AGEALBENCH: BENCHMARKING DEEP ACTIVE LEARNING FOR FACIAL AGE ESTIMATION WITH CROWD-SOURCED ANNOTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Age estimation from facial images is a problem of practical importance, with a variety of real-world applications. While deep neural networks have depicted commendable performance for this task, they require large amounts of hand-annotated training data, acquiring which is a time-consuming and labor-intensive process. Active Learning (AL) algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data, and tremendously reduce human annotation effort in training deep neural network models. However, facial age estimation poses a unique challenge as it is almost impossible to estimate the exact correct age of a person merely from a facial image; moreover, the perceived age is inherently subjective, and different human annotators can have very different perceptions about the age of the same subject. This necessitates a thorough analysis of the performance of AL techniques for facial age estimation, where each image is annotated by multiple annotators to capture annotation variability. In this research, we first collect crowd-sourced age annotations of two benchmark datasets: UTKFace and AAFD; each image is annotated by three annotators, who provide exact age estimates, as well as age range estimates. We then analyze the performance of 13 single annotator, as well as multi-annotator AL algorithms using the collected annotations. To the best of our knowledge, this is the first research effort to analyze the performance of deep AL algorithms (with crowd-sourced annotations) for the task of facial age estimation, where annotators can only provide an approximate label to the queried instances, rather than the exact label. We hope this research will be a step toward bridging the gap between AL and real-world applications, and our findings will be useful for the development of next generation AL algorithms under real-world annotation challenges. Our code and the collected annotations will be open-sourced upon acceptance of our paper.

1 INTRODUCTION

Facial age estimation from a single RGB image has applications in many computer vision tasks, such as, video surveillance, demographic analysis, customer profiling, human-computer interaction, age-aware content delivery, etc (Geronimo et al., 2009; Angulu et al., 2018; Pei et al., 2019). Given a facial image, the goal is to infer a person's age from visual appearance. Deep neural networks (DNNs) have depicted commendable performance for this application by learning discriminative age-related facial representations, ranging from convolutional and ordinal-regression models (Levi & Hassner, 2015; Rothe et al., 2015; Niu et al., 2016) to more recent transformer-based architectures (Shi et al., 2023; Chen et al., 2023; Qin et al., 2023; Liu et al., 2025). Alongside advances in model design, benchmark datasets such as APPA-REAL (Agustsson et al., 2017), MORPH-II (Ricanek & Tesafaye, 2006), CACD (Chen et al., 2014), and AFAD (Niu et al., 2016) have enabled systematic evaluation of facial age estimation.

Modern deep learning based facial age estimation models necessitate large amounts of labeled data (Paplhám et al., 2024; Liu et al., 2025) to be trained reliably. However, obtaining such data is an expensive process in terms of time, labor and human expertise. *Active Learning (AL)* algorithms automatically identify the salient and exemplar samples from large amounts of unlabeled data and tremendously reduce human annotation effort in training a DNN model Ren et al. (2021);

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

Settles (2009). Deep active learning has been successfully studied across computer vision (Sinha et al., 2019; Yoo & Kweon, 2019), natural language processing (Chakraborty & Singh, 2022; Zhang et al., 2022), medical diagnosis (Gorriz et al., 2017), and anomaly detection (Pimentel et al., 2020) among others, demonstrating its potential to reduce human labeling burden across diverse learning problems. Active learning is thus a promising solution in inducing a robust DNN model for facial age estimation with minimal human annotation effort.

However, such an application poses a unique challenge, as it is almost impossible to estimate the exact age of a person merely from a facial image. Further, the perceived age is inherently subjective, and different annotators can have noticeably different perceptions about the age of the same subject (Agustsson et al., 2017). Figure 1 depicts sample illustrations from two benchmark datasets UTK-Face (Zhang et al. (2017)) and AAFD (Cheng et al. (2019)), where for each image, we report the ground truth (GT) label from the dataset, the exact age estimates from three crowd-sourced human annotators and three state-of-the-art vision-language models (VLMs): Qwen2-VL Bai et al. (2025), InternVL2 Chen et al. (2025), LLaVA-1.6 Liu et al. (2023); we also report an age range estimated by the three human annotators and the VLMs. We note that both the exact age estimates and the range estimates vary quite significantly among the human and the VLM annotators, corroborating the difficulty of obtaining accurate annotations in such an application. This motivates the study of AL algorithms for facial age estimation where each image is annotated by multiple human annotators, to capture the inter-annotator variability.

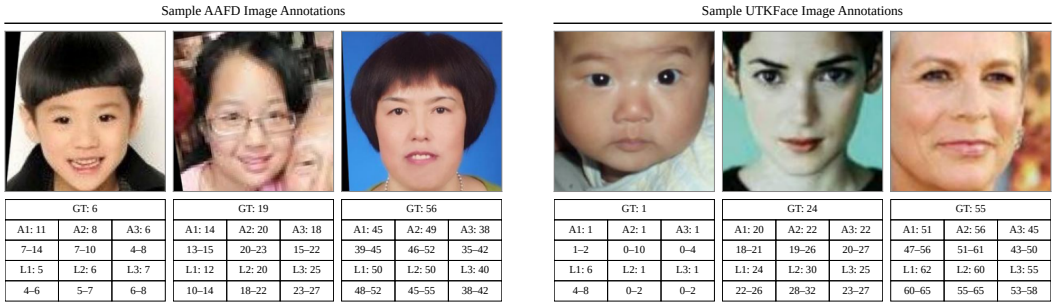


Figure 1: Annotation illustrations on AAFD (left) and UTKFace (right) datasets. The figure shows the ground truth age value for each image, the exact and range age estimates provided by the three human annotators A_1, A_2 and A_3 and the same estimates provided by three VLMs: L_1 (Qwen2-VL), L_2 (InternVL2) and L_3 (LLaVA-1.6).

Existing AL methods with multiple imperfect annotators are studied only for the classification setup Chakraborty (2020); Huang et al. (2017); Yan et al. (2016); their extension to a regression application (such as facial age estimation) is not straightforward. Modern pre-trained vision-language models (VLMs), such as Qwen2.5-VL (Bai et al., 2025), InternVL2.5 (Chen et al., 2025), and LLaVA (Liu et al., 2023), offer a scalable alternative by inferring labels directly from images. However, benchmarks like CDALBench (Werner et al., 2024) show that the relative rankings of AL methods can vary substantially across domains and experimental settings. A comprehensive study on the stability of AL methods with changes in the source of annotation is still unexplored. This motivates examining whether such rankings remain stable when the annotation source shifts from ground-truth labels to human annotations and VLM-generated labels in regression.

These observations necessitate a thorough validation of deep AL algorithms with crowd-sourced annotations, for facial age estimation, where annotators can only provide an approximate estimate of the label, rather than the exact label. In this paper, we attempt to address this relevant problem, which to the best of our knowledge, is unexplored in the literature. Our contributions in this work can be summarized as follows:

- We collect crowd-sourced annotations for two benchmark age estimation datasets through a crowd-sourcing platform. Each image is annotated by three annotators (to capture the inter-annotator variability), who provide an exact age estimate and a range estimate. These annotations will help in the development of deep AL algorithms for regression applications with multiple imperfect human annotators, which has not been studied in the literature. The

collected annotations will be made publicly available to promote further research on this important topic.

- We analyze the performance of 13 deep AL algorithms for regression, which include single annotator as well as multi-annotator methods, for vision-based facial age estimation. We observe significant shifts in the performance of the algorithms based on the source of annotations (human / VLM / GT), and with changes in the backbone architecture, fine-tuning technique, query batch size and initial training set size. Our analysis provides valuable insights into the performance of AL algorithms under real-world annotation constraints, which will hopefully be helpful for the development of next generation AL algorithms and their deployment in practical applications involving facial age estimation, such as security, surveillance and healthcare. We term our research *AgeALBench*, as it benchmarks the performance of AL algorithms for facial age estimation.

2 Related Work

We present a brief survey of active learning algorithms in general, followed by a survey of AL in the presence of noisy oracles, and AL for facial age estimation.

Active Learning: Active Learning (AL) is a well-studied problem in the machine learning literature Settles (2009). The main objective is to automatically select the salient and exemplar samples from large amounts of unlabeled data, to learn discriminative feature representations using a deep neural network Ren et al. (2021). The most popular query strategy is uncertainty sampling, where unlabeled samples with maximal predictive uncertainties are queried for manual annotation. Common strategies to compute uncertainty include predictive entropy Guo & Schuurmans (2007), disagreement among a committee of learners Seung et al. (1992), temporal output discrepancy Huang et al. (2021), Fisher information Ash et al. (2021), and mutual information Guo (2010) among others. Diversity based AL methods are also popular, where the goal is to select a diverse batch of samples, such as a coreset of samples that are representative of the entire data Sener & Savarese (2018), and sampling based on greedy distance maximization in the largest cluster Holzmüller et al. (2023). Methods combining uncertainty and diversity criteria have also been studied with promising results Chakraborty et al. (2015); Ash et al. (2019). Deep AL methods based on adversarial learning have particularly shown promising performance Sinha et al. (2019); Mayer & Timofte (2020); Zhu & Bento (2017); Ducoffe & Precioso (2018).

Active Learning with Imperfect Annotators: A branch of AL research has focused on the development of algorithms in the presence of noisy annotators, where the goal is to select informative samples, as well as the best annotators for labeling them Chen et al. (2022). Initial efforts in this direction have assumed a very simplistic setting, such as a single oracle which can provide incorrect labels and can also abstain from labeling Yan et al. (2016), or two labeling oracles, one of which always returns the correct label and the other returns incorrect labels with a fixed probability Zhang & Chaudhuri (2015); Donmez & Carbonell (2008). For multiple noisy annotators, a common approach is to relabel an actively queried sample multiple times, and derive the final label using majority voting Donmez et al. (2009); Zheng et al. (2010); Ipeirotis et al. (2014); Zhao et al. (2011). The *ActiveLab* algorithm Goh & Mueller (2023) is also based on a similar rationale, and balances between relabeling already labeled samples vs. labeling completely new samples. A cost effective AL framework (*CEAL*) was proposed by Huang *et al.* (Huang et al. (2017)) for active sample selection in the presence of multiple noisy oracles. An LP based formulation was proposed in (Chakraborty (2020)) to identify the informative unlabeled samples and the most optimal annotator to annotate each sample.

However, all these methods simulate the imperfect annotators using machine learning models, which may not be able to capture the complex nuances of human annotation behavior. Further, these methods have been studied in the context of classification problems, where the possible class options for a sample are provided in advance; regression problems (such as vision based facial age estimation) present a different challenge where it is almost impossible to estimate the exact label for a given sample. Yan *et al.* (Yan et al. (2011; 2012)) analyzed the performance of AL algorithms on scientific text data annotated by human experts. However, they studied only binary classification problems, and the study was conducted using simple ML models, as opposed to state-of-the-art

deep neural networks. Very recently, another analysis of AL algorithms for text classification using crowd-sourced annotations was presented in (Totakura et al. (2026)).

Active Learning for Facial Age Estimation: Although AL has been widely studied in computer vision, its application to facial age estimation is limited. Wang *et al.* (Wang et al. (2010; 2011)) proposed an active sampling mechanism for age estimation based on the furthest nearest neighbor (FNN) that generalizes the margin-based uncertainty to the multi-class case. An AL method for facial age estimation was proposed in (Singh & Chakraborty (2021)) which queried pairs of unlabeled samples, and the annotators provided relative labels (whether subject A is older / younger than subject B) rather than exact labels. Another AL technique was proposed in (Bhattacharya & Chakraborty (2022)) where the annotators provided an estimated range of the age of a subject, instead of exact age labels.

However, none of these methods conducted a thorough analysis of state-of-the-art AL methods for facial age estimation with crowd-sourced annotations. In this paper, we investigate this important research question with the goal of bridging the gap between AL and real-world annotation challenges, thereby advancing the practical applicability of AL methods.

3 Datasets, Annotations and Analysis

In this section, we provide an overview of the datasets used in our research, describe the annotation process, and provide statistical analysis of the human annotations, and VLM-generated age estimates.

Table 1: Summary of the obtained crowd-sourced annotations for the UTKFace and AAFD datasets

Source Dataset	# Images	Annotators/Image	Point Labels	Range Labels	Total Human Annotations
UTKFace	10,000	3	10,000	10,000	30,000
AAFD	6,637	3	6,637	6,637	19,911
Total	16,637	3	16,637	16,637	49,911

3.1 Data Annotation

Datasets: We used two popular age-estimation benchmarks in this research, UTKFace (Zhang et al. (2017)) and AAFD (Cheng et al. (2019)), which contain images covering large variations in pose, facial expression, illumination, occlusion, resolution, etc.

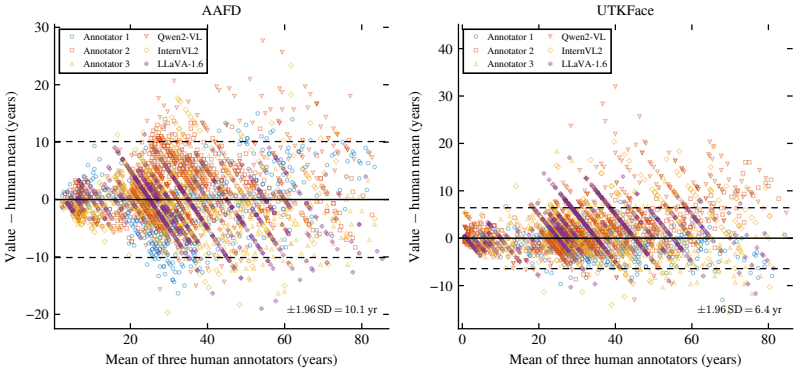
Annotator Recruitment: We used the Upwork crowd-sourcing platform (https://www.upwork.com/) to recruit human annotators for our task. The job was made publicly available to eligible workers worldwide, with the condition that they be comfortable reading and understanding English, since the instructions were all written in English. Consent was obtained from each worker to participate in the study. No information about the identity of the workers was recorded.

