QuMAB: Query-based Multi-annotator Behavior Pattern Learning

Liyun Zhang[1*], Zheng Lian[2], Hong Liu[3], Takanori Takebe[4], Shozo Nishii[5], Yuta Nakashima[6]

[1]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
[2]National Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai, China
[3]School of Informatics, Xiamen University, Fujian, China
[4]Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati OH, USA
[5]Advanced Medical Research Center, Yokohama City University, Kanagawa, Japan
[6]Institute of Scientific and Industrial Research, The University of Osaka, Osaka, Japan
liyun.zhang@lab.ime.cmc.osaka-u.ac.jp

Abstract

Multi-annotator learning traditionally aggregates diverse annotations to approximate a single "ground truth", treating disagreements as noise. However, this paradigm faces fundamental challenges: subjective tasks often lack absolute ground truth, and sparse annotation coverage makes aggregation statistically unreliable. We introduce a paradigm shift from sample-wise aggregation to annotator-wise behavior modeling. By treating annotator disagreements as valuable information rather than noise, modeling annotator-specific behavior patterns can reconstruct unlabeled data to reduce annotation cost, enhance aggregation reliability, and explain annotator decision behavior. To this end, we propose QuMAB (**Qu**ery-based **M**ulti-**A**nnotator **B**ehavior Pattern Learning), which uses lightweight queries to model individual annotators while capturing inter-annotator correlations as implicit regularization, preventing overfitting to sparse individual data while maintaining individualization and improving generalization, with a visualization of annotator focus regions offering an explainable analysis of behavior understanding. We contribute two large-scale datasets with dense per-annotator labels: STREET (4,300 labels/annotator) and AMER (average 3,118 labels/annotator), the first multimodal multi-annotator dataset. Extensive experiments demonstrate the superiority of our QuMAB in modeling individual annotators' behavior patterns, their utility for consensus prediction, and applicability under sparse annotations.

Code & Datasets — https://github.com/zly9120/QuMAB
Extended version — https://arxiv.org/abs/2507.17653

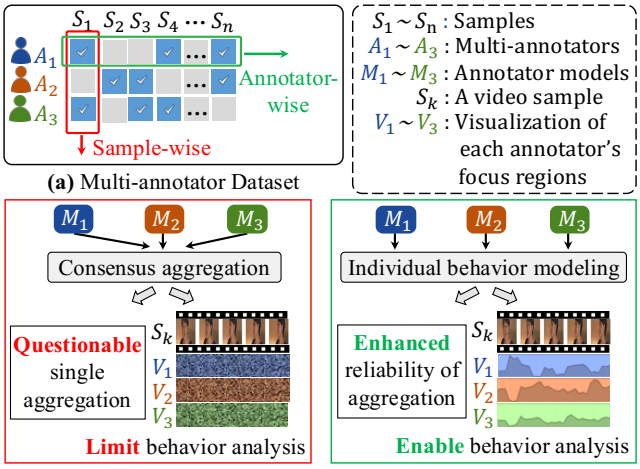

Figure 1: Paradigm shift from sample-wise aggregation to annotator-wise behavior modeling. **(a)**: Sparse annotation matrix showing each annotator labels a small subset of samples with disjoint coverage. **(b)**: Traditional sample-wise aggregation makes a questionable single "ground truth" prediction, potentially losing individual information to limit behavior analysis. **(c)**: Our annotator-wise behavior modeling captures each annotator's behavior patterns longitudinally across their labeled samples, enhancing reliability of aggregation via reconstructed unlabeled data in annotation matrix, offering explainable analysis of behavior understanding.

Introduction

In real-world multi-annotation scenarios, such as medical analysis (Liao et al. 2024; Wang et al. 2024), sentiment analysis (Lian et al. 2023; Zhang et al. 2024), and visual perception (Zhang et al. 2023a), different annotators often provide different labels to the same sample (Raykar et al. 2010) due to different personal backgrounds, subjective interpretations, and preferences. Traditional multi-annotator learning focuses on learning different characteristics (e.g., confusion mode (Tanno et al. 2019a), agreement (Wang et al. 2023), expertise level (Ji et al. 2021a)) from multiple annotators, then treating these discrepancies as bias or noise to elimant for achieving aggregation to approximate a single "ground truth" prediction (Yan et al. 2014; Jinadu et al. 2023).

