}_{k=1}^K$.

Finalization via Configurable Non-Maximal Suppression: The candidate set A^* represents high-agreement detections but may still contain spatial overlaps. We employ a final, configurable filtering module based on advanced NMS variants to resolve these conflicts. The consensus score \mathcal{S} is used as the primary sorting criterion. Our framework integrates several strategies to handle diverse object distributions:

- **DIoU-NMS:** We utilize a metric that penalizes for the distance between box centers in addition to IoU. This is our default, as it yields more robust suppression for occluded or proximate objects by considering the centrality of the detections.
- **Soft-NMS:** For dense scenes, we employ a non-destructive suppression where the scores of overlapping boxes are decayed as a Gaussian function of their IoU, preserving plausible but overlapping detections.

The final output of this two-phase pipeline, $A_{final} = \text{NMS}(A^*, \tau_{nms})$, is a single, programmatically-generated annotation file for each image, synthesized from a robust process of evidence aggregation, voting, and refinement.

IV. RESULTS AND EVALUATION

Evaluating a long-running, autonomous agent like LABELING COPILOT presents a non-trivial benchmarking challenge. Its purpose is to generate datasets, a task with a vast and open-ended action space. Consequently, a traditional end-to-end evaluation based on a single, fixed downstream task would fail to capture the agent’s general-purpose utility across diverse domains. To address this, we adopt a **data-centric evaluation protocol**: Instead of measuring a single downstream outcome, we perform a rigorous, component-wise analysis focused on the quality and characteristics of the data artifacts produced at each stage of the agentic workflow. This approach allows us to thoroughly validate the performance of LABELING COPILOT’s core tools on both academic and industry-specific datasets. The following sections present the key findings from our experiments, demonstrating the effectiveness of each component.

A. Protocol for Scalability and Efficacy Evaluation

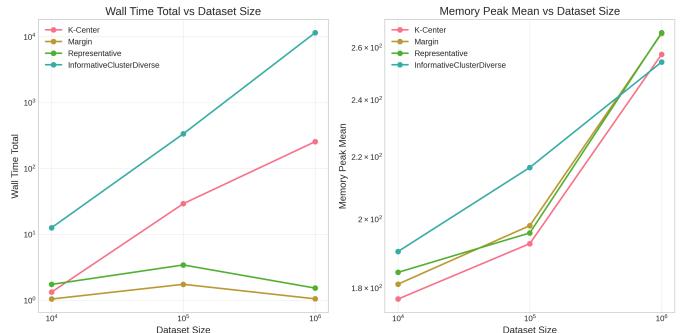


Fig. 4. Computational scaling behavior of active learning algorithms across dataset sizes from 10K to 1M samples. The left panel shows wall-clock time scaling on log-log axes, revealing fundamental algorithmic complexity differences: InformedClusterDiverse exhibits super-linear growth (15s to 10,000s), K-Center demonstrates favorable sub-linear scaling (1s to 250s), while Margin and Representative show flat profiles. The right panel displays memory consumption scaling, with all methods following similar power-law growth patterns that converge at large scales, suggesting memory bottlenecks shift from vector storage to auxiliary computational structures. The parallel scaling curves indicate that algorithmic choice, rather than data structure optimization, becomes the primary determinant of computational feasibility at production scales. Note that these measurements aggregate across different index architectures (Flat, IVF, HNSW) that were varied with dataset size.

We designed a rigorous experimental protocol to evaluate the performance of our scalable AL framework under diverse conditions. The independent variables include: (i) *dataset scale*, ranging from $N = 10^4$ to 10^8+ samples with dimensionality $d \in \{64, 128, 256\}$; (ii) *index architecture*, which progresses with scale from Flat to IVF, Flat and ultimately to compressed IVF, PQ structures; and (iii) *AL hyperparameters*, specifically the candidate pool size N_c for K-Center and the local neighborhood size K_s for localized methods. Performance is assessed using two dependent measures: *sample efficiency*, defined as downstream classifier accuracy versus the number of acquired labels, and *computational cost*, measured by wall-clock batch selection time, CPU usage, and peak memory footprint.