Annotation Protocol: We selected $10,000$ images from UTKFace and $6,637$ images from AAFD at random. The workers were asked to provide two annotations for each image: (i) an estimated exact age of the person (point label); (ii) an estimated age range within a maximum span of 10 years (range label). They were instructed to exercise their best judgment in providing the annotations and not to use any kind of AI models for the same. Each worker was given 30 days' time to submit their annotations. Once a job was submitted by a worker, the research team manually reviewed the file for errors (in particular, missed annotations); in case of any issues, the worker was contacted to fix the error. Upon successful completion of a job, each worker was compensated for their efforts. Each image from each dataset was annotated by 3 different crowd-sourced workers. Table 1 depicts a summary of the obtained annotations. Figure 1 shows sample annotation illustrations from the two datasets, provided by the three human annotators A_1, A_2 and A_3 and three VLMs: Qwen2-VL, InternVL2 and L3 LLaVA-1.6. *The collected annotations will be made publicly available to promote further research on this important topic.*

3.2 Annotation Analysis

3.2.1 Inter-annotator agreement for point age estimates

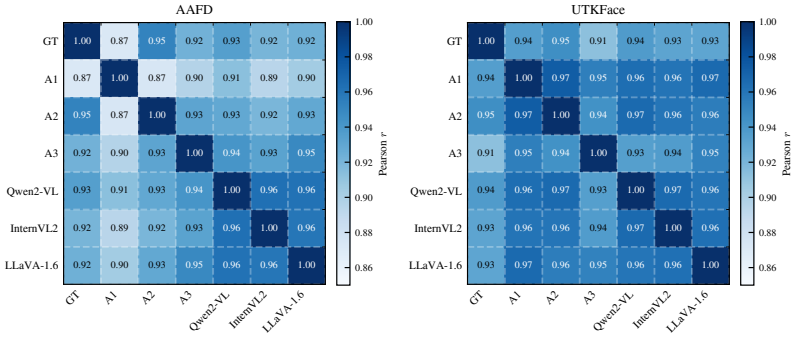
We first analyze the agreement among the three human annotators (A_1, A_2, A_3) and three vision-language models (VLMs): Qwen2-VL Bai et al. (2025), InternVL2 Chen et al. (2025) and LLaVA-1.6 Liu et al. (2023) for the point age estimates. In Figure 2, the x-axis denotes the mean age ($\bar{a}$) provided by the three human annotators for a given sample; the y-axis denotes the difference between the estimate provided by one of the annotators a_i (human / VLM) and the mean age estimated by the three human annotators ($a_i - \bar{a}$). We note that the annotations are more consistent for UTKFace (where the $a_i - \bar{a}$ difference is 6.4 years for 95% of the samples) than AAFD (where the difference is 10.1 years for 95% of the samples). Also, for both datasets, annotations are more consistent for lower values of the age (< 20 years) and diverge for higher values. This shows that it is probably easier to estimate the age of younger subjects than older subjects.



(a) AAFD (b) UTKFace

Figure 2: Annotator agreement for point age estimates. The x-axis denotes the mean age ($\bar{a}$) provided by the three human annotators for a given sample; the y-axis denotes the difference between the estimate provided by each annotator a_i (human / VLM) and the mean age estimated by the three human annotators ($a_i - \bar{a}$).

We also compute the mutual agreement between the point age estimates provided by the annotators using the Pearson's correlation coefficient. Figure 3 shows the pairwise correlation heatmap for the three human annotators A_1, A_2, A_3 and the three pre-trained VLMs, together with the ground truth. We observe consistently strong correlations across annotators, but with noticeable differences between the datasets. In AAFD, correlations among the three human annotators range from 0.87 to 0.93, indicating greater variability in the perceived age. In contrast, UTKFace exhibits stronger human agreement, with correlations ranging from 0.94 to 0.97. Agreement with ground-truth age also varies across annotators, reaching 0.87–0.95 on AAFD and 0.91–0.95 on UTKFace. These results show that annotator agreement is high overall, yet remains dataset and annotator-dependent.



(a) AAFD (b) UTKFace

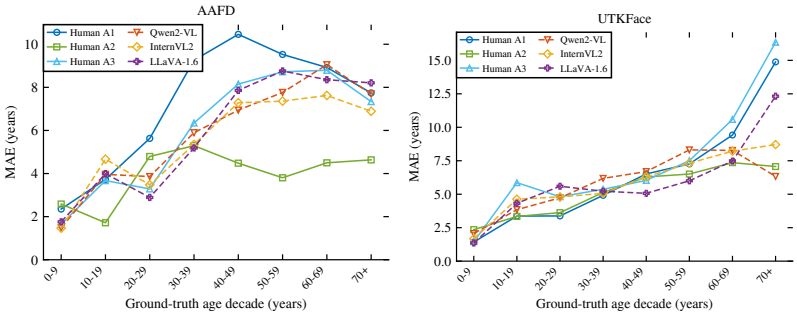
Figure 3: Pearson's correlation between ground-truth ages (GT), human annotators (A_1, A_2, A_3), and VLM annotators. Higher correlation indicates stronger agreement in age estimates between annotators.

3.2.2 Estimation error for point and range age estimates

We now analyze the error in estimation for the point and range age estimates with respect to the ground truth (GT) age. We first partition the samples into of 10-year age groups and compute the mean absolute error (MAE) for the three human annotators and the three VLMs separately for each group. As shown in Figure 4 (a–b), the annotation error varies substantially across age groups, annotators, and datasets. We note that the estimation error remains moderately low for younger

subjects ($0-29$ years) for both datasets. In AAFD, the annotators exhibit non-uniform error patterns, with larger errors for middle-aged and older subjects.

In UTKFace, errors generally increase with age, with the largest errors observed in the 60–69 and 70+ age groups for some annotators. These results suggest that human age estimation errors exhibit heteroscedasticity, rather than being well modeled as uniform annotation noise. We also note that human annotators can often furnish better age estimates than pre-trained VLMs (such as annotator A_2 for both the datasets). Thus, it may not be a good idea to replace human annotators with pre-trained VLMs for data annotation in AL. We further note that no VLM annotator depicts the best performance consistently across all age groups. For instance, in AAFD, LLaVA depicts better performance than InternVL2 for age groups $10-19$ and $20-29$; however, InternVL2 outperforms LLaVA for age groups $40-49$ and higher.



(a) AAFD: MAE by age groups (b) UTKFace: MAE by age groups

Figure 4: MAE of Annotators and VLMs at different age groups for UTKFace and AAFD.

We analyze range annotations using two measures: *range coverage*, defined as the percentage of times the ground-truth age falls within the predicted age range given by the annotator (denoted by the vertical bars in Figure 5), and *midpoint MAE*, computed as the MAE between the midpoint of each predicted range and the ground-truth age, averaged across all samples (denoted by the red curve (axis on the right) in Figure 5). From Figure 5, we note that humans are often better than VLMs in predicting the age range of a subject. For instance, for UTKFace, we note that the age range predicted by A_2 substantially outperforms the ranges predicted by all the VLMs. The midpoint MAE errors also vary substantially across human and VLM annotators. However, better covering ranges do not necessarily imply more accurate central age estimates. For instance, for UTKFace, human annotator A_3 has a higher midpoint MAE than Qwen2-VL, although it has a more accurate range coverage (46%) than Qwen2-VL (33%).

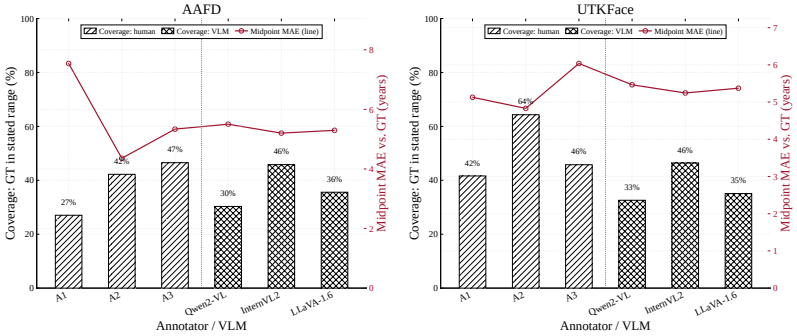


Figure 5: Range annotation analysis for human annotators ($A_1 - A_3$) and VLM annotators on AAFD (left) and UTKFace (right).

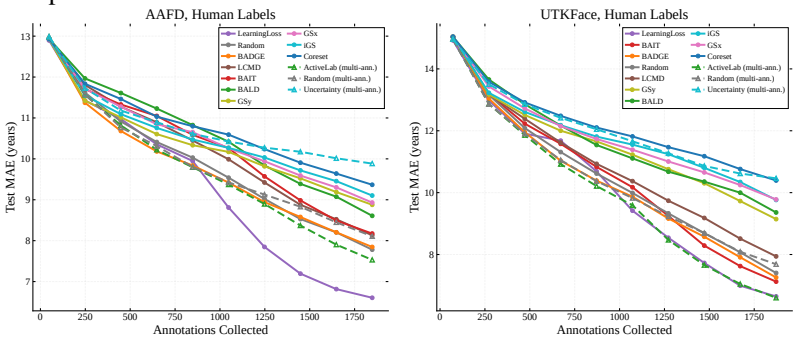
4 Experiments and Results

We provide a detailed explanation of the Experimental Setup in Appendix A.3, Implementation Details in Appendix A.4, and 13 AL algorithms we studied in this research, including single annotator as well as multi-annotator methods in Appendix A.5.

4.1 Active Learning Performance: Crowd-sourced Annotations

The active learning performance using crowd-sourced annotations for single and multi-annotator settings is depicted in Figure 6. In each graph, the x-axis denotes the number of unlabeled samples annotated, and the y-axis denotes the MAE on the test set. On UTKFace, *LearningLoss* and *ActiveLab* outperform the other baselines, while on AAFD, *LearningLoss* achieves the lowest MAE in the later AL rounds by a significant margin, with *ActiveLab* depicting the second best perfor-

mance. Across both datasets, other AL strategies lag, remain close to, or only modestly outperform random sampling. These results indicate that existing AL methods are not robust to crowd-sourced supervision, with substantial performance variations across datasets.



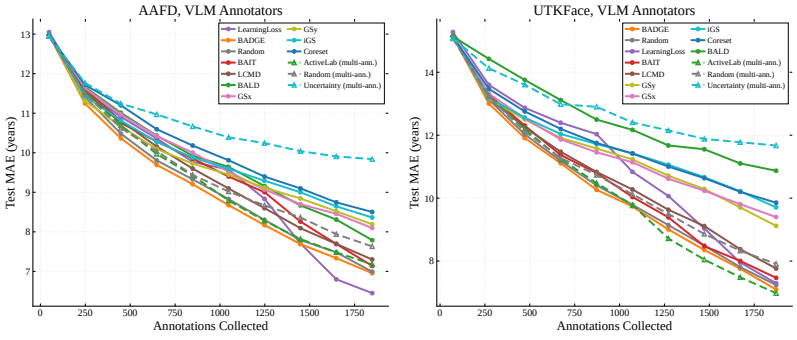
(a) AAFD (b) UTKFace

Figure 6: AL Performance using crowd-sourced annotations on UTKFace and AAFD under single and multi-annotator settings. Best viewed in color.

Further, while multi-annotator methods like *ActiveLab* and *Random (multi-ann)* outperform most of the methods, uncertainty sampling with multiple annotators depicts much worse performance. Thus, despite access to richer supervision, AL methods based on multiple annotators do not consistently outperform conventional single-annotation strategies for this application. The results suggest that current multi-annotator AL methods fail to fully exploit annotator diversity and may even be adversely affected by disagreement and annotation noise. These findings motivate the development of AL methods specifically designed for crowd-sourced, multi-annotator regression applications.

We also assess whether performance differences between AL methods are statistically significant. Since the goal is to achieve low test MAE, we use the area under the learning curve (AULC) of the MAE curve (averaged across all the runs) to compute the statistical significance. Further, since we would like to test the hypothesis in both directions (method A better than method B, and vice versa), we use a one-sided paired t-test on the AULC values. Table 2 reports pairwise p-values, where the entry in row i and column j corresponds to the hypothesis that the method in the row is statistically significantly better than the method in the column. On UTKFace, *LearningLoss* shows statistically significant performance improvement ($p < 0.05$) against all other AL methods, whereas on AAFD, *ActiveLab* exhibits significant performance relative to the other baselines. For the remaining method pairs, significance varies across datasets without a consistent pattern. These results further highlight the instability of existing AL methods under crowd-sourced, multi-annotator regression settings.

4.2 Active Learning Performance: VLM Annotations



(a) AAFD (b) UTKFace

Figure 7: AL performance using VLM-generated annotations. Best viewed in color.

In Figure 7, we show the AL performance using VLM-generated annotations under single and multi-annotator settings. *ActiveLab* achieves the lowest MAE in the later AL rounds on UTKFace. On AAFD, *BADGE* attains the lowest MAE in the first three quarters AL round of, but, the performance of *LearningLoss* is the best in the last quarter. Notably, the uncertainty-based multi-annotator AL strategy performs worst on both datasets. This suggests that conventional uncertainty estimates may become unreliable when supervision is distributed across multiple, and potentially disagreeing, annotators. To quantify the statistical significance of the observed performance differences, we conducted on-sided paired t-tests between AL methods and report the results in Appendix A.1.1, Table 4. Overall, these findings highlight that the performance of AL methods differs from one dataset to another when using VLMs as annotators.

Table 2: Pairwise p-values corresponding to one-sided t-test. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column.

(a) AAFD

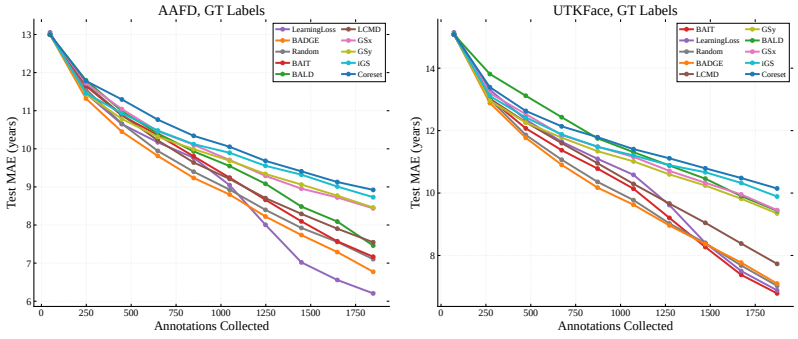
	LL	MA-ALB	BDG	RND	MA-RND	LCM	BAIT	GSY	IGS	GSX	BLD	CST	MA-UNC
LL	–	2.53e-05	7.23e-06	1.77e-06	6.10e-06	3.51e-08	1.44e-06	7.53e-09	3.00e-09	1.20e-09	2.98e-05	2.79e-10	5.31e-08
MA-ALB	1.00e+00	–	1.09e-01	1.59e-03	2.41e-03	2.54e-09	1.19e-04	8.71e-11	8.52e-12	1.17e-10	5.01e-04	2.14e-10	4.18e-07
BDG	1.00e+00	8.91e-01	–	2.41e-03	8.08e-04	1.72e-09	1.16e-04	8.27e-10	1.09e-10	4.39e-10	7.67e-04	2.73e-10	2.29e-07
RND	1.00e+00	9.98e-01	9.98e-01	–	2.83e-01	3.44e-06	5.98e-04	3.09e-08	5.84e-09	3.14e-08	1.16e-03	3.37e-09	4.85e-07
MA-RND	1.00e+00	9.98e-01	9.99e-01	7.17e-01	–	1.34e-06	1.74e-03	2.53e-07	2.05e-08	2.38e-08	2.91e-03	8.00e-10	1.53e-06
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	1.96e-01	4.97e-05	4.47e-08	4.80e-08	3.91e-02	1.28e-08	3.73e-05
BAIT	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.98e-01	8.04e-01	–	3.14e-01	2.89e-02	3.12e-02	5.09e-02	9.64e-04	1.35e-03
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.86e-01	–	1.92e-07	9.21e-05	1.14e-01	2.33e-07	3.35e-04
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.71e-01	1.00e+00	–	2.39e-01	3.32e-01	2.26e-06	5.30e-03
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.69e-01	1.00e+00	7.61e-01	–	3.64e-01	1.66e-07	9.58e-03
BLD	1.00e+00	9.99e-01	9.99e-01	9.99e-01	9.97e-01	9.61e-01	9.49e-01	8.86e-01	6.68e-01	6.36e-01	–	2.07e-01	1.88e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	1.00e+00	7.93e-01	–	4.60e-01
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	9.95e-01	9.90e-01	8.12e-01	5.40e-01	–