However, the reliability of this paradigm faces two fundamental challenges: (1) In subjective domains such as emotional or impression assessment, there often exists no absolute ground truth—making this aggregation itself questionable (Raykar et al. 2010; Snow et al. 2008). (2) In real-world crowdsourcing, each annotator labels only a small fraction of the data, with most samples receiving annotations from

*Corresponding author.

different, often disjoint annotator subsets. This sparse and fragmented coverage makes aggregation statistically unreliable, as there is insufficient overlap to establish robust consensus patterns (Nørregaard and Derczynski 2022; Zhang et al. 2025d).

Therefore, we argue for a shift in focus, i.e., from sample-wise to annotator-wise (Figure 1): instead of treating sample-wise annotator disagreements as noise to be averaged away, we model annotator-wise behavior patterns as valuable information to capture consistent differences in judgment arising from expertise or preference. By longitudinally learning annotator-specific models across their labeled samples rather than aggregating multiple annotators per sample, we unlock three advantages: (1) Cost reduction: reconstructing unlabeled data for more annotation coverage; (2) Enhanced reliability: aggregating over sufficient reconstructed coverage yields a more robust consensus than fragmented annotations; (3) Behavioral insights: understanding the behavior patterns underlying annotators' decisions to explain disagreement sources.

Existing research on multi-annotator behavior pattern modeling with explainable analysis remains limited. While some works model individual annotators through various techniques, e.g., MaDL's confusion matrices (Herde, Huseljic, and Sick 2023) or PADL's Gaussian distributions (Liao et al. 2024), their aggregation-oriented mechanisms (e.g., PADL's meta-learning, MaDL's joint optimization) average annotator perspectives, losing individual information, and influencing individual behavior modeling. Conversely, completely independent modeling preserves individuality but suffers from overfitting on sparse labels. Moreover, existing models either lack explainability (Chen et al. 2019) or only implicitly reveal annotator importance (Cheng et al. 2023).

Our key insight is that annotators, despite their individual differences, often share behavioral structures. By capturing inter-annotator correlations, we can leverage the collective patterns as implicit regularization, constraining individual models from overfitting while preserving unique characteristics. To this end, we propose a novel query-based architecture QuMAB (**Qu**ery-based **M**ulti-**A**nnotator **B**ehavior Pattern Learning), hypothesize that annotator judgment differences arise from their varying degrees of focus on different regions of the input content (e.g., focusing on different image patches). Each annotator is represented by a learnable query that interacts with input features via cross-attention to effectively model individual behavior patterns. Lightweight query significantly reduces computational cost compared to separate conventional models.

Crucially, all annotator queries also interact through shared self-attention to capture inter-annotator correlations as a form of implicit structural regularization. This constrains inter-annotator representations to follow similarity patterns derived from annotations, preventing individual representations from drifting too far from the group and promoting mutual enhancement. This mechanism prevents overfitting to sparse individual data while maintaining individualization, improving generalization and robustness in individual annotator modeling, particularly under sparse annotations. Additionally, the cross-attention weights provide a visualization of annotator focus regions, offering an explainable analysis of behavior understanding (Zhang et al. 2025c).

Furthermore, we contribute two new large-scale datasets with dense per-annotator labels: STREET (city impression assessment, 4,300 labels/annotator) and AMER (video emotion recognition, average 3,118 labels/annotator). These datasets provide a high-value longitudinal annotation perspective for understanding and evaluating individual annotator behavior patterns, offering valuable data to the community and further researchers. It is worth noting that AMER is the first multi-annotator multimodal dataset in this field. Our work makes the following contributions:

- **A paradigm focus shift in multi-annotator learning**: We introduce a paradigm shift from sample-wise consensus aggregation to annotator-wise behavior modeling. By treating annotator disagreements as valuable information rather than noise, modeling annotator-specific behavior patterns can reconstruct unlabeled data to reduce annotation cost, enhance aggregation reliability, and explain annotator behavior.
- **A novel query-based architecture**: We propose QuMAB, which uses lightweight queries to model individual annotators while capturing inter-annotator correlations as implicit regularization, preventing overfitting to sparse individual data while maintaining individualization and improving generalization, with a visualization of annotator focus regions offering an explainable analysis of behavior understanding.
- **Two new large-scale datasets**: We contribute STREET (4,300 labels/annotator) and AMER (average 3,118 labels/annotator) datasets with denser per-annotator labels than existing resources, offering a longitudinal perspective for understanding individual annotator behaviors. AMER is the first multimodal multi-annotator dataset.