Our analysis reveals that algorithmic architecture is the dominant factor in system performance, far outweighing vector database optimization. At the 10M sample scale, the choice of algorithm created a distinct performance hierarchy, while different index architectures all converged to a similar accuracy (0.843-0.851 AUC, $<1\%$ variance). The FAISS-optimized K-Center algorithm emerges as the optimal choice, matching the sample efficiency of the more complex InformativeClusterDiverse (both reaching an AUC of 0.90) while being 40 times more computationally efficient (250 vs. 10,000+ seconds). This efficiency is critical, as InformativeClusterDiverse exhibits a debilitating super-linear time complexity, scaling from 15 to over 10,000 seconds between 10,000 and 1 million samples. Simpler methods like Margin and Representative failed to scale, indicated by flat performance curves that suggest systematic experimental failures. Ultimately, these findings establish that for extreme-scale active learning, optimization should prioritize algorithmic robustness, where K-Center provides the best balance of sample efficiency and computational feasibility.

B. Quantitative Analysis of Annotation Quality

To validate the efficacy and scalability of our automated annotation framework, we conducted a comprehensive evaluation across three canonical object detection benchmarks: COCO, Open Images, and Pascal VOC. These datasets were strategically selected to represent a broad spectrum of challenges, including varying object scales, annotation densities, and class distributions. The programmatic annotations generated by our framework were evaluated against the original ground truth labels, with the results detailed in Table I. Our analysis compares the performance of six distinct Non-Maximal Suppression (NMS) algorithms, examining their impact on final annotation quality across key metrics, including precision, recall, F1-score, and mean Average Precision (mAP).

1) *Comparative Performance of NMS Algorithms:* The choice of NMS algorithm proves to be an essential factor in optimizing the precision-recall trade-off inherent in our high-recall, ensemble-based proposal generation system. While several methods demonstrate strong performance, nuanced differences emerge that highlight their suitability for different data characteristics.

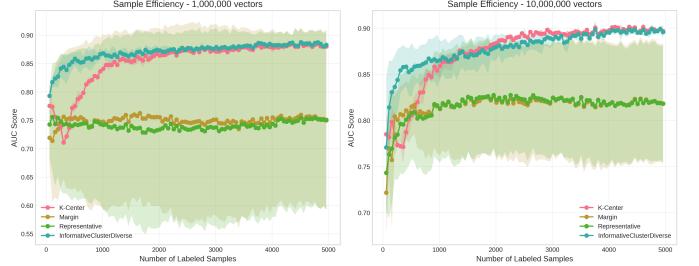


Fig. 5. Sample efficiency curves for active learning algorithms at 1M and 10M vector scales. Learning curves display AUC performance as a function of labeled sample count across four sampling strategies: K-Center (diversity-based), Margin (uncertainty-based), Representative (combining uncertainty and diversity), and Informative Cluster Diverse (combining uncertainty, clustering, and diversity). At 1M scale (left), InformativeClusterDiverse and K-Center achieve superior performance (0.87-0.88 AUC), while Margin and Representative plateau at lower performance levels (0.74-0.75 AUC). At 10M scale (right), K-Center matches InformativeClusterDiverse’s final performance (both reaching 0.90 AUC) while maintaining faster convergence. The performance gap between sophisticated and lightweight methods widens at extreme scale, with robust algorithms demonstrating consistent improvement throughout the labeling budget while simpler methods exhibit early saturation. Curves represent mean performance across multiple random seeds with standard deviation bands, using synthetic binary classification tasks with 128-dimensional feature vectors.

A defining strength of our framework is its exceptionally high recall, a direct consequence of our ensemble-based weak supervision strategy. The system is intentionally prolific, generating a rich superset of candidate detections. For instance, on COCO, the pipeline generated an average of 14.20 proposals per image using standard NMS, compared to only 7.43 ground-truth objects. This results in outstanding overall recall scores (e.g., **0.802 on COCO, 0.939 on Pascal VOC** with Soft-NMS). The primary role of the NMS algorithm is therefore to distill this high-recall, lower-precision proposal set into a high-quality final annotation set.

Soft-NMS consistently emerges as a top-tier performer, particularly in recall-centric metrics. It achieves the highest recall on all three datasets (COCO: **0.802**, Open Images: **0.456**, Pascal VOC: **0.939**) and secures the highest mAP across all thresholds on COCO and Open Images. Its strategy of decaying the scores of overlapping boxes rather than performing hard elimination is demonstrably superior for preserving valid detections, especially in dense scenes where objects have significant spatial overlap.