(b) UTKFace

	MA-ALB	LL	BDG	MA-RND	BAIT	RND	LCM	GSY	BLD	GSX	IGS	MA-UNC	CST
MA-ALB	–	7.64e-04	7.38e-06	7.19e-07	2.72e-04	4.60e-08	6.81e-10	4.71e-11	2.70e-04	2.01e-11	4.95e-11	5.35e-05	9.37e-13
LL	9.99e-01	–	1.37e-03	4.29e-04	2.69e-03	4.26e-05	1.98e-07	2.57e-09	4.73e-04	1.06e-09	9.46e-10	6.66e-05	1.21e-10
BDG	1.00e+00	9.99e-01	–	2.77e-01	1.49e-01	1.26e-03	2.28e-08	4.52e-12	2.84e-03	1.78e-10	5.20e-12	5.96e-04	2.50e-12
MA-RND	1.00e+00	1.00e+00	7.23e-01	–	2.29e-01	1.78e-03	2.63e-07	2.86e-10	2.48e-03	6.89e-11	3.96e-11	5.06e-04	1.34e-13
BAIT	1.00e+00	9.97e-01	8.51e-01	7.71e-01	–	2.62e-01	7.70e-04	1.82e-06	4.10e-03	2.80e-08	1.05e-07	8.65e-04	1.51e-08
RND	1.00e+00	1.00e+00	9.99e-01	9.98e-01	7.38e-01	–	1.14e-06	1.10e-09	3.93e-03	6.02e-11	5.26e-10	7.95e-04	3.84e-13
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	–	3.30e-09	2.45e-02	2.01e-10	1.63e-10	3.97e-03	3.67e-12
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	3.68e-01	2.69e-04	2.98e-07	9.75e-02	2.34e-09
BLD	1.00e+00	1.00e+00	9.97e-01	9.98e-01	9.96e-01	9.96e-01	9.75e-01	6.32e-01	–	3.59e-01	3.21e-01	4.60e-02	9.45e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.41e-01	–	1.66e-01	2.51e-01	2.78e-07
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.79e-01	8.34e-01	–	3.01e-01	2.23e-06
MA-UNC	1.00e+00	1.00e+00	9.99e-01	9.99e-01	9.99e-01	9.99e-01	9.96e-01	9.02e-01	9.54e-01	7.49e-01	6.99e-01	–	3.71e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.06e-01	1.00e+00	1.00e+00	6.29e-01	–

4.3 Active Learning Performance: Ground Truth (GT) Annotations

To establish an idealized reference for AL performance, we further evaluate all AL methods using the ground-truth age labels as annotations. The multi-annotator setting is not applicable in this case as the oracle always returns the true label. Figure 8 depicts that *LearningLoss, BADGE* and *BAIT* depict impressive performance in both the datasets. LearningLoss and BAIT respectively attain the lowest MAE in UTKFace and AAFD at the end of all the AL iterations. *Coreset* and *iGS* depict the worst performance in both the datasets. To quantify the statistical significance of the observed performance differences, we conducted one-sided paired t-tests between the AL methods and report the results in Appendix A.1.2, Table 5.



(a) AAFD (b) UTKFace

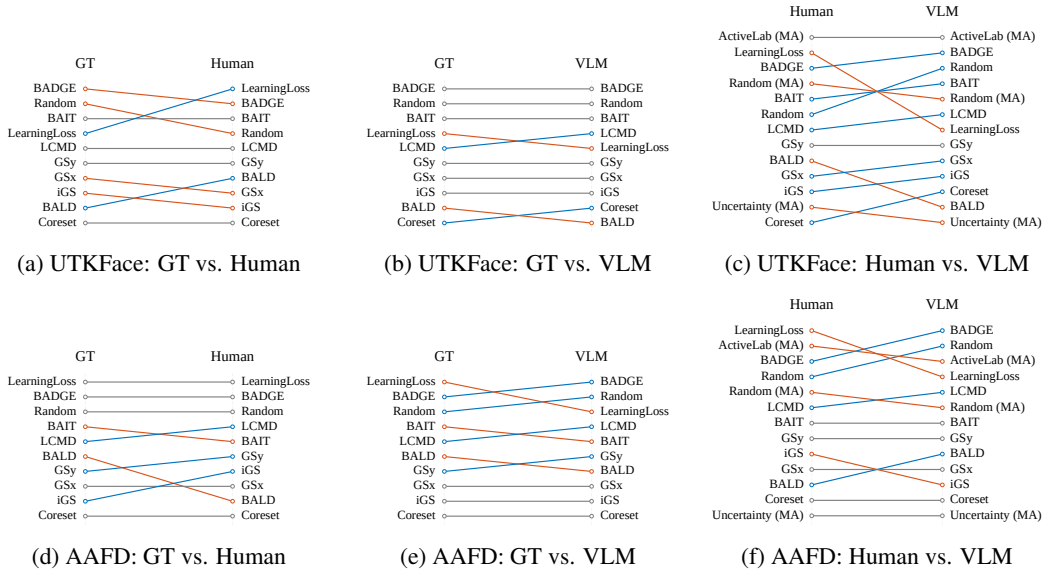
Figure 8: AL performance using GT annotations. Best viewed in color.

4.4 Rank Consistency Analysis

In Table 3, we compare the ranking consistency of AL methods across ground-truth (GT), human, and VLM-generated annotations. *Moved* denotes the number of methods (as a percentage of the total number of methods) whose rankings change due to the change in the source of annotation. This parameter depicts how much a method's rank changes due to a change in the source of annotation. *Friedman p (A/B)* reports the Friedman-test p-values for statistical test of significance of the obtained rankings of the AL methods for a given source of annotation A/B (as depicted in Figure 9). We

Table 3: Rank consistency analysis across annotation sources. n denotes the number of baselines ($n = 10$ for GT as the multi-annotator methods are not applicable here, since the oracle always returns the ground truth).

Dataset	A vs. B	n	Moved	Friedman p (A / B)
AAFD	Human vs. VLM	13	8/13 (62%)	1.7e-16 / 1.6e-16
AAFD	GT vs. Human	10	5/10 (50%)	2.4e-12 / 4.9e-12
AAFD	GT vs. VLM	10	7/10 (70%)	2.4e-12 / 2.1e-12
UTKFace	Human vs. VLM	13	11/13 (85%)	7.9e-17 / 1.0e-16
UTKFace	GT vs. Human	10	6/10 (60%)	4.7e-12 / 3.6e-13
UTKFace	GT vs. VLM	10	4/10 (40%)	4.7e-12 / 4.7e-13



(a) UTKFace: GT vs. Human (b) UTKFace: GT vs. VLM (c) UTKFace: Human vs. VLM

(d) AAFD: GT vs. Human (e) AAFD: GT vs. VLM (f) AAFD: Human vs. VLM

Figure 9: Rank shifts of active learning methods across annotation sources on UTKFace and AAFD.

note that the Friedman p-values are all less than 0.05, implying that the rankings of the AL methods obtained for each source of annotation (human / VLM / GT) are statistically significant, and are not affected by noise. Such rankings can thus be used for analysis, and to draw meaningful conclusions. On AAFD, at least $5/10(50\%)$ of methods change rank across annotation sources, with the highest change of $8/13(62\%)$ for *Human vs. VLM* comparison. On UTKFace, the largest shift occurs for *Human vs. VLM*, where $11/13(85\%)$ methods change rank. Figure 9 illustrates the ranking shifts of AL methods across different annotation sources for both the datasets. We observe notable fluctuations in the relative ordering as the source of supervision changes. Overall, these results reveal that the relative rankings of AL methods can change when supervision shifts from curated ground truth to crowd-sourced or VLM-generated annotations.

5 Conclusion

In this paper, we introduced *AgeALBench*, a benchmark for deep active learning for vision-based facial age estimation using crowd-sourced annotations. Using this benchmark, we systematically evaluated 13 AL strategies under ground-truth, human, and VLM-generated supervision. Our extensive evaluation revealed three key findings: (i) existing AL methods exhibit limited robustness to crowd-sourced supervision; (ii) current multi-annotator AL methods do not consistently outperform conventional single-annotator strategies for facial age estimation; and (iii) the relative rankings of AL methods can change significantly with the annotation source. We hope the insights gained from this research will contribute to the development of next generation AL techniques for a variety of real-world applications involving facial age estimation, such as security and surveillance, robotics, augmented reality and healthcare among others.

AI USE STATEMENT

Large Language Models were used to polish the writing of the manuscript and improve overall readability. Research design, experiments, analysis, and interpretation were fully carried out and decided by the authors.

REFERENCES

Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 87–94. IEEE, 2017.

Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, 2018.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8927–8939, 2021.

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.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Aditya Bhattacharya and Shayok Chakraborty. Deep active learning with range feedback for facial age estimation. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2022.

Shayok Chakraborty. Asking the right questions to the right users: Active learning with imperfect oracles. In *AAAI Conference on Artificial Intelligence*, 2020.

Shayok Chakraborty and Ankita Singh. Active sampling for text classification with subinstance level queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6150–6158, 2022.

Shayok Chakraborty, Vineeth Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(10):1945–1958, 2015.

Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pp. 768–783. Springer, 2014.

Ping Chen, Xingpeng Zhang, Ye Li, Ju Tao, Bin Xiao, Bing Wang, and Zongjie Jiang. DAA: A delta age AdaIN operation for age estimation via binary code transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15836–15845, 2023.

Yifang Chen, Karthik Sankararaman, Alessandro Lazaric, Matteo Pirotta, Dmytro Karamshuk, Qifan Wang, Karishma Mandyam, Sinong Wang, and Han Fang. Improved adaptive algorithm for scalable active learning with weak labeler. In *arXiv:2211.02233v1*, 2022.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019.

Pinar Donmez and Jaime Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *ACM Conference on Information and Knowledge Management (CIKM)*, 2008.

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. In *International Conference on Machine Learning (ICML)*, 2018.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *nternational Conference on Machine Learning (ICML)*, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*, pp. 1183–1192. PMLR, 2017.

David Geronimo, Antonio M Lopez, Angel D Sappa, and Thorsten Graf. Survey of pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(7):1239–1258, 2009.

Hui Wen Goh and Jonas Mueller. ActiveLab: Active learning with re-labeling by multiple annotators. *ICLR Workshop on Trustworthy ML*, 2023.

Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giro-i Nieto. Cost-effective active learning for melanoma segmentation. *arXiv preprint arXiv:1711.09168*, 2017.

Yuhong Guo. Active instance sampling via matrix partition. *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

Yuhong Guo and Dale Schuurmans. Discriminative batch mode active learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 20, 2007.

David Holzmüller, Viktor Zaverkin, Johannes Kästner, and Ingo Steinwart. A framework and benchmark for deep batch active learning for regression. *Journal of Machine Learning Research (JMLR)*, 24(164):1–81, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *https://arxiv.org/abs/2106.09685*, 2021.

Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. Cost-effective active learning from diverse labelers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

Panagiotis Ipeirotis, Foster Provost, Victor Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28, 2014.

Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–42, 2015.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023.

Xueli Liu, Mingyan Qiu, Ziqun Zhang, Yuxuan Shi, Zhen Li, Xiao Chen, Yinlong Liu, Xinrong Chen, and Hongmeng Yu. Enhancing facial age estimation with local and global multi-attention mechanisms. *Pattern Recognition Letters*, 189:71–77, 2025.

Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 3060–3068, 2020.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4920–4928, 2016.

Jakub Paplhám, Vojt Franc, et al. A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1196–1205, 2024.

Wenjie Pei, Hamdi Dibeklioğlu, Tadas Baltrušaitis, and David MJ Tax. Attended end-to-end architecture for age estimation from facial expression videos. *IEEE Transactions on Image Processing (TIP)*, 29:1972–1984, 2019.

Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 341–345, 2006.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV-W)*, pp. 10–15, 2015.

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

Burr Settles. Active learning literature survey. In *Technical Report, Department of Computer Science, University of Wisconsin-Madison*, 2009.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational Learning Theory (COLT)*, pp. 287–294, 1992.

Chaojun Shi, Shiwei Zhao, Ke Zhang, Yibo Wang, and Longping Liang. Face-based age estimation using improved swin transformer with attention-based convolution. *Frontiers in Neuroscience*, 17:1136934, 2023.

Ankita Singh and Shayok Chakraborty. Deep active learning with relative label feedback: An application to facial age estimation. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2021.

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

Varun Totakura, Ankita Singh, Yushun Dong, and Shayok Chakraborty. An analysis of active learning algorithms using real-world crowd-sourced text annotations. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2026.

Jian-Gang Wang, Eric Sung, and Wei-Yun Yau. Active learning with the furthest nearest neighbor criterion for facial age estimation. In *Asian Conference on Computer Vision (ACCV)*, 2010.

Jian-Gang Wang, Eric Sung, and Wei-Yun Yau. Active learning for solving the incomplete data problem in facial age classification by the furthest nearest-neighbor criterion. *IEEE Transactions on Image Processing (TIP)*, 20(7):2049 – 2062, 2011.

Thorben Werner, Johannes Burchert, Maximilian Stubbemann, and Lars Schmidt-Thieme. A cross-domain benchmark for active learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:62875–62911, 2024.

Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. In *Neural Information Processing Systems (NeurIPS)*, 2016.

Yan Yan, Romer Rosales, Glenn Fung, and Jennifer Dy. Active learning from crowds. In *International Conference on Machine Learning (ICML)*, 2011.

Yan Yan, Romer Rosales, Glenn Fung, Faisal Farooq, Bharat Rao, and Jennifer Dy. Active learning from multiple knowledge sources. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

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

Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In *Neural Information Processing Systems (NeurIPS)*, 2015.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6166–6190, 2022.

Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *International Conference on Social Computing*, 2011.

Yaling Zheng, Stephen Scott, and Kun Deng. Active learning from multiple noisy labelers with varied costs. In *IEEE International Conference on Data Mining (ICDM)*, 2010.

Jia-Jie Zhu and José Bento. Generative adversarial active learning. In *arXiv:1702.07956*, 2017.