Related Work

Multi-annotator Behavior Modeling Paradigm

To the best of our knowledge, the multi-annotator behavior modeling paradigm problem has not yet been investigated. Traditional multi-annotator learning focuses on estimating consensus or ground-truth labels from multiple noisy annotations. These include early probabilistic models (Dawid and Skene 1979), EM algorithms (Whitehill et al. 2009), Gaussian models (Rodrigues, Pereira, and Ribeiro 2014), and biased estimation (Welinder et al. 2010). Tanno et al. (Tanno et al. 2019b) proposed modeling annotator confusion matrices as learnable parameters in neural networks. Cao et al. (Cao et al. 2019) introduced max-MIG to learn from multiple annotators. NEAL (Chen et al. 2023) employs neural expectation-maximization to jointly learn annotator expertise and true labels. Later methods used probabilistic frameworks to aggregate multiple annotations into a consensus or ground-truth label by confusion matrix (Tanno et al. 2019a), agreement distribution (Wang et al. 2023), and Gaussian distributions (Liao et al. 2024). This sample-wise

Dataset	Dataset description	Modality	# samples per annotator
QUBIQ-kidney (Menze et al. 2020)	kidney image	image	24
QUBIQ-tumor (Menze et al. 2020)	brain tumor image	image	32
QUBIQ-growth (Menze et al. 2020)	brain growth image	image	39
QUBIQ-prostate (Menze et al. 2020)	prostate image	image	55
CIFAR-10H (Peterson et al. 2019b)	object recognition	image	200
MUSIC (Rodrigues, Pereira, and Ribeiro 2013a)	music genre classification	audio	2∼368
MURA (Rajpurkar et al. 2017)	radiographic image	image	556
RIGA (Almazroa et al. 2017)	retinal cup and disc segmentation	image	750
LIDC-IDRI (Armato III et al. 2011)	lung nodule image	image	1,018
STREET (Ours)	city impression evaluation	image	4,300
AMER (Ours)	video emotion recognition	audio, video, text	970∼5,202

Table 1: Dataset comparison. Compared to existing datasets, our datasets contain a greater number of samples annotated by each annotator, helping promote multi-annotator behavior pattern modeling. AMER is the first multimodal multi-annotator dataset.

aggregation paradigm often treats annotator disagreements as noise to be averaged away rather than valuable information (Yan et al. 2014; Jinadu et al. 2023). In contrast, our introduced annotator-wise modeling paradigm treats annotator disagreements as valuable information for modeling annotator-specific behavior patterns, enhancing aggregation reliability, and explaining annotator behavior (Zhang et al. 2025a).

Multi-annotator Behavior Modeling Architecture

Previous studies model individual annotators in various techniques: D-LEMA (Mirikharaji et al. 2021) trains on non-contradictory subsets with spatial weights, PADL (Liao et al. 2024) infers Gaussian preference distributions, MaDL (Herde, Huseljic, and Sick 2023) jointly learns ground-truth and annotator embeddings, and embedding-based architectures (Gordon et al. 2022) or multi-task heads (Davani, Díaz, and Prabhakaran 2022). However, their aggregation-oriented mechanisms compromise individual behavior modeling by averaging annotator perspectives: D-LEMA dilutes annotator-specific patterns, PADL converges individual distributions, and MaDL smooths behaviors to minimize consensus loss.

Meanwhile, existing efforts on explainable analysis of annotator behavior understanding remain limited. Some works provide insights, e.g., TAX (Cheng et al. 2023) associates convolutional kernels with prototype libraries for pixel-level annotation decisions, MAGI (Zhang et al. 2023b) leverages annotator explanations to address noisy annotations, and Schaekermann et al. (Schaekermann et al. 2019) analyze factors contributing to disagreements. However, they only reveal annotators' trends in aggregation or analyze isolated factors without behavioral analysis. In contrast, our method models individual annotators via lightweight queries, leveraging inter-annotator correlations as regularization against overfitting while preserving individualization, with attention visualization analyzing behavioral patterns.

Multi-annotator Datasets

Most existing multi-annotator datasets often have only a small subset of samples with consecutive annotations from consistent annotator IDs. CIFAR-10H (Peterson et al. 2019a), based on the CIFAR-10 (Krizhevsky 2009) dataset, includes 10,000 test samples labeled by 2,571 annotators, but each annotator ID has on average only about 200 consecutive labels. LabelMe (Rodrigues and Pereira 2018) includes an average of approximately 42.4 consecutive labels per annotator ID. Audio dataset Music (Rodrigues, Pereira, and Ribeiro 2013b) contains an average of about 46.1 consecutive labels. The medical datasets are commonly used in multi-annotator studies, and consecutive annotator labels are even sparser. QUBIQ (Ji et al. 2021b), a dataset for quantifying uncertainty in biomedical image segmentation, includes four distinct segmentation datasets with an average of only 40 samples, and even then, annotator IDs have only around 8 consecutive labeled samples each. For longitudinal annotator behavior understanding, we contribute two new large-scale datasets with dense per-annotator labels: STREET (city impression assessment, 4,300 labels/annotator) and AMER (video emotion recognition, average 3,118 labels/annotator). AMER is the first multimodal multi-annotator dataset.