Adaptive NMS shows strong performance in balancing precision and recall, achieving the highest F1-score on Pascal VOC (**0.517**) and the highest precision on both COCO (**0.494**) and Open Images (**0.235**). This suggests its dynamic thresholding mechanism is effective in contexts where object densities vary.

Conversely, **Cluster NMS** consistently underperforms in F1-score while achieving the highest Average IoU across all datasets (COCO: **0.134**, Open Images: **0.128**, Pascal VOC: **0.190**). This suggests that it suppresses too many true positives—reducing recall—yet the boxes it retains are more precisely localized. Such a trade-off makes it less suitable for our goal of maximizing object discovery. In contrast, **NMS**,

TABLE I
COMPREHENSIVE ANNOTATION QUALITY COMPARISON OF NMS METHODS ACROSS COCO, OPEN IMAGES, AND PASCAL VOC DATASETS

Dataset	Method	Precision	Recall	F1	mAP @.5	mAP @.75	mAP @.5:.95	Avg IoU	Overlap Ratio	Sm% (<1%)	Med% (1-5%)	Lg% (>5%)	Imbalance Ratio	Avg GT Obj/Img	Avg Pred Obj/Img	Avg Correct Obj/Img	Coverage GT %	Discovered New Classes
COCO	NMS	0.486	0.799	0.565	0.460	0.397	0.368	0.118	0.920	48.4	26.6	25.0	13325	7.43	14.20	5.24	92.4	598
	Adaptive NMS	0.494	0.771	0.564	0.445	0.384	0.356	0.112	0.912	49.2	26.5	24.3	12753	7.43	13.43	5.06	92.4	598
	Soft NMS	0.481	0.802	0.562	0.461	0.401	0.371	0.123	0.923	48.1	26.8	25.2	13489	7.43	14.55	5.26	92.4	598
	DIoU NMS	0.485	0.799	0.565	0.460	0.397	0.368	0.119	0.921	48.3	26.7	25.0	13344	7.43	14.24	5.24	92.4	598
	Weighted NMS	0.486	0.799	0.565	0.460	0.397	0.368	0.118	0.920	48.4	26.6	25.0	13325	7.43	14.20	5.24	92.4	598
	Cluster NMS	0.488	0.762	0.554	0.449	0.392	0.362	0.134	0.904	41.1	28.8	30.1	10178	7.43	11.93	4.62	92.4	598
Open Images	NMS	0.234	0.454	0.252	0.174	0.148	0.142	0.116	0.983	45.6	26.5	27.9	65813	8.15	22.55	3.48	57.7	903
	Adaptive NMS	0.235	0.442	0.251	0.170	0.144	0.138	0.115	0.982	46.4	26.4	27.2	64196	8.15	21.74	3.39	57.7	903
	Soft NMS	0.230	0.456	0.249	0.175	0.150	0.143	0.119	0.984	45.1	26.5	28.4	65877	8.15	23.27	3.51	57.7	903
	DIoU NMS	0.233	0.455	0.252	0.174	0.148	0.142	0.117	0.983	45.6	26.5	28.0	65824	8.15	22.63	3.49	57.7	903
	Weighted NMS	0.234	0.454	0.252	0.174	0.148	0.142	0.116	0.983	45.6	26.5	27.9	65813	8.15	22.55	3.48	57.7	903
	Cluster NMS	0.233	0.445	0.250	0.173	0.149	0.142	0.128	0.980	39.5	27.3	33.2	46988	8.15	19.25	3.21	57.7	903
Pascal VOC	NMS	0.404	0.937	0.516	0.420	0.390	0.363	0.171	0.904	27.0	24.6	48.4	6562	2.29	6.13	2.03	98.0	82
	Adaptive NMS	0.410	0.919	0.517	0.413	0.383	0.358	0.163	0.895	27.3	24.7	48.0	6270	2.29	5.83	1.97	98.0	82
	Soft NMS	0.401	0.939	0.513	0.420	0.392	0.365	0.176	0.907	26.8	24.7	48.4	6627	2.29	6.26	2.04	98.0	82
	DIoU NMS	0.404	0.937	0.516	0.420	0.390	0.364	0.171	0.904	27.0	24.6	48.4	6566	2.29	6.15	2.03	98.0	82
	Weighted NMS	0.404	0.937	0.516	0.420	0.390	0.363	0.171	0.904	27.0	24.6	48.4	6562	2.29	6.13	2.03	98.0	82
	Cluster NMS	0.404	0.930	0.515	0.417	0.389	0.362	0.190	0.898	24.1	24.4	51.5	5721	2.29	5.83	1.98	98.0	82

DIoU NMS, and **Weighted NMS** deliver nearly identical results across most metrics, providing reliable baselines though rarely surpassing adaptive methods like Soft-NMS.