A APPENDIX

TABLE OF CONTENTS

A.1 AL Performance: Statistical test of significance (main paper contd.) 15
A.1.1 VLMs as Annotators 15
A.1.2 Ground Truth as Annotation Source 15
A.2 AL Performance based on AULC and final test MAE 15
A.2.1 Crowd-sourced Human Annotators 15
A.2.2 VLM Annotators 16
A.2.3 GT Annotators 17
A.3 Experimental Setup 17
A.4 Implementation Details 18
A.5 Active Learning Algorithms and Adaptations 19
A.5.1 Single-annotator AL Methods 19
A.5.2 Multi-annotator AL Methods 20
A.6 Ablation Studies 20
A.6.1 Effect of Initial Labeled Set Size 20
A.6.2 Effect of Query Size 20
A.6.3 Effect of Fine-Tuning Strategy 22
A.6.4 Effect of Backbone Architecture 24
A.7 VLM Prompt 24
A.8 Limitations and Future Work 25

A.1 AL PERFORMANCE: STATISTICAL TEST OF SIGNIFICANCE (MAIN PAPER CONTD.)

A.1.1 VLMS AS ANNOTATORS

Table 4 presents the pairwise p-values corresponding to one-sided t-test for the experiment with VLM-generated annotations. On AAFD, *BADGE* achieves statistically significant improvements ($p < 0.05$) over all other AL methods. On UTKFace, *ActiveLab* and *BADGE* form the strongest group, with no statistically significant difference between them ($p > 0.05$), while both significantly outperform the remaining methods. For the other method pairs, statistical significance varies across datasets without a consistent pattern. These results further demonstrate that the relative effectiveness of AL methods depends strongly on the dataset, even under the same VLM-based annotation setting.

Table 4: Pairwise p-values corresponding to one-sided t-test for the experiment with VLMs as annotators. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column. The AL performance results are depicted in Figure 7 in the main paper.

(a) AAFD

	BDG	RND	MA-ALB	LL	LCM	MA-RND	BAIT	GSY	BLD	GSX	IGS	CST	MA-UNC
BDG	–	1.78e-05	4.57e-06	5.60e-04	6.06e-09	5.70e-08	5.22e-04	3.93e-12	1.40e-04	9.58e-12	2.28e-13	2.11e-11	1.26e-06
RND	1.00e+00	–	7.71e-03	8.97e-03	5.13e-06	1.27e-05	5.00e-03	1.25e-10	3.75e-04	2.27e-09	2.59e-11	1.28e-09	1.81e-06
MA-ALB	1.00e+00	9.92e-01	–	4.39e-02	6.94e-05	2.51e-04	1.70e-02	1.48e-09	9.10e-04	1.30e-09	4.69e-10	4.67e-09	5.71e-06
LL	9.99e-01	9.91e-01	9.56e-01	–	2.28e-01	1.75e-01	1.65e-01	5.03e-04	4.99e-03	2.67e-05	2.88e-05	6.31e-06	1.21e-05
LCM	1.00e+00	1.00e+00	1.00e+00	7.72e-01	–	3.59e-01	2.82e-01	2.64e-06	8.82e-03	1.76e-07	1.37e-07	1.07e-08	1.40e-05
MA-RND	1.00e+00	1.00e+00	1.00e+00	8.25e-01	6.41e-01	–	3.08e-01	3.31e-05	1.01e-02	3.53e-07	2.78e-07	1.96e-08	8.17e-06
BAIT	9.99e-01	9.95e-01	9.83e-01	8.35e-01	7.18e-01	6.92e-01	–	1.67e-02	6.09e-02	1.99e-03	1.69e-03	2.84e-05	4.34e-05
GSY	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	9.83e-01	–	2.48e-01	2.62e-04	3.73e-06	5.39e-07	1.29e-04
BLD	1.00e+00	1.00e+00	9.99e-01	9.95e-01	9.91e-01	9.90e-01	9.39e-01	7.52e-01	–	4.65e-01	3.76e-01	6.72e-02	9.48e-05
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.98e-01	1.00e+00	5.35e-01	–	7.34e-02	3.63e-06	4.14e-04
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.98e-01	1.00e+00	6.24e-01	9.27e-01	–	3.68e-05	3.46e-04
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.33e-01	1.00e+00	1.00e+00	–	3.18e-03
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.97e-01	–

(b) UTKFace

	MA-ALB	BDG	RND	BAIT	MA-RND	LCM	LL	GSY	GSX	IGS	CST	BLD	MA-UNC
MA-ALB	–	4.72e-01	4.64e-02	1.26e-03	2.41e-05	3.20e-05	3.79e-04	2.13e-09	4.53e-10	4.29e-10	1.26e-10	4.32e-05	6.54e-06
BDG	5.28e-01	–	9.26e-03	2.85e-03	5.99e-07	2.24e-06	2.91e-04	9.55e-11	2.67e-11	7.20e-11	9.35e-12	3.60e-05	6.99e-06
RND	9.54e-01	9.91e-01	–	4.77e-02	4.54e-06	1.46e-08	1.20e-03	1.02e-12	7.45e-13	6.72e-13	1.92e-13	6.00e-05	1.47e-05
BAIT	9.99e-01	9.97e-01	9.52e-01	–	2.66e-02	6.86e-03	5.40e-03	1.74e-07	1.05e-07	7.09e-09	3.94e-09	1.12e-04	1.55e-05
MA-RND	1.00e+00	1.00e+00	1.00e+00	9.73e-01	–	6.38e-03	9.13e-03	9.91e-10	9.32e-10	3.14e-10	1.28e-10	1.49e-04	3.61e-05
LCM	1.00e+00	1.00e+00	1.00e+00	9.93e-01	9.94e-01	–	3.24e-02	4.67e-11	1.33e-09	2.97e-11	3.33e-11	2.37e-04	5.98e-05
LL	1.00e+00	1.00e+00	9.99e-01	9.95e-01	9.91e-01	9.68e-01	–	4.38e-02	4.36e-02	7.34e-03	3.62e-03	3.24e-04	7.54e-05
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.56e-01	–	4.43e-01	6.98e-07	5.22e-08	4.75e-03	9.53e-04
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.56e-01	5.57e-01	–	1.32e-05	3.02e-06	5.30e-03	1.15e-03
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.93e-01	1.00e+00	1.00e+00	–	8.00e-04	1.55e-02	3.18e-03
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.96e-01	1.00e+00	1.00e+00	9.99e-01	–	2.16e-02	4.19e-03
BLD	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.95e-01	9.95e-01	9.84e-01	9.78e-01	–	1.63e-01
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.99e-01	9.97e-01	9.96e-01	8.37e-01	–

A.1.2 GROUND TRUTH AS ANNOTATION SOURCE

In Table 5, we report the pairwise p-values corresponding to one-sided t-test for the experiment with ground truth annotations. On AAFD, *LearningLoss* achieves the best overall performance and significantly outperforms all methods except *BADGE* ($p = 0.218$), while *BADGE* performs significantly better than the remaining baselines. On UTKFace, *BADGE* achieves the strongest performance and shows statistically significant improvements ($p < 0.05$) over all methods except *BAIT* ($p = 0.0623$). These results show that even under clean ground-truth supervision, the statistically strongest AL methods differ across datasets, further highlighting the dataset-dependent behavior of existing AL strategies.

A.2 AL PERFORMANCE BASED ON AULC AND FINAL TEST MAE

A.2.1 CROWD-SOURCED HUMAN ANNOTATORS

Table 6 reports the mean area under the learning curve (AULC) and final test MAE, averaged across random seeds, for all 13 AL methods with crowd-sourced human annotations (methods are ranked from low to high AULC). On AAFD, *LearningLoss* achieves the best overall performance with the lowest mean AULC (9.24 ± 0.18) and final MAE (6.60 ± 0.14), followed by *ActiveLab* and *BADGE*.

Table 5: Pairwise p-values corresponding to one-sided t-test for the experiment with ground truth (GT) annotations. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column. The AL performance results are depicted in Figure 8 in the main paper.

(a) AAFD

	LL	BDG	RND	BAIT	LCM	BLD	GSY	GSX	IGS	CST
LL	–	2.18e-01	1.82e-03	2.96e-03	4.39e-05	1.56e-03	3.92e-07	1.59e-07	9.41e-08	4.17e-08
BDG	7.82e-01	–	1.28e-05	2.00e-03	1.69e-08	1.62e-03	1.06e-11	1.11e-11	6.37e-12	8.14e-11
RND	9.98e-01	1.00e+00	–	4.28e-02	3.66e-06	1.46e-02	2.07e-10	5.12e-10	1.01e-10	5.90e-10
BAIT	9.97e-01	9.98e-01	9.57e-01	–	1.33e-01	8.57e-02	2.18e-04	1.50e-04	2.18e-05	1.79e-06
LCM	1.00e+00	1.00e+00	1.00e+00	8.67e-01	–	1.45e-01	3.13e-10	1.88e-10	6.81e-13	2.46e-12
BLD	9.98e-01	9.98e-01	9.85e-01	9.14e-01	8.55e-01	–	1.72e-01	9.86e-02	3.62e-02	5.85e-03
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	8.28e-01	–	6.55e-04	6.20e-07	9.67e-08
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.01e-01	9.99e-01	–	1.27e-04	6.26e-07
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.64e-01	1.00e+00	1.00e+00	–	7.46e-06
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.94e-01	1.00e+00	1.00e+00	1.00e+00	–

(b) UTKFace

	BDG	RND	BAIT	LL	LCM	GSY	GSX	IGS	BLD	CST
BDG	–	3.06e-03	6.23e-02	2.77e-03	3.27e-08	1.46e-11	4.78e-12	3.10e-12	3.60e-04	2.15e-12
RND	9.97e-01	–	1.69e-01	9.06e-03	8.74e-07	2.02e-10	7.56e-11	3.50e-11	5.65e-04	6.35e-12
BAIT	9.38e-01	8.31e-01	–	6.45e-02	6.72e-03	4.93e-06	1.41e-06	1.01e-06	9.24e-04	2.14e-07
LL	9.97e-01	9.91e-01	9.36e-01	–	1.67e-01	1.07e-04	1.76e-05	2.00e-05	1.73e-03	2.26e-06
LCM	1.00e+00	1.00e+00	9.93e-01	8.33e-01	–	3.97e-11	1.03e-08	6.48e-10	4.83e-03	9.47e-11
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	6.33e-04	4.12e-07	1.17e-01	1.03e-08
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	–	1.10e-03	2.14e-01	5.65e-08
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	–	3.20e-01	8.71e-06
BLD	1.00e+00	9.99e-01	9.99e-01	9.98e-01	9.95e-01	8.83e-01	7.86e-01	6.80e-01	–	4.44e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	5.56e-01	–

Table 6: Mean AULC and final test MAE ($\pm$ SD across seeds) for the 13 AL methods with crowd-sourced human annotations, sorted best (lowest mean AULC) to worst.

(a) AAFD

Method	**Mean AULC $\pm$ SD**	**Final MAE $\pm$ SD**
LearningLoss	9.24 ± 0.18	6.60 ± 0.14
ActiveLab (MA)	9.69 ± 0.07	7.53 ± 0.16
BADGE	9.73 ± 0.06	7.85 ± 0.22
Random	9.85 ± 0.07	7.78 ± 0.16
Random (MA)	9.87 ± 0.10	8.11 ± 0.20
LCMD	10.20 ± 0.06	8.13 ± 0.15
BAIT	10.28 ± 0.30	8.17 ± 0.63
GSy	10.33 ± 0.04	8.88 ± 0.12
iGS	10.49 ± 0.04	9.10 ± 0.10
GSx	10.50 ± 0.07	8.93 ± 0.19
BALD	10.57 ± 0.58	8.61 ± 0.84
Coreset	10.74 ± 0.09	9.37 ± 0.09
Uncertainty (MA)	10.75 ± 0.24	9.88 ± 0.44

(b) UTKFace

Method	**Mean AULC $\pm$ SD**	**Final MAE $\pm$ SD**
ActiveLab (MA)	9.93 ± 0.10	6.60 ± 0.22
LearningLoss	10.08 ± 0.18	6.64 ± 0.15
BADGE	10.34 ± 0.10	7.26 ± 0.27
Random (MA)	10.37 ± 0.07	7.69 ± 0.26
BAIT	10.44 ± 0.30	7.12 ± 0.47
Random	10.50 ± 0.08	7.40 ± 0.24
LCMD	10.82 ± 0.07	7.94 ± 0.25
GSy	11.49 ± 0.08	9.14 ± 0.17
BALD	11.62 ± 1.08	9.36 ± 1.73
GSx	11.75 ± 0.11	9.78 ± 0.19
iGS	11.80 ± 0.10	9.76 ± 0.25
Uncertainty (MA)	11.99 ± 1.09	10.46 ± 1.68
Coreset	12.11 ± 0.04	10.39 ± 0.21

On UTKFace, *ActiveLab* obtains the lowest mean AULC (9.93 ± 0.10), while *LearningLoss* follows closely and attains a comparable final MAE. In contrast, *Coreset* and multi-annotator *Uncertainty* rank among the weakest methods on both datasets. Overall, the results further show that the relative effectiveness of AL methods varies across datasets, even under the same crowd-sourced human annotation setting.

A.2.2 VLM ANNOTATORS

In Table 7, we report the mean AULC and final test MAE, averaged across random seeds, for all 13 AL methods with VLM-generated annotations. On AAFD, *BADGE* achieves the lowest mean AULC (9.15 ± 0.03), followed by *Random* and *ActiveLab*, while *LearningLoss* attains the lowest final MAE (6.45 ± 0.18). On UTKFace, *ActiveLab* and *BADGE* achieve the lowest mean AULC (10.25), with *ActiveLab* obtaining the lowest final MAE (6.98 ± 0.59). In contrast, multi-annotator *Uncertainty* performs worst on both datasets. These results show that the relative ranking of AL

Table 7: Mean AULC and final test MAE (± SD across seeds) for the 13 AL methods with VLM generated annotations, sorted best (lowest mean AULC) to worst.

(a) AAFD

Method	Mean AULC ± SD	Final MAE ± SD
BADGE	9.15 ± 0.03	6.95 ± 0.12
Random	9.27 ± 0.06	6.99 ± 0.15
ActiveLab (MA)	9.34 ± 0.05	7.18 ± 0.17
LearningLoss	9.47 ± 0.22	6.45 ± 0.18
LCMD	9.53 ± 0.08	7.30 ± 0.12
Random (MA)	9.54 ± 0.09	7.63 ± 0.22
BAIT	9.59 ± 0.28	7.14 ± 0.57
GSy	9.81 ± 0.04	8.19 ± 0.11
BALD	9.91 ± 0.44	7.79 ± 0.38
GSx	9.92 ± 0.05	8.10 ± 0.11
iGS	9.96 ± 0.04	8.36 ± 0.09
Coreset	10.16 ± 0.09	8.50 ± 0.10
Uncertainty (MA)	10.74 ± 0.50	9.84 ± 0.94

(b) UTKFace

Method	Mean AULC ± SD	Final MAE ± SD
ActiveLab (MA)	10.25 ± 0.17	6.98 ± 0.59
BADGE	10.25 ± 0.11	7.10 ± 0.26
Random	10.35 ± 0.08	7.24 ± 0.26
BAIT	10.49 ± 0.20	7.47 ± 0.62
Random (MA)	10.63 ± 0.08	7.91 ± 0.24
LCMD	10.73 ± 0.09	7.77 ± 0.27
LearningLoss	11.12 ± 0.54	7.30 ± 0.30
GSy	11.47 ± 0.08	9.12 ± 0.17
GSx	11.47 ± 0.10	9.40 ± 0.19
iGS	11.70 ± 0.08	9.71 ± 0.18
Coreset	11.77 ± 0.08	9.85 ± 0.19
BALD	12.59 ± 1.05	10.87 ± 1.21
Uncertainty (MA)	12.80 ± 0.95	11.67 ± 1.39

Table 8: Mean AULC and final test MAE (± SD across seeds) for the 10 AL methods with GT annotations, sorted best (lowest mean AULC) to worst.