Dataset Construction

We contribute two new large-scale datasets: STREET (city impression assessment) and AMER (multi-modal emotion recognition) in this paper. Table 1 compares the current multi-annotator datasets. We observe that in existing datasets, the number of samples annotated by each annotator is relatively small, and there is a lack of multi-annotator multimodal datasets. For example, in the RIGA dataset (Almazroa et al. 2017), each annotator labels 750 samples, while in the CIFAR-10H dataset (Peterson et al. 2019b), each annotator labels 200 samples.

(1) STREET is an urban perception dataset with multi-annotators, which contains $4,300$ high-resolution images covering various urban elements, such as streets, public spaces, and infrastructure. The images were captured during a series of city strolling surveys, which aim to analyze emotions in relation to various factors associated with the city. The surveys were conducted by an organization to which one of our co-authors belongs. Voluntary participants walked around their own familiar city, took photos of various factors that may affect their subjective feelings (i.e.,

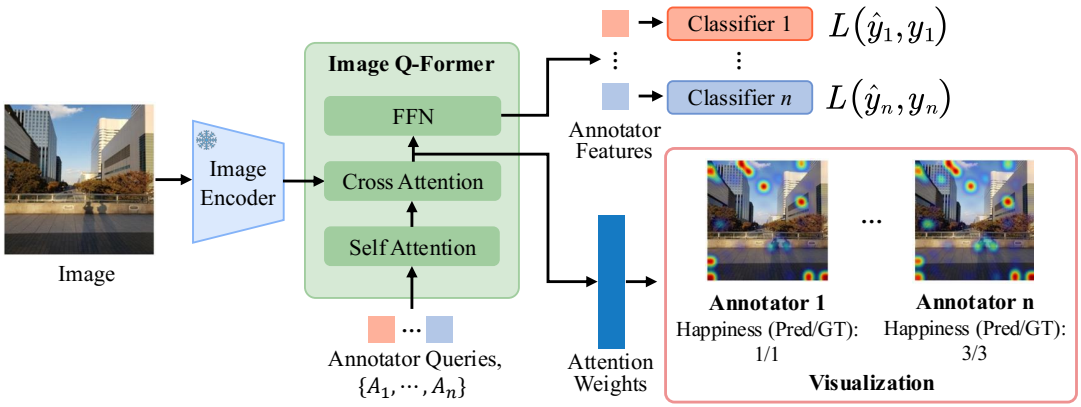

Figure 2: QuMAB for image. A frozen pre-trained image encoder extracts features, which interact with annotator-specific learnable queries in the Image Q-Former through cross-attention, producing annotator-specific features for classification. Different annotators' cross-attention weights reflect the differences in the image patches they focus on and support visual interpretation.

happiness/health), and assigned labels related to these feelings to each image (though we do not use these labels but ones assigned by crowd workers in our experiments). Thirteen survey sessions were conducted in five different cities (three urban areas and two suburban areas). A total of 327 participants, ranging in age from their 10s to 60s, took part in the survey. Each session lasted about one hour, with each participant taking an average of 12.6 photos.

We outsourced the annotation process to a company, which selected 10 annotators with balanced age, gender, and location diversity on a platform similar to Amazon Mechanical Turk, assessing five perception dimensions: happiness, healthiness, safety, liveliness, and orderliness, using a 6-point scale (-3 to $+3$). Each annotator spent approximately three weeks on their annotations. This multi-annotator dataset provides comprehensive human perception data for urban environments, enabling the quantitative analysis of environmental features and their emotional impact.

(2) AMER is a multimodal emotion dataset; the raw data is sourced from MER2024 (Lian et al. 2024), which contains $5,207$ video samples from movies and TV series, with multi-annotator emotion labels. Each sample typically contains one person, with relatively complete speech content. We annotated AMER using the open-source software Label Studio (Tkachenko et al. 2020-2025). We hired 15 annotators, who were students of our co-author's institution, and underwent a training session with 10 samples. We retained 13 annotators after screening out careless and irresponsible ones. Each annotator completed the task in approximately two weeks, with scheduled breaks to maintain annotation quality, where each annotator selects the most likely label from 8 candidate labels, i.e., worry, happiness, neutrality, anger, surprise, sadness, other, and unknown. Among all annotators, 10 annotators show consistent participation, each providing approximately 970 to $1,096$ labels, while the remaining 3 annotators contribute over $5,000$ labels each. This rich multi-annotator setup provides reliable emotion annotation results and allows for a robust evaluation of emotion recognition performance.