C. Performance on Challenging Data Characteristics

A key objective of our evaluation was to assess the framework’s robustness to the complexities inherent in large-scale, real-world datasets.

COCO: Handling High Annotation Density and Small Objects. The COCO dataset, characterized by its high density of objects (7.43 GT Obj/Img) and a significant proportion of small objects (48.4% categorized as ‘Sm%’), presents a formidable challenge. Our framework’s ability to generate a dense proposal set (14.20 Pred Obj/Img for NMS) is crucial in this context. The success of **Soft-NMS**, which achieves the highest mAP@.75 (**0.401**) and overall mAP (**0.371**), underscoring its efficacy in resolving ambiguous overlaps in cluttered scenes without erroneously suppressing valid, closely packed objects. The high recall achieved on COCO further suggests that the multi-model ensemble is capable of identifying small objects that a single detector might miss, and the consensus mechanism effectively preserves these fine-grained detections.

Open Images: Resilience to Extreme Scale and Class Imbalance. The Open Images dataset serves as a proxy for web-scale data, with its 903 discovered classes and an extreme class imbalance ratio (up to **65,877**). In this regime, our framework’s high-recall nature becomes an important discovery mechanism. While the overall F1-score is lower than on other datasets, the framework maintains a respectable recall (e.g., **0.456** with Soft-NMS). This indicates that the heterogeneous ensemble of weak labelers is effective at discovering instances even from rare, long-tail classes. However, this comes at the cost of precision, as evidenced by the high number of predicted objects per image (23.27 for Soft-NMS) compared to the number of correct detections (3.51). This highlights the fundamental challenge of managing the precision-recall trade-off in massively multi-class environments, a task where the choice of NMS algorithm is paramount. **Soft-NMS** again

achieves the highest mAP, indicating a better-ranked list of detections despite the noise.

Pascal VOC: Performance on High-Coverage, Simpler Scenes. On the Pascal VOC dataset, which features fewer objects per image (2.29) and a very high ground-truth coverage rate (98.0%), the framework demonstrates exceptional performance. The recall is outstanding, peaking at **0.939** with Soft-NMS, indicating near-complete object discovery. In this less-cluttered environment, the task of NMS is simpler. **Adaptive NMS** achieves the highest F1-score (**0.517**) by attaining the best balance between its high precision (**0.410**) and recall (**0.919**), making it the optimal choice for datasets with these characteristics.

V. DISCUSSION

The LABELING COPILOT framework demonstrates the power of an agentic approach to solve the complex, multi-stage challenge of computer vision data curation. Rather than a rigid, linear pipeline, our system is a dynamic workflow managed by a central orchestrator agent that intelligently deploys a suite of specialized tools for data discovery, synthesis, and annotation. This design philosophy is intentionally built around the core primitives of data-centric AI, ensuring the framework’s longevity and adaptability.

A key architectural principle of LABELING COPILOT is the separation of fundamental data operations from their underlying model implementations. We treat Discovery, Synthesis, and Annotation as timeless primitives in the data curation lifecycle. The true innovation is not the specific models used today, but the agentic workflow that orchestrates these primitives. This separation is what makes the system robust, versatile, and future-proof. This philosophy is realized through a modular, “hotswappable” engineering design. The agent interacts with data through standardized interfaces, such as cloud storage buckets and PASCAL VOC annotation formats. Each tool—Discovery, Synthesis, and Annotation—is a containerized module that adheres to this contract. This has profound practical implications:

- **Effortless Upgrades:** The framework is not architecturally dependent on any single foundation model. For example,

the current Consensus Annotation tool uses an ensemble including GroundingDINO and DETIC. Should a superior open-vocabulary detector emerge, it can be integrated simply by wrapping it in a new container that conforms to the established input/output format. The core agentic workflow requires no re-architecting.