(a) AAFD

Method	Mean AULC ± SD	Final MAE ± SD
LearningLoss	9.14 ± 0.22	6.21 ± 0.15
BADGE	9.19 ± 0.07	6.77 ± 0.19
Random	9.38 ± 0.06	7.10 ± 0.12
BAIT	9.54 ± 0.27	7.17 ± 0.60
LCMD	9.65 ± 0.04	7.55 ± 0.11
BALD	9.84 ± 0.53	7.46 ± 0.66
GSy	10.01 ± 0.04	8.46 ± 0.10
GSx	10.08 ± 0.05	8.44 ± 0.16
iGS	10.18 ± 0.04	8.73 ± 0.06
Coreset	10.38 ± 0.08	8.92 ± 0.22

(b) UTKFace

Method	Mean AULC ± SD	Final MAE ± SD
BADGE	10.17 ± 0.09	7.10 ± 0.29
Random	10.25 ± 0.10	7.03 ± 0.27
BAIT	10.37 ± 0.35	6.79 ± 0.69
LearningLoss	10.61 ± 0.37	6.88 ± 0.36
LCMD	10.74 ± 0.10	7.73 ± 0.19
GSy	11.36 ± 0.07	9.35 ± 0.19
GSx	11.50 ± 0.09	9.45 ± 0.15
iGS	11.61 ± 0.09	9.89 ± 0.21
BALD	11.77 ± 1.04	9.43 ± 1.76
Coreset	11.82 ± 0.06	10.15 ± 0.20

methods in terms of the mean AULC and final MAE depends on the source of annotation for a given dataset.

A.2.3 GT Annotators

Table 8 presents the mean AULC and final test MAE, averaged across random seeds, for the 10 AL methods with ground-truth annotations. On AAFD, *LearningLoss* achieves the lowest mean AULC (9.14 ± 0.22) and the lowest final MAE (6.21 ± 0.15), followed closely by *BADGE*. On UTKFace, *BADGE* achieves the lowest mean AULC (10.17 ± 0.09), while *BAIT* attains the lowest final MAE (6.79 ± 0.69). In contrast, *Coreset* performs worst on both datasets. These results further show that the relative rankings of AL methods depends on the source of annotation (human / VLM / GT) and varies from one dataset to another.

A.3 Experimental Setup

Overall pipeline. Figure 10 depicts an overview of our deep active learning framework for facial age estimation. Starting from an initial labeled pool $\mathcal{L}$, the model is trained and then used to select samples from the unlabeled pool $\mathcal{U}$ using a chosen AL strategy. Once a batch of samples is selected in a given AL round, the labels are queried from either human annotators ($A_1 - A_3$) or VLM annotators ($L_1 - L_3$); the queried samples together with their annotations are then appended to $\mathcal{L}$, the model is updated and its performance is evaluated on a held-out test set. The process is repeated until a stopping criterion is satisfied. This formulation supports both single-annotator and multi-annotator AL and enables a controlled comparison of different AL strategies under human and VLM-generated annotations. Each method is evaluated over 10 independent runs and the results are averaged to rule out the effects of randomness.

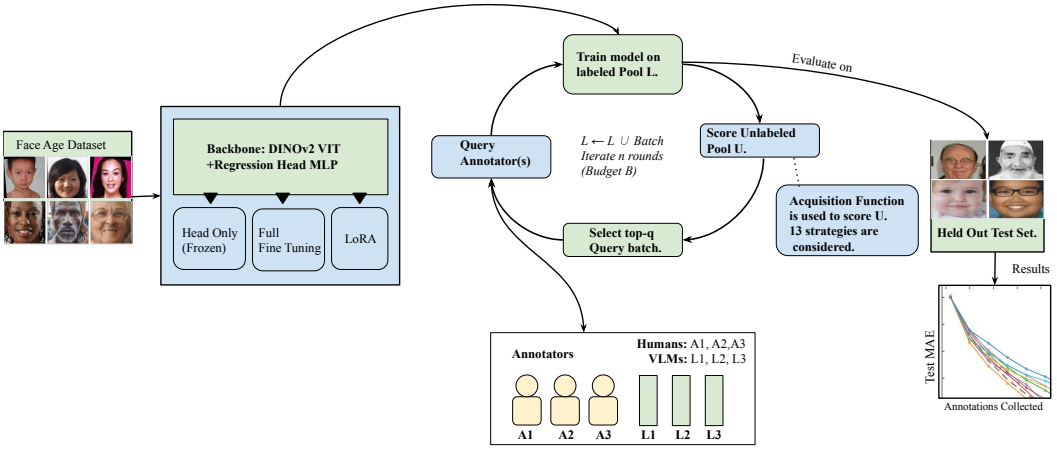


Figure 10: Overview of the deep active learning pipeline for facial age estimation.

The default query budget is taken as 200 annotations per active learning round, and we perform 10 AL rounds. We additionally ablate the initial labeled training set over $\{2\%, 5\%, 10\%\}$ of the data, and the query size over $\{100, 500\}$. In all the experiments with single annotator, we use the labels given by human annotator A_3 and the VLM Qwen-2.5.

Dataset splits. For each dataset, we construct a fixed unlabeled pool / validation / test split and reuse it across all acquisition methods, annotation settings, and ablation studies for consistency. For AAFD, the 6,637 images are divided into 4,652 pool, 660 validation, and 1,325 test images; for UTKFace, the 10,000 images are divided into 7,007 pool, 996 validation, and 1,997 test images. From each dataset's pool, we draw an initial labeled set equal to 1% of the pool (48 images for AAFD and 72 for UTKFace), with the remaining pool images (4,604 for AAFD and 6,935 for UTKFace) forming the unlabeled pool at the start of active learning.

Evaluation metric. Since we are dealing with a regression problem, we used the mean absolute error (MAE) on the test set as the evaluation metric in our work.

A.4 Implementation Details

Backbone and regression head. We use DINOv2 ViT-B/14 as the feature backbone with 224×224 input resolution. The pretrained model is loaded through torch.hub . The regression head is DNN with the following layers:

$$\text{LayerNorm} \to \text{Linear}(768, 256) \to \text{GELU} \to \text{Dropout}(0.1) \to \text{Linear}(256, 1).$$

A ResNet-18 backbone is additionally evaluated in the backbone ablation.

Fine-tuning strategies. We consider three fine-tuning regimes for the DINOv2 backbone. Our default setting is *full fine-tuning*, implemented using a two-phase training procedure. In the first phase, the backbone is frozen and only the regression head is optimized for up to 5 epochs with early-stopping patience of 3. In the second phase, the complete backbone is unfrozen and trained for up to 15 epochs with patience 5. We use differential learning rates of 2×10^{-5} for the regression head and 2×10^{-6} for the backbone. A fresh model is initialized from the same pre-trained checkpoint in every active learning round.

For *LoRA* Hu et al. (2021), adapters are inserted into the attention `qkv` and `proj` modules. We use rank $r = 16$, scaling parameter $\alpha = 32$, adapter dropout 0.05, and a learning rate of 2×10^{-4}. Training proceeds for up to 15 epochs, with only the LoRA adapters and regression head optimized while the original backbone parameters remain frozen.

For *head-only* fine-tuning, the backbone remains frozen throughout training and only the regression head is optimized, using a learning rate of 10^{-3} for up to 50 epochs with patience 10.

Across all regimes, we use the *Adam* optimizer with weight decay 10^{-4} and mean absolute error (MAE) as the training objective. Training uses a batch size of 32, while validation and test loaders use a batch size of 64. Predictive uncertainty used during in the acquisition functions is estimated using 20 Monte Carlo dropout forward passes with dropout probability $p = 0.1$.

Compute resources. All experiments were carried out using NVIDIA `B200` GPUs with 8 cores and 64 GB memory, and PyTorch 2.7.

A.5 Active Learning Algorithms and Adaptations

We evaluate 13 active-learning (AL) methods, 10 in single-annotator and 3 in multi-annotator settings. We use native regression-based active learning methods without modification and manually adapt classification-based methods for the facial age estimation regression setting.

A.5.1 Single-annotator AL Methods

We benchmark a diverse set of single-annotator AL strategies, covering random sampling, representation-based selection, uncertainty-based selection, loss prediction, gradient-based selection, and regression-specific methods. Methods originally designed for regression are applied directly, while classification-based approaches are adapted to continuous age prediction as described below.

Random. Samples are selected uniformly at random from the unlabeled pool without using model predictions or representations. Random sampling serves as a task-agnostic baseline for evaluating the effectiveness of different AL strategies.

Coreset.(Sener & Savarese (2018)) Samples are selected to improve coverage of the unlabeled data distribution by solving a k-center greedy objective in the learned feature space. Since sample selection depends only on learned representations, the method naturally extends to regression without modifying its selection criterion.

BADGE.(Ash et al. (2019)) BADGE jointly captures uncertainty and diversity through gradient-based sample representations. We adapt its gradient embedding construction to the regression objective and apply k-means++ initialization over the resulting embeddings to select a diverse batch of samples.

BALD.(Houlsby et al. (2011); Gal et al. (2017)) BALD estimates model uncertainty through Bayesian inference. For regression, we use Monte-Carlo (MC) dropout to obtain predictive uncertainty from multiple stochastic forward passes and prioritize samples with high uncertainty.

BAIT.(Ash et al. (2021)) BAIT performs sample selection using a Fisher-information-based criterion. We adapt the method to the regression setting of facial age estimation by computing gradients with respect to the regression loss instead of the classification objective.

LearningLoss.(Yoo & Kweon (2019)) LearningLoss trains an auxiliary module to predict the loss of each unlabeled sample and prioritizes samples expected to incur high loss, for selection. In our setting, the auxiliary module predicts the per-sample regression loss for age estimation.

LCMD(Holzmüller et al. (2023)) LCMD is a regression-oriented sample selection strategy that combines model uncertainty with diversity in the learned representation space. We follow the regression-specific formulation of Holzmüller *et al.*(Holzmüller et al. (2023)) for sample acquisition.

GSx.(Wu et al. (2019)) GSx promotes diversity in the input feature space by selecting unlabeled samples that are distant from the currently labeled examples. This encourages broader coverage of the input distribution during sampling.

GSy.(Wu et al. (2019)) GSy performs greedy selection in the predicted output space. It prioritizes samples whose predicted age values improve coverage of the continuous target space.

iGS.(Wu et al. (2019)) iGS jointly considers diversity in the input and predicted output spaces. It favors samples that are informative with respect to both their learned representations and predicted age values.

A.5.2 MULTI-ANNOTATOR AL METHODS

To the best of our knowledge, multi-annotator AL has not been studied for regression problems. We therefore use algorithms proposed for classification problems, and adapt them to the regression setup. We evaluate three AL algorithms in the multi-annotator setting: *Random, Reliability-weighted Uncertainty*, and *ActiveLab*(Goh & Mueller (2023)). Unlike conventional AL, these methods jointly determine both *which* image to acquire and *which* annotator to query.

Random. The multi-annotator Random baseline randomly samples an $(\text{image}, \text{annotator})$ pair from the set of available image–annotator combinations in each AL round.

Reliability-weighted Uncertainty. We first select the image with the highest predictive uncertainty, estimated using MC-dropout (Gal & Ghahramani (2016)). Among annotators who have not previously labeled the selected image, we query the annotator with the highest estimated reliability. Annotator reliability is derived from the running consensus model using the estimated annotator-specific bias b_j and variance v_j.

ActiveLab.(Goh & Mueller (2023)) We adapt *ActiveLab* from classification to the regression application of age estimation. We replace its categorical agreement and confidence measures with precision-weighted consensus and residual variance estimates from the regression annotator model. Each annotator is characterized by an estimated bias b_j and variance v_j, while the regression model is incorporated as a pseudo-annotator following the general *ActiveLab* framework.

Regression adaptation. For classification-based acquisition functions, we modify only the components that depend explicitly on categorical predictions. In particular, entropy or class posterior based uncertainty is replaced with predictive uncertainty appropriate for continuous outputs, while feature-space and regression-native sample selection criteria are retained without modification. This allows all methods to be evaluated under a common regression setting while preserving their original acquisition principles as closely as possible.

A.6 ABLATION STUDIES

We conduct ablation studies to examine the AL performance across different sizes of the initial labeled set, query size, different fine-tuning strategies, and backbone architectures.

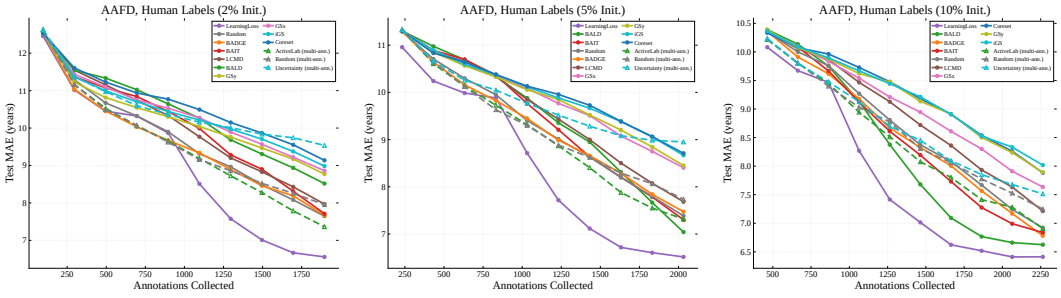
A.6.1 EFFECT OF INITIAL LABELED SET SIZE

The effect of the initial labeled set size on AL performance is shown in Figure 11. We evaluate initial labeled set sizes of 2%, 5%, and 10% on AAFD using crowd-sourced human annotations, while keeping the acquisition budget and evaluation protocol fixed. Across all three settings, test MAE decreases steadily as additional annotations are acquired, although the relative performance of the AL methods remains noticeably different. *LearningLoss* consistently achieves the lowest MAE in the later acquisition rounds, with a significant performance gap for all 3 settings. We observe the performance of AL strategies fluctuates depending upon the initial size of labeled set. Table 9 reports the pairwise p-values across AL methods for different initial labeled set sizes. We note that *LearningLoss* achieves statistically significant improvements ($p < 0.05$) over all other AL methods across all evaluated initial labeled set sizes.