Methodology

We propose QuMAB, a query-based architecture designed to model the behavior patterns of individual annotators. We take the image classification task to illustrate our image-specific architecture for image inputs from the STREET dataset, which consists of a frozen image encoder, annotator-specific learnable queries, an image Q-Former, and annotator-wise classifiers, as shown in Figure 2. For video inputs from the AMER dataset, we describe a video-specific architecture in the supplementary material (Zhang et al. 2025b).

Given an image input $I \in \mathbb{R}^{H \times W \times 3}$, a frozen pre-trained image encoder (Sun et al. 2023) first extracts image features, which are then fed into the Image Q-Former (Li et al. 2023). For each annotator A_k ($k = 1, \ldots, n$), we assign a learnable query token for modeling. These queries are first processed by a shared self-attention layer to allow them interact with each other to implicitly capture inter-annotator correlations, and then interact with image features through multi-head cross-attention (typically with 12 heads), enabling each query to access diverse attention perspectives and produce individualized representations that reflect each annotator's potential focus and decision process. The resulting representations are mapped through a fully connected layer and passed to each annotator's classifier for prediction, e.g., for the same image on the Happiness dimension, A_1 predicts score (Pred) 1 while A_n predicts score (Pred) 3, both matching their respective ground-truth (GT) scores.

This query-based design is motivated by the hypothesis that annotator judgment differences arise from their varying degrees of focus on different regions of the input content (e.g., focusing on different image patches). Through learnable queries and cross-attention, our model effectively captures these individualized behavior patterns. Additionally, representing each annotator with a lightweight query significantly reduces computational cost compared to separate conventional models.

Meanwhile, the mechanism of capturing inter-annotator correlations acts as a form of implicit structural regulariza-

tion. This constrains inter-annotator representations to follow similarity patterns derived from annotations, preventing individual representations from drifting too far from the group and promoting mutual enhancement. It also prevents overfitting to annotator-specific noise while preserving individual differences of behavior patterns. As a result, the model achieves improved generalization and robustness in individual annotator modeling, particularly under sparse annotations.

For qualitative understanding, we visualize the cross-attention weights from the Image Q-Former to reveal annotator-specific focus regions. These weights indicate which image patches different annotators focus on when making predictions. Figure 2 illustrates results on the STREET dataset (city impression classification), annotators exhibit distinct spatial focus: *annotator n* attends more strongly to the two people holding hands in the center. This contrast reflects how annotators may interpret emotional cues differently based on their focus: *annotator n* assigns a higher score to the "happiness" dimension compared to *annotator 1*, suggesting that variations in focus on semantically positive regions may contribute to their differing judgments.

Loss Function

Finally, as shown in Figure 2, the total training loss $\mathcal{L}_{\text{total}}$ for the proposed multi-annotator classification model, QuMAB, is defined as the sum of individual cross-entropy losses for each annotator:

$$\mathcal{L}_{\text{total}} = \sum_{k=1}^{n} \mathcal{L}(\hat{y}_k, y_k), \tag{1}$$

where each annotator A_k has a specific predicted probabilities $\hat{y}_k \in [0,1]^C$, a reference label $y_k \in \{0,1\}^C$ in one-hot vector representation, and C is the number of classes.

Experiment

We conduct extensive experiments to evaluate our QuMAB, including modeling individual annotators' behavior patterns, assessing their utility for consensus prediction, testing applicability under sparse annotations, and complementing with qualitative visualization analysis. We compare against three representative baselines: D-LEMA (Mirikharaji et al. 2021), an ensemble-based multi-annotator learner; PADL (Liao et al. 2024), which fits Gaussian distributions for each annotator; and MaDL (Herde, Huseljic, and Sick 2023), which models annotator-specific confusion matrices. To ensure our model captures annotator-specific patterns rather than shared encoder features, we also include a Base variant with only the encoder and classifiers. Results are reported on two novel datasets. Accuracy and F_1 score (Zhan et al. 2019) are used as evaluation metrics. Note that additional experiments, including model efficiency, a faithfulness-oriented interpretability analysis, extended results, and further