- **Extensibility:** New capabilities can be added as new primitive tools. One could easily envision adding a “Data Repair” tool that uses models to find and fix labeling errors or a “Data Privacy” tool that automatically blurs sensitive information. The agent’s capabilities can be extended without altering the existing components.
- **Decoupled Intelligence:** The agent’s reasoning—deciding when to synthesize more data based on the output of the discovery tool, for example—is separate from the execution of the tools themselves. This allows the agent’s strategic intelligence to be improved independently of the tools’ capabilities. This engineering also allows us to not overload the primary orchestrator agents’ context window, allowing longer-term coherence.

The agent’s ability to plan, incorporate feedback from tools, and flexibly execute these tools—sequentially, in parallel, or in iterative cycles—is the cornerstone of the framework. For example, if the Data Discovery Tool returns a sparse dataset for a particular class, the agent can autonomously decide to invoke the Synthetic Data Tool to generate new examples before proceeding to annotation. This decision-making capability allows for continuous, targeted refinement of the dataset.

This agent-driven design meets the dual requirements of generality and specificity. The agent and its core toolkit are broadly applicable to a wide range of CV tasks, from object detection to panoptic segmentation. Simultaneously, the agent’s strategies and tool parameters can be customized for specific domains. By combining the outputs of its tools with human-in-the-loop feedback, the orchestrator agent learns and adapts, progressively improving the dataset in an iterative cycle. This transforms data curation from a manual, fragmented process into an intelligent, automated, and versatile solution for diverse AI challenges.

VI. CONCLUSION

To address the critical bottleneck of data curation in computer vision, we introduced LABELING COPILOT, the first deep research agent for automating this complex task. Our system replaces rigid pipelines with a dynamic agentic workflow that intelligently unifies data discovery, synthesis, and annotation. Large-scale validation confirmed our approach: the Calibrated Discovery tool employs techniques that are up to 40 times more computationally efficient than optimized alternatives, and the Consensus Annotation module achieves a 37.1% mAP on the dense COCO dataset, discovering 598 new classes and labeling twice as many objects per image. These results validate that an agentic framework built on scalable, individually optimized tools provides a robust foundation for creating industrial-scale datasets. The modular design ensures LABELING COPILOT is an extensible and versatile solution,

representing a significant step forward in solving critical challenges for data-centric AI.

REFERENCES

- [1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [3] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallochi, A. Kolesnikov *et al.*, “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,” *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.
- [4] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis) contents: A survey of dataset development and use in machine learning research,” *Patterns*, vol. 2, no. 11, 2021.
- [5] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, “Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 17612–17625, 2022.
- [6] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan, “Florence-2: Advancing a unified representation for a variety of vision tasks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829.
- [7] H.-a. Gao, J. Geng, W. Hua, M. Hu, X. Juan, H. Liu, S. Liu, J. Qiu, X. Qi, Y. Wu *et al.*, “A survey of self-evolving agents: On path to artificial super intelligence,” *arXiv preprint arXiv:2507.21046*, 2025.
- [8] J. Fang, Y. Peng, X. Zhang, Y. Wang, X. Yi, G. Zhang, Y. Xu, B. Wu, S. Liu, Z. Li *et al.*, “A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems,” *arXiv preprint arXiv:2508.07407*, 2025.
- [9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” in *European conference on computer vision*. Springer, 2024, pp. 38–55.
- [10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [11] T. Sumers, S. Yao, K. Narasimhan, and T. Griffiths, “Cognitive architectures for language agents,” *Transactions on Machine Learning Research*, 2023.
- [12] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *Advances in neural information processing systems*, vol. 35, pp. 25278–25294, 2022.
- [13] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, “Datacomp: In search of the next generation of multimodal datasets,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 27092–27112, 2023.
- [14] A. Birhane, V. U. Prabhu, and E. Kahembwe, “Multimodal datasets: misogyny, pornography, and malignant stereotypes,” *arXiv preprint arXiv:2110.01963*, 2021.
- [15] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8340–8349.
- [16] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, “Wilds: A benchmark of in-the-wild distribution shifts,” in *International conference on machine learning*. PMLR, 2021, pp. 5637–5664.
- [17] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferante, “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12592–12594, 2020.