As reported in table 10, across different initial labeled set sizes on AAFD, at least $7/13$ (54%) of the AL methods change ranks, with the largest shift of $11/13$ (85%) observed between the 2% and 10% initialization settings. The remaining comparisons also exhibit considerable ranking changes, ranging from $8/13$ (62%) for 1% vs. 2% to $10/13$ (77%) for several other pairs. Moreover, the Friedman tests yield $p < 0.05$ for all settings, indicating the ranking of the AL methods obtained for each initial set size is statistically significant. Figure 12 further illustrates these pairwise rank shifts, showing notable fluctuations in the relative ordering of methods as the initial labeled set size changes. These results indicate that AL methods are sensitive not only to the annotation source, but also to the size of the initial labeled set.

A.6.2 EFFECT OF QUERY SIZE

The effect of query size on AL performance is shown in Figure 13. We evaluate query sizes of 100, 200, and 500 on AAFD using crowd-sourced human annotations while keeping the remaining



(a) Initial labeled fraction: 2% (b) Initial labeled fraction: 5% (c) Initial labeled fraction: 10%

Figure 11: Active learning performance on AAFD with crowd-sourced human labels under different sizes of the initial labeled set. Best viewed in color.

Table 9: Pairwise p-values corresponding to one-sided t-test on AAFD with crowd-sourced annotations for different sizes of the initial labeled set. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column.

(a) Initial labeled fraction: 2%

	LL	MA-ALB	BDG	MA-RND	RND	BAIT	LCM	GSY	BLD	GSX	IGS	MA-UNC	CST
LL	–	2.63e-07	3.06e-08	4.55e-08	4.76e-08	7.29e-07	4.61e-09	2.35e-11	6.25e-07	1.69e-10	8.70e-11	2.37e-07	1.73e-11
MA-ALB	1.00e+00	–	5.38e-03	6.70e-04	3.81e-04	4.30e-05	3.03e-07	5.43e-10	3.02e-06	4.64e-11	6.68e-10	4.75e-05	1.67e-10
BDG	1.00e+00	9.95e-01	–	1.91e-01	5.39e-03	5.93e-05	1.02e-06	1.65e-09	1.03e-05	3.59e-10	4.41e-09	6.23e-05	5.14e-11
MA-RND	1.00e+00	9.99e-01	8.09e-01	–	1.22e-02	2.06e-04	5.20e-08	1.70e-10	1.65e-05	2.93e-10	3.72e-10	7.57e-05	5.98e-11
RND	1.00e+00	1.00e+00	9.95e-01	9.88e-01	–	1.15e-03	3.89e-07	8.52e-09	2.40e-05	8.86e-09	3.38e-09	1.56e-04	5.20e-10
BAIT	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	–	4.19e-01	1.27e-02	8.50e-03	2.55e-04	7.48e-04	1.54e-02	7.51e-06
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	5.81e-01	–	2.29e-05	3.18e-03	1.88e-06	7.91e-08	8.63e-03	1.64e-08
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.87e-01	1.00e+00	–	9.39e-02	1.44e-04	2.09e-06	5.72e-02	3.71e-08
BLD	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.92e-01	9.97e-01	9.06e-01	–	4.08e-01	3.87e-01	3.37e-01	2.40e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	5.92e-01	–	4.01e-01	3.16e-01	9.69e-07
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	6.13e-01	5.99e-01	–	3.24e-01	4.03e-06
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.85e-01	9.91e-01	9.43e-01	6.63e-01	6.84e-01	6.76e-01	–	1.59e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.76e-01	1.00e+00	1.00e+00	8.41e-01	–

(b) Initial labeled fraction: 5%

	LL	MA-ALB	MA-RND	BDG	RND	BAIT	BLD	LCM	MA-UNC	GSX	GSY	IGS	CST
LL	–	8.99e-08	2.43e-09	8.57e-11	2.58e-09	6.02e-09	3.20e-08	6.46e-12	2.07e-08	4.71e-13	1.03e-12	8.56e-13	2.26e-12
MA-ALB	1.00e+00	–	9.87e-03	3.66e-03	2.93e-03	7.75e-04	5.17e-04	1.20e-07	3.71e-05	2.32e-09	2.35e-09	8.09e-10	1.74e-10
MA-RND	1.00e+00	9.90e-01	–	3.20e-01	1.68e-01	5.12e-03	1.31e-03	1.24e-07	6.88e-06	4.25e-09	2.90e-10	3.02e-10	6.11e-10
BDG	1.00e+00	9.96e-01	6.80e-01	–	3.91e-01	1.37e-02	1.95e-03	7.49e-07	6.99e-05	1.09e-10	5.76e-10	5.46e-10	2.94e-10
RND	1.00e+00	9.97e-01	8.32e-01	6.09e-01	–	1.63e-02	3.69e-04	1.05e-06	1.70e-05	2.74e-08	1.27e-09	2.39e-11	9.25e-11
BAIT	1.00e+00	9.99e-01	9.95e-01	9.86e-01	9.84e-01	–	1.77e-01	5.21e-03	4.94e-04	1.70e-05	9.58e-06	2.31e-06	2.23e-06
BLD	1.00e+00	9.99e-01	9.99e-01	9.98e-01	1.00e+00	8.23e-01	–	8.21e-02	4.75e-03	9.07e-05	1.86e-05	1.82e-06	2.05e-06
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.95e-01	9.18e-01	–	2.74e-02	1.14e-08	1.70e-08	2.71e-09	2.37e-09
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.95e-01	9.73e-01	–	9.01e-02	4.09e-02	4.09e-03	4.01e-03
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.10e-01	–	7.47e-02	2.64e-04	4.86e-05
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.59e-01	9.25e-01	–	1.75e-05	5.15e-06
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.96e-01	1.00e+00	1.00e+00	–	2.58e-02
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.96e-01	1.00e+00	1.00e+00	9.74e-01	–

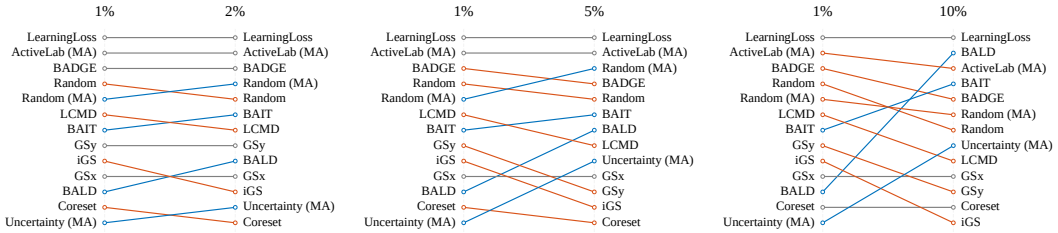
(c) Initial labeled fraction: 10%

	LL	BLD	MA-ALB	BAIT	BDG	MA-RND	RND	MA-UNC	LCM	GSX	GSY	CST	IGS
LL	–	4.81e-11	9.28e-09	1.43e-08	2.12e-12	3.60e-10	4.13e-12	4.38e-07	1.53e-12	2.05e-14	1.95e-16	4.09e-14	2.21e-13
BLD	1.00e+00	–	1.04e-04	9.29e-05	1.40e-08	1.98e-06	3.90e-08	2.70e-04	2.79e-09	4.11e-11	4.67e-13	2.63e-12	3.57e-12
MA-ALB	1.00e+00	1.00e+00	–	3.90e-02	2.72e-03	2.19e-03	3.02e-04	8.37e-03	5.21e-06	1.04e-08	1.70e-09	5.09e-09	3.03e-09
BAIT	1.00e+00	1.00e+00	9.61e-01	–	5.84e-02	2.84e-02	8.17e-03	1.17e-02	1.94e-05	5.99e-08	1.87e-08	2.01e-08	2.80e-08
BDG	1.00e+00	1.00e+00	9.97e-01	9.42e-01	–	1.52e-01	1.14e-03	9.73e-02	5.14e-07	1.39e-09	5.68e-11	5.48e-11	7.27e-10
MA-RND	1.00e+00	1.00e+00	9.98e-01	9.72e-01	8.48e-01	–	1.45e-01	2.26e-01	5.83e-05	2.10e-07	2.59e-08	1.75e-08	1.05e-08
RND	1.00e+00	1.00e+00	1.00e+00	9.92e-01	9.99e-01	8.55e-01	–	3.99e-01	4.52e-06	8.63e-08	3.43e-10	2.57e-09	8.64e-09
MA-UNC	1.00e+00	1.00e+00	9.92e-01	9.88e-01	9.03e-01	7.74e-01	6.01e-01	–	1.69e-02	5.73e-04	3.83e-05	6.51e-05	7.39e-05
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.83e-01	–	1.17e-04	8.29e-08	1.28e-08	6.71e-08
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	–	3.25e-07	9.86e-07	6.66e-06
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	7.98e-02	9.67e-02
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.20e-01	–	3.90e-01
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.03e-01	6.10e-01	–

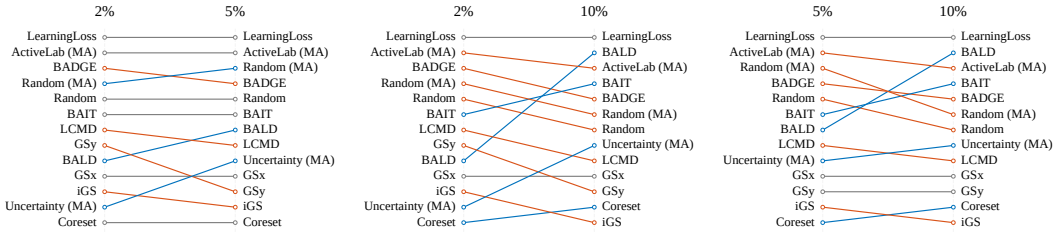
experimental parameters fixed. Across all three settings *LearningLoss* consistently achieves the lowest MAE in the later AL rounds, with a clear performance gap across all three query sizes. We further observe that several other AL strategies change their relative performance as the query size varies, indicating sensitivity to the number of samples acquired at each round. Table 11 reports the pairwise p-values among AL methods for each query size. *LearningLoss* achieves statistically

Table 10: Ranking consistency of AL methods on AAFD across different initial labeled set sizes.

Dataset	A vs. B	n	Moved	Friedman p (A / B)
AAFD	1% vs. 2%	13	8/13 (62%)	1.7e-16 / 2.8e-17
AAFD	1% vs. 5%	13	10/13 (77%)	1.7e-16 / 7.3e-18
AAFD	1% vs. 10%	13	10/13 (77%)	1.7e-16 / 4.2e-18
AAFD	2% vs. 5%	13	7/13 (54%)	2.8e-17 / 7.3e-18
AAFD	2% vs. 10%	13	11/13 (85%)	2.8e-17 / 4.2e-18
AAFD	5% vs. 10%	13	10/13 (77%)	7.3e-18 / 4.2e-18



(a) 1% vs. 2% (b) 1% vs. 5% (c) 1% vs. 10%



(d) 2% vs. 5% (e) 2% vs. 10% (f) 5% vs. 10%

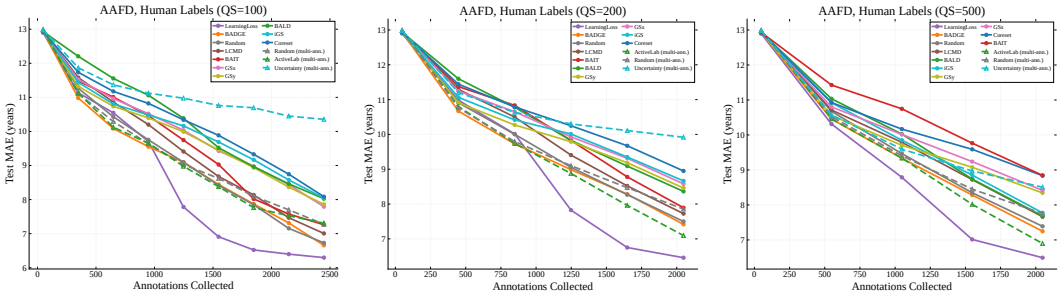
Figure 12: Pairwise rank shifts of AL methods on AAFD across different initial labeled set sizes.

significant improvements ($p < 0.05$) over all other AL methods for query sizes of 100, 200, and 500, demonstrating consistent performance across different acquisition batch sizes.

Table 12 summarizes the stability of AL method rankings across different query sizes on AAFD. We observe that $5/13$ (38%) of the methods change rank between QS100 and QS200, while the largest shift of $7/13$ (54%) occurs between QS100 and QS500. In contrast, only $2/13$ (15%) of the methods change position between QS200 and QS500, indicating comparatively greater stability at larger query sizes. The Friedman tests yield $p < 0.05$ for all query-size settings, confirming that the rankings of the AL methods for each query size is statistically significant. Figure 14 visualizes these pairwise changes and shows that, although most methods preserve their relative ordering, some exhibit clear rank shifts as the acquisition batch size changes. Overall, the results show that AL methods remain sensitive to the choice of query size, particularly when comparing smaller and larger acquisition batches.

A.6.3 Effect of Fine-Tuning Strategy

The effect of fine-tuning strategy on AL performance is shown in Figure 15. We compare LoRA, head-only, and full fine-tuning on UTKFace using crowd-sourced human annotations while keeping the other experimental parameters fixed. Across all three settings, test MAE decreases as additional annotations are acquired; however, both the absolute performance and relative ordering of AL methods vary considerably with the fine-tuning strategy. For LoRA and head-only fine-tuning, *Random (multi-ann.)* achieves the lowest overall error, whereas for full fine-tuning, *ActiveLab* and *LearningLoss* perform best in the later AL rounds. Table 13 reports the corresponding pairwise p-values among the AL methods for each fine-tuning strategy. *Random (multi-ann.)* shows statistically significant improvements ($p < 0.05$) over the other methods for LoRA and head-only fine-tuning, while



(a) Query size: 100 (b) Query size: 200 (c) Query size: 500

Figure 13: Active learning performance on AAFD with crowd-sourced human labels for different query sizes.

Table 11: Pairwise p-values corresponding to one-sided t-test on AAFD with crowd-sourced annotations for different query sizes. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column.

(a) Query size: 100

	LL	BDG	MA-ALB	RND	MA-RND	LCM	BAIT	GSY	GSX	IGS	BLD	CST	MA-UNC
LL	–	8.41e-07	2.19e-05	1.22e-07	9.49e-08	9.97e-10	6.22e-08	1.84e-10	1.24e-11	2.32e-10	2.69e-06	1.83e-11	1.65e-07
BDG	1.00e+00	–	1.19e-01	2.69e-04	2.18e-04	4.28e-07	1.10e-06	2.78e-10	2.35e-10	4.46e-11	2.08e-05	1.47e-11	9.59e-07
MA-ALB	1.00e+00	8.81e-01	–	4.38e-01	3.40e-02	9.37e-05	3.48e-05	1.64e-08	3.42e-08	1.15e-09	6.98e-05	1.89e-09	7.72e-07
RND	1.00e+00	1.00e+00	5.62e-01	–	2.03e-03	5.11e-07	4.68e-06	2.69e-12	1.06e-11	1.38e-13	4.94e-05	1.82e-12	1.15e-06
MA-RND	1.00e+00	1.00e+00	9.66e-01	9.98e-01	–	4.63e-05	4.03e-05	5.58e-10	1.16e-10	8.92e-11	2.07e-04	1.99e-10	1.73e-06
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	3.31e-02	6.69e-08	7.50e-09	2.11e-08	9.06e-04	4.30e-11	5.59e-06
BAIT	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.67e-01	–	1.01e-04	3.33e-05	9.82e-06	3.61e-03	3.89e-07	3.59e-06
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	8.78e-05	6.68e-07	3.76e-02	2.63e-08	3.66e-05
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	3.69e-02	8.64e-02	6.43e-07	7.23e-05
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.63e-01	–	1.14e-01	1.73e-06	1.28e-04
BLD	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.96e-01	9.62e-01	9.14e-01	8.86e-01	–	4.39e-01	1.05e-02
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	5.61e-01	–	1.02e-03
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.90e-01	9.99e-01	–

(b) Query size: 200

	LL	MA-ALB	BDG	RND	MA-RND	LCM	BAIT	GSY	IGS	BLD	GSX	CST	MA-UNC
LL	–	6.23e-06	4.33e-07	8.40e-07	7.97e-08	2.98e-09	4.33e-09	5.46e-11	6.59e-11	4.69e-05	1.77e-10	2.36e-11	1.74e-08
MA-ALB	1.00e+00	–	1.49e-03	2.32e-04	8.50e-05	5.78e-08	4.48e-06	9.73e-10	1.13e-09	4.82e-04	9.16e-11	1.58e-10	2.28e-08
BDG	1.00e+00	9.99e-01	–	4.95e-04	1.68e-04	1.45e-08	3.36e-06	6.08e-10	8.09e-11	1.04e-03	4.31e-13	3.00e-11	1.22e-08
RND	1.00e+00	1.00e+00	1.00e+00	–	4.34e-02	1.60e-06	4.22e-05	2.18e-07	1.23e-09	2.43e-03	6.25e-11	2.78e-10	1.07e-08
MA-RND	1.00e+00	1.00e+00	1.00e+00	9.57e-01	–	9.67e-06	7.57e-06	1.02e-07	5.01e-09	5.70e-03	2.83e-09	8.34e-10	2.94e-07
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	8.12e-03	2.34e-05	3.50e-09	4.12e-02	3.33e-09	2.13e-09	8.27e-07
BAIT	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.92e-01	–	3.15e-01	2.73e-03	1.87e-01	2.59e-03	4.62e-06	5.30e-05
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.85e-01	–	9.49e-06	1.67e-01	1.20e-05	2.70e-08	2.94e-05
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.97e-01	1.00e+00	–	4.71e-01	1.71e-01	5.16e-08	1.71e-04
BLD	1.00e+00	1.00e+00	9.99e-01	9.98e-01	9.94e-01	9.59e-01	8.13e-01	8.33e-01	5.29e-01	–	4.94e-01	1.16e-01	2.90e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.97e-01	1.00e+00	8.29e-01	5.06e-01	–	6.36e-07	1.06e-04
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	8.84e-01	1.00e+00	–	2.70e-02
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.71e-01	1.00e+00	9.73e-01	–

(c) Query size: 500

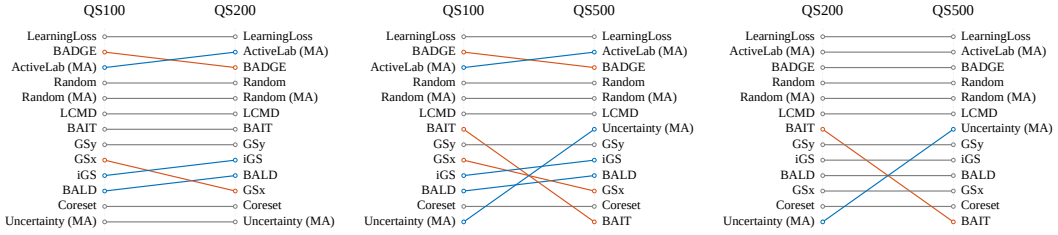
	LL	MA-ALB	BDG	RND	MA-RND	LCM	MA-UNC	GSY	IGS	BLD	GSX	CST	BAIT
LL	–	2.28e-06	6.38e-06	7.78e-08	3.05e-07	2.83e-08	3.23e-06	7.77e-09	8.10e-10	5.21e-05	1.36e-09	1.70e-10	3.78e-06
MA-ALB	1.00e+00	–	1.83e-02	1.52e-06	2.35e-05	3.21e-07	5.31e-05	3.08e-08	2.02e-07	1.96e-03	3.61e-09	8.33e-10	1.56e-05
BDG	1.00e+00	9.82e-01	–	2.61e-03	4.51e-03	7.44e-07	5.78e-04	2.10e-06	2.92e-06	3.67e-03	3.75e-09	6.30e-09	9.06e-06
RND	1.00e+00	1.00e+00	9.97e-01	–	2.45e-01	3.54e-06	1.32e-03	7.91e-07	1.36e-06	1.70e-02	1.24e-08	2.07e-09	4.90e-05
MA-RND	1.00e+00	1.00e+00	9.95e-01	7.55e-01	–	5.47e-04	5.81e-03	6.55e-06	1.34e-04	2.93e-02	1.95e-06	2.49e-07	1.25e-04
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	–	1.53e-01	1.41e-03	4.77e-04	1.60e-01	6.04e-06	4.10e-07	3.46e-04
MA-UNC	1.00e+00	1.00e+00	9.99e-01	9.99e-01	9.94e-01	8.47e-01	–	4.09e-01	4.06e-01	3.39e-01	2.04e-02	8.37e-04	1.75e-03
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	5.91e-01	–	4.92e-01	4.21e-01	6.78e-04	2.55e-06	1.30e-03
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	5.94e-01	5.08e-01	–	4.12e-01	8.40e-04	4.49e-06	1.94e-03
BLD	1.00e+00	9.98e-01	9.96e-01	9.83e-01	9.71e-01	8.40e-01	6.61e-01	5.79e-01	5.88e-01	–	1.67e-01	2.72e-02	1.57e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.80e-01	9.99e-01	9.99e-01	8.33e-01	–	1.29e-06	9.13e-03
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	9.73e-01	1.00e+00	–	1.07e-01
BAIT	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.98e-01	9.99e-01	9.98e-01	9.84e-01	9.91e-01	8.93e-01	–

ActiveLab exhibits significant improvements for full fine-tuning. These results further demonstrate that the performance of AL algorithms can depend on the underlying fine-tuning strategy.

Table 14 summarizes the stability of AL method rankings across different fine-tuning strategies on UTKFace. We observe large ranking shifts in all comparisons: $10/13$ (77%) methods change position between full fine-tuning and LoRA, all $13/13$ (100%) methods shift between full fine-tuning and head-only training, and $12/13$ (92%) change rank between LoRA and head-only fine-

Table 12: Ranking consistency of AL methods on AAFD across different query sizes.

Dataset	A vs. B	n	Moved	Friedman p (A / B)
AAFD	QS100 vs. QS200	13	5/13 (38%)	1.5e-18 / 2.8e-17
AAFD	QS100 vs. QS500	13	7/13 (54%)	1.5e-18 / 2.3e-15
AAFD	QS200 vs. QS500	13	2/13 (15%)	2.8e-17 / 2.3e-15



(a) QS100 vs. QS200 (b) QS100 vs. QS500 (c) QS200 vs. QS500

Figure 14: Pairwise rank shifts of AL methods on AAFD across different query sizes.

tuning. The Friedman tests yield $p < 0.05$ for all settings, indicating that the ranking of AL methods across each fine-tuning strategy is statistically significant. Figure 16 further illustrates these changes, with several methods moving substantially in relative order as the fine-tuning strategy changes. These results show that the relative performance of AL methods are sensitive to the underlying fine-tuning technique, reinforcing that results can vary considerably even when the dataset and annotation source remain fixed.

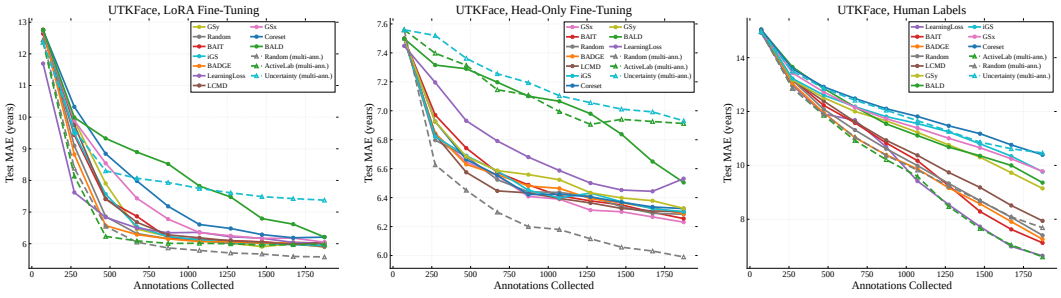
A.6.4 Effect of Backbone Architecture

The effect of backbone architecture on AL performance is shown in Figure 17. We compare DINOv2 ViT and ResNet18 backbones on AAFD using crowd-sourced human annotations while other experimental settings remain fixed. With DINOv2 ViT, *LearningLoss* achieves the lowest MAE in the later acquisition rounds by a clear margin, whereas with ResNet18, *BADGE* and *LearningLoss* emerge as the strongest methods. Table 15 reports the corresponding pairwise p-values among the AL methods for each fine-tuning strategy. With DINOv2 ViT as the backbone, *LearningLoss* significantly outperforms all other methods ($p < 0.05$), while with ResNet18, *BADGE* achieves the best overall performance but is not significantly different from *LearningLoss* ($p = 0.0516$). These results demonstrate that the performance of AL methods are dependent on the underlying model architecture.

Figure 18 illustrates the change in AL method rankings when the backbone is switched from DINOv2 ViT to ResNet18 on AAFD. We observe that all $13/13$ (100%) methods change rank, indicating a complete reordering of the benchmark under the two architectures. For example, *LearningLoss* ranks first with ViT but moves to second with ResNet18, while *BADGE* rises to the top position under ResNet18. Table 16 reports Friedman p-values below 0.05 for both backbones, confirming that the rankings of the AL methods for both the backbone architectures are statistically significant. These results show that the performance of AL methods are sensitive to the choice of the backbone model architecture.

A.7 VLM Prompt

In figure 19 we present the prompt used to collect VLM-based age annotations. Each VLM is instructed to provide both a point age estimate and an uncertainty-aware age range in a structured JSON format.



(a) LoRA (b) Head-only (c) Full fine-tuning

Figure 15: Active learning performance on UTKFace with crowd-sourced human annotations for different fine-tuning strategies.

Table 13: Pairwise p-values corresponding to one-sided t-test on UTKFace with crowd-sourced annotations for different fine-tuning strategies. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column.

(a) LoRA

	MA-RND	MA-ALB	LL	BDG	RND	BAIT	IGS	GSY	LCM	GSX	CST	MA-UNC	BLD
MA-RND	–	1.53e-04	4.82e-05	2.31e-05	7.39e-07	4.91e-05	7.32e-08	4.46e-09	1.01e-08	3.07e-11	5.58e-10	3.65e-03	3.86e-04
MA-ALB	1.00e+00	–	2.53e-03	4.52e-04	4.13e-05	2.38e-04	8.22e-07	3.15e-09	1.01e-08	2.23e-10	4.08e-09	4.66e-03	5.73e-04
LL	1.00e+00	9.97e-01	–	1.01e-01	1.73e-02	2.37e-02	3.50e-04	1.73e-05	7.51e-05	8.40e-08	3.87e-08	9.20e-03	1.41e-03
BDG	1.00e+00	1.00e+00	8.99e-01	–	1.19e-01	2.41e-02	3.90e-04	2.04e-05	1.03e-05	7.33e-08	1.34e-09	9.46e-03	1.36e-03
RND	1.00e+00	1.00e+00	9.83e-01	8.81e-01	–	8.25e-02	2.01e-03	9.71e-04	3.48e-04	5.26e-10	1.69e-08	1.10e-02	1.51e-03
BAIT	1.00e+00	1.00e+00	9.76e-01	9.76e-01	9.17e-01	–	3.54e-01	1.50e-01	1.11e-01	1.02e-04	7.83e-06	1.46e-02	1.97e-03
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.98e-01	6.46e-01	–	1.30e-01	1.02e-01	1.97e-07	2.92e-07	1.98e-02	3.99e-03
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	8.50e-01	8.70e-01	–	4.52e-01	9.86e-06	2.03e-07	2.18e-02	3.73e-03
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	8.89e-01	8.98e-01	5.48e-01	–	2.60e-06	5.02e-08	2.16e-02	3.43e-03
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	2.92e-04	7.05e-02	1.85e-02
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	1.60e-01	5.76e-02
MA-UNC	9.96e-01	9.95e-01	9.91e-01	9.91e-01	9.89e-01	9.85e-01	9.80e-01	9.78e-01	9.78e-01	9.29e-01	8.40e-01	–	3.87e-01
BLD	1.00e+00	9.99e-01	9.99e-01	9.99e-01	9.98e-01	9.98e-01	9.96e-01	9.96e-01	9.97e-01	9.82e-01	9.42e-01	6.13e-01	–

(b) Head-only

	MA-RND	LCM	GSX	RND	BDG	IGS	CST	BAIT	GSY	LL	BLD	MA-ALB	MA-UNC
MA-RND	–	1.74e-05	1.44e-06	5.55e-09	6.49e-07	3.14e-07	3.94e-06	3.02e-05	7.13e-07	2.55e-08	1.64e-09	2.60e-08	1.27e-08
LCM	1.00e+00	–	4.93e-01	1.35e-01	1.13e-01	8.59e-02	1.39e-01	6.54e-02	4.76e-03	9.04e-06	2.62e-07	1.78e-06	3.25e-07
GSX	1.00e+00	5.07e-01	–	1.11e-01	8.40e-02	2.51e-02	3.93e-02	5.67e-02	1.61e-03	4.10e-05	1.45e-07	1.02e-06	2.50e-07
RND	1.00e+00	8.65e-01	8.89e-01	–	2.79e-01	2.38e-01	2.25e-01	1.62e-01	6.23e-03	4.08e-05	8.48e-08	4.33e-07	1.25e-07
BDG	1.00e+00	8.87e-01	9.16e-01	7.21e-01	–	4.67e-01	4.27e-01	3.39e-01	1.14e-01	2.61e-04	2.33e-07	8.51e-07	3.51e-07
IGS	1.00e+00	9.14e-01	9.75e-01	7.62e-01	5.33e-01	–	4.50e-01	3.72e-01	3.46e-02	5.56e-06	4.35e-07	2.81e-06	4.64e-07
CST	1.00e+00	8.61e-01	9.61e-01	7.75e-01	5.73e-01	5.50e-01	–	3.72e-01	1.15e-01	5.43e-04	5.42e-08	2.78e-07	2.27e-07
BAIT	1.00e+00	9.35e-01	9.43e-01	8.38e-01	6.61e-01	6.28e-01	6.28e-01	–	1.85e-01	2.86e-03	7.03e-07	1.64e-06	7.06e-07
GSY	1.00e+00	9.95e-01	9.98e-01	9.94e-01	8.86e-01	9.65e-01	8.85e-01	8.15e-01	–	3.76e-04	2.78e-06	4.87e-06	4.24e-07
LL	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.97e-01	1.00e+00	–	5.94e-06	5.15e-05	4.52e-06
BLD	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	1.20e-01	9.67e-03
MA-ALB	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	8.80e-01	–	8.06e-03
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.90e-01	9.92e-01	–