- [18] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.
- [19] H. Su, J. Deng, and L. Fei-Fei, “Crowdsourcing annotations for visual object detection.” *HCOMP@ AAAI*, vol. 1, 2012.
- [20] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, “Training object class detectors with click supervision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6374–6383.
- [21] C. Northcutt, L. Jiang, and I. Chuang, “Confident learning: Estimating uncertainty in dataset labels,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.
- [22] B. Settles, “Active learning literature survey,” 2009.
- [23] J. Yang, K. Zhou, Y. Li, and Z. Liu, “Generalized out-of-distribution detection: A survey,” *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.
- [24] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” *arXiv preprint arXiv:2112.10741*, 2021.
- [25] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.
- [26] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
- [27] O. Sener and S. Savarese, “Active learning for convolutional neural networks: A core-set approach,” *arXiv preprint arXiv:1708.00489*, 2017.
- [28] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, “Deep batch active learning by diverse, uncertain gradient lower bounds,” *arXiv preprint arXiv:1906.03671*, 2019.
- [29] D. Ganguly, W. R. Morningstar, A. S. Yu, and V. Chaudhary, “Forte : Finding outliers with representation typicality estimation,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=7XNgVPxCiA>
- [30] T. Brooks, A. Holynski, and A. A. Efros, “Instructpix2pix: Learning to follow image editing instructions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18392–18402.
- [31] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3836–3847.
- [32] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, “Detecting twenty-thousand classes using image-level supervision,” in *European conference on computer vision*. Springer, 2022, pp. 350–368.
- [33] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-nms—improving object detection with one line of code,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5561–5569.
- [34] D. Zhou, J. Fang, X. Song, C. Guan, J. Yin, Y. Dai, and R. Yang, “Io loss for 2d/3d object detection,” in *2019 international conference on 3D vision (3DV)*. IEEE, 2019, pp. 85–94.
- [35] Y. Roh, G. Heo, and S. E. Whang, “A survey on data collection for machine learning: a big data-ai integration perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp. 1328–1347, 2019.
- [36] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [37] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “Cutmix: Regularization strategy to train strong classifiers with localizable features,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.
- [38] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation strategies from data,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.
- [39] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “RandAugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.
- [40] J.-N. Chen, S. Sun, J. He, P. H. Torr, A. Yuille, and S. Bai, “Transmix: Attend to mix for vision transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12135–12144.
- [41] Q. Zhao, Y. Huang, W. Hu, F. Zhang, and J. Liu, “Mixpro: Data augmentation with maskmix and progressive attention labeling for vision transformer,” *arXiv preprint arXiv:2304.12043*, 2023.
- [42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [43] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
- [44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [45] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov, “Effective data augmentation with diffusion models,” *arXiv preprint arXiv:2302.07944*, 2023.
- [46] M. F. Burg, F. Wenzel, D. Zietlow, M. Horn, O. Makansi, F. Locatello, and C. Russell, “Image retrieval outperforms diffusion models on data augmentation,” *arXiv preprint arXiv:2304.10253*, 2023.
- [47] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196.
- [48] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.
- [49] Z.-H. Zhou and M. Li, “Tri-training: Exploiting unlabeled data using three classifiers,” *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.
- [50] D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.
- [51] H. S. Seung, M. Opper, and H. Sompolinsky, “Query by committee,” in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.
- [52] Q. Bouinot, A. Loesch, R. Audigier, and A. Habrard, “Towards few-annotation learning for object detection: Are transformer-based models more efficient?” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 75–84.
- [53] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, “Snorkel: Rapid training data creation with weak supervision,” in *Proceedings of the VLDB endowment. International conference on very large data bases*, vol. 11, no. 3, 2017, p. 269.
- [54] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, “Solving the multiple instance problem with axis-parallel rectangles,” *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.
- [55] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [56] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, “Holoclean: Holistic data repairs with probabilistic inference,” *arXiv preprint arXiv:1702.00820*, 2017.
- [57] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg, “Active-clean: Interactive data cleaning for statistical modeling,” *Proceedings of the VLDB Endowment*, vol. 9, no. 12, pp. 948–959, 2016.
- [58] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, “Automating large-scale data quality verification,” *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, 2018.
- [59] A. Java, A. Khandelwal, S. Midigeshi, A. Halfaker, A. Deshpande, N. Goyal, A. Gupta, N. Natarajan, and A. Sharma, “Characterizing deep research: A benchmark and formal definition,” *arXiv preprint arXiv:2508.04183*, 2025.
- [60] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvassy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, “The faiss library,” 2024.
- [61] G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem, “Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 3732–3784, 2023.