(c) Full fine-tuning

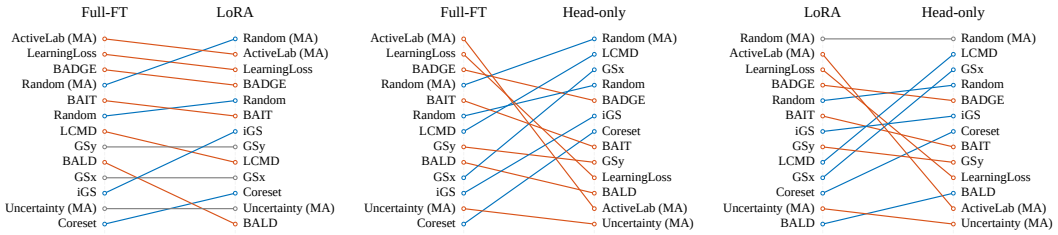
	MA-ALB	LL	BDG	MA-RND	BAIT	RND	LCM	GSY	BLD	GSX	IGS	MA-UNC	CST
MA-ALB	–	7.64e-04	7.38e-06	7.19e-07	2.72e-04	4.60e-08	6.81e-10	4.71e-11	2.70e-04	2.01e-11	4.95e-11	5.35e-05	9.37e-13
LL	9.99e-01	–	1.37e-03	4.29e-04	2.69e-03	4.26e-05	1.98e-07	2.57e-09	4.73e-04	1.06e-09	9.46e-10	6.66e-05	1.21e-10
BDG	1.00e+00	9.99e-01	–	2.77e-01	1.49e-01	1.26e-03	2.28e-08	4.52e-12	2.84e-03	1.78e-10	5.20e-12	5.96e-04	2.50e-12
MA-RND	1.00e+00	1.00e+00	7.23e-01	–	2.29e-01	1.78e-03	2.63e-07	2.86e-10	2.48e-03	6.89e-11	3.96e-11	5.06e-04	1.34e-13
BAIT	1.00e+00	9.97e-01	8.51e-01	7.71e-01	–	2.62e-01	7.70e-04	1.82e-06	4.10e-03	2.80e-08	1.05e-07	8.65e-04	1.51e-08
RND	1.00e+00	1.00e+00	9.99e-01	9.98e-01	7.38e-01	–	1.14e-06	1.10e-09	3.93e-03	6.02e-11	5.26e-10	7.95e-04	3.84e-13
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	–	3.30e-09	2.45e-02	2.01e-10	1.63e-10	3.97e-03	3.67e-12
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	3.68e-01	2.69e-04	2.98e-07	9.75e-02	2.34e-09
BLD	1.00e+00	1.00e+00	9.97e-01	9.98e-01	9.96e-01	9.96e-01	9.75e-01	6.32e-01	–	3.59e-01	3.21e-01	4.60e-02	9.45e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.41e-01	–	1.66e-01	2.51e-01	2.78e-07
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.79e-01	8.34e-01	–	3.01e-01	2.23e-06
MA-UNC	1.00e+00	1.00e+00	9.99e-01	9.99e-01	9.99e-01	9.99e-01	9.96e-01	9.02e-01	9.54e-01	7.49e-01	6.99e-01	–	3.71e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.06e-01	1.00e+00	1.00e+00	6.29e-01	–

A.8 Limitations and Future Work

Our work focuses on a single regression task, facial age estimation, with three human and three VLM annotators per image. Although our annotations include both point and range annotations, our current AL experiments primarily use point estimates and do not exploit the range-based annotations. In addition, the benchmark inherits the demographic biases of the underlying facial datasets, which

Table 14: Ranking consistency of active learning methods on UTKFace across different fine-tuning strategies.

Dataset	**A vs. B**	n	**Moved**	**Friedman p (A / B)**
UTKFace	Full-FT vs. LoRA	13	10/13 (77%)	7.9e-17 / 1.5e-14
UTKFace	Full-FT vs. Head-only	13	13/13 (100%)	7.9e-17 / 7.9e-15
UTKFace	LoRA vs. Head-only	13	12/13 (92%)	1.5e-14 / 7.9e-15



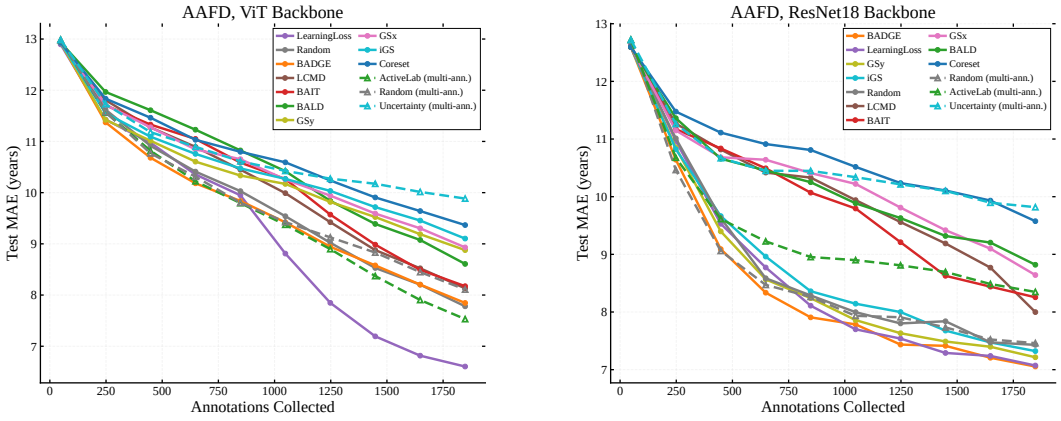
(a) Full-FT vs. LoRA (b) Full-FT vs. Head-only (c) LoRA vs. Head-only

Figure 16: Pairwise rank shifts of active learning methods on UTKFace across different fine-tuning strategies.

may affect both the human and VLM age estimates. In future, we plan to develop multi-annotator AL methods that explicitly model annotator disagreement, age-dependent uncertainty, and range-based supervision. We also plan to study cost-aware annotation strategies that jointly decide which samples to query and whether to obtain supervision from human annotators or VLMs.

Our work provides a foundation for future research on active learning under realistic, multi-source supervision. The benchmark can support the development of multi-annotator AL methods that explicitly model disagreement, age-dependent annotation noise, range-based uncertainty, and cost-aware selections. It also enables studying whether annotation source dependent ranking instability extends to other subjective regression tasks. At the same time, because facial age estimation can affect sensitive applications, future use of *AgeALBench* should consider fairness, privacy, and demographic bias when evaluating or deploying age-estimation systems.

Lastly, the images in the UTKFace dataset were annotated using a hybrid approach combining automatic labeling and manual verification. The raw age labels were initially estimated automatically using DEX (Deep Expectation); human annotators subsequently reviewed and manually verified or cleaned the labels to correct major discrepancies. We acknowledge that there might be some approximation errors in the original dataset's ground truth themselves.



(a) ViT (DINOv2)

(b) ResNet18

Figure 17: Active learning performance on AAFD with crowd-sourced human labels with different backbone architectures.

Table 15: Pairwise p-values corresponding to one-sided t-test on AAFD with crowd-sourced annotations for different backbone architectures. The entry at row i and column j denotes the p-value corresponding to the hypothesis that the method in the row is better than the method in the column.

(a) ViT (DINOv2)

	LL	MA-ALB	BDG	RND	MA-RND	LCM	BAIT	GSY	IGS	GSX	BLD	CST	MA-UNC
LL	–	2.53e-05	7.23e-06	1.77e-06	6.10e-06	3.51e-08	1.44e-06	7.53e-09	3.00e-09	1.20e-09	2.98e-05	2.79e-10	5.31e-08
MA-ALB	1.00e+00	–	1.09e-01	1.59e-03	2.41e-03	2.54e-09	1.19e-04	8.71e-11	8.52e-12	1.17e-10	5.01e-04	2.14e-10	4.18e-07
BDG	1.00e+00	8.91e-01	–	2.41e-03	8.08e-04	1.72e-09	1.16e-04	8.27e-10	1.09e-10	4.39e-10	7.67e-04	2.73e-10	2.29e-07
RND	1.00e+00	9.98e-01	9.98e-01	–	2.83e-01	3.44e-06	5.98e-04	3.09e-08	5.84e-09	3.14e-08	1.16e-03	3.37e-09	4.85e-07
MA-RND	1.00e+00	9.98e-01	9.99e-01	7.17e-01	–	1.34e-06	1.74e-03	2.53e-07	2.05e-08	2.38e-08	2.91e-03	8.00e-10	1.53e-06
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	–	1.96e-01	4.97e-05	4.47e-08	4.80e-08	3.91e-02	1.28e-08	3.73e-05
BAIT	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.98e-01	8.04e-01	–	3.14e-01	2.89e-02	3.12e-02	5.09e-02	9.64e-04	1.35e-03
GSY	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	6.86e-01	–	1.92e-07	9.21e-05	1.14e-01	2.33e-07	3.35e-04
IGS	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.71e-01	1.00e+00	–	2.39e-01	3.32e-01	2.26e-06	5.30e-03
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.69e-01	1.00e+00	7.61e-01	–	3.64e-01	1.66e-07	9.58e-03
BLD	1.00e+00	9.99e-01	9.99e-01	9.99e-01	9.97e-01	9.61e-01	9.49e-01	8.86e-01	6.68e-01	6.36e-01	–	2.07e-01	1.88e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	1.00e+00	7.93e-01	–	4.60e-01
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	1.00e+00	9.95e-01	9.90e-01	8.12e-01	5.40e-01	–

(b) ResNet18

	BDG	LL	GSY	MA-RND	RND	IGS	MA-ALB	BAIT	LCM	BLD	GSX	MA-UNC	CST
BDG	–	5.16e-02	3.11e-03	8.17e-04	7.40e-07	2.05e-02	1.26e-03	1.71e-05	1.55e-07	4.11e-04	2.59e-08	5.62e-05	1.64e-12
LL	9.48e-01	–	3.66e-01	3.11e-01	3.25e-02	1.50e-01	6.66e-03	1.50e-04	2.40e-06	1.65e-03	2.73e-06	8.22e-05	1.18e-10
GSY	9.97e-01	6.34e-01	–	3.96e-01	2.21e-02	1.17e-01	1.56e-03	4.63e-05	5.22e-08	1.10e-03	2.83e-07	8.74e-05	1.99e-10
MA-RND	9.99e-01	6.89e-01	6.04e-01	–	8.34e-03	1.75e-01	3.31e-03	6.24e-05	3.88e-08	8.27e-04	1.25e-07	8.25e-05	1.02e-11
RND	1.00e+00	9.68e-01	9.78e-01	9.92e-01	–	3.97e-01	1.54e-02	1.59e-04	8.58e-07	1.61e-03	7.17e-08	2.45e-04	3.44e-12
IGS	9.79e-01	8.50e-01	8.83e-01	8.25e-01	6.03e-01	–	2.09e-02	8.48e-04	1.21e-04	2.69e-03	2.22e-05	9.56e-04	1.43e-06
MA-ALB	9.99e-01	9.93e-01	9.98e-01	9.97e-01	9.85e-01	9.79e-01	–	4.02e-02	2.72e-03	4.59e-02	4.14e-03	1.07e-03	8.54e-05
BAIT	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.60e-01	–	2.40e-01	2.92e-01	9.80e-02	8.05e-02	3.79e-03
LCM	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.97e-01	7.60e-01	–	3.96e-01	1.42e-01	9.96e-02	2.38e-04
BLD	1.00e+00	9.98e-01	9.99e-01	9.99e-01	9.98e-01	9.97e-01	9.54e-01	7.08e-01	6.04e-01	–	4.18e-01	2.66e-01	8.41e-02
GSX	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.96e-01	9.02e-01	8.58e-01	5.82e-01	–	2.37e-01	1.66e-03
MA-UNC	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.99e-01	9.99e-01	9.19e-01	9.00e-01	7.34e-01	7.63e-01	–	3.04e-01
CST	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.96e-01	1.00e+00	9.16e-01	9.98e-01	6.96e-01	–

Table 16: Ranking consistency of active learning methods on AAFD across different backbone architectures.

Dataset	**A vs. B**	n	**Moved**	**Friedman p (A / B)**
AAFD	ViT vs. ResNet18	13	13/13 (100%)	1.7e-16 / 1.3e-14

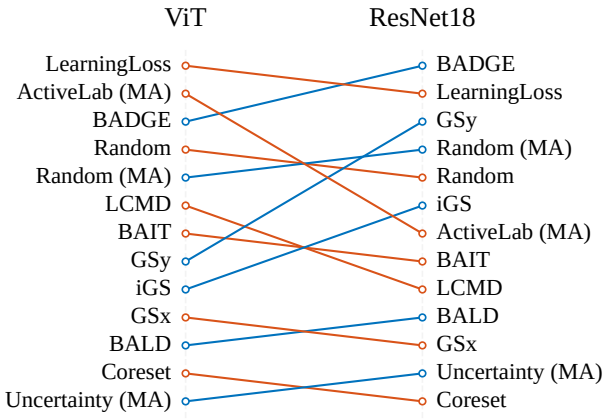


Figure 18: Rank shifts of active learning methods on AAFD when changing the backbone from ViT (DINOv2) to ResNet18.

VLM Age Estimation Prompt

Look carefully at the face in this image and estimate the person's age.

Response format. Respond **only** with a valid JSON object—no explanation, no markdown, and no extra text. The JSON must contain exactly the following three keys:

`"age"`: an integer representing your best single age estimate.

`"age_range_low"`: an integer representing the lower bound of your confidence interval.

`"age_range_high"`: an integer representing the upper bound of your confidence interval.

Uncertainty. The interval should reflect your uncertainty: use a narrow range (e.g., $\pm$1–4 years) when confident and a wider range (e.g., $\pm$5–9 years) when uncertain. The total range must be at most 10 years wide.

Example outputs

```
{"age": 27, "age_range_low": 25, "age_range_high": 28}
{"age": 55, "age_range_low": 53, "age_range_high": 58}
{"age": 77, "age_range_low": 76, "age_range_high": 79}
{"age": 41, "age_range_low": 35, "age_range_high": 43}
```

Now produce the JSON for the image.

Figure 19: **Prompt used for VLM-based age annotation.** The VLM provides a point age estimate together with an uncertainty-aware age interval in a structured JSON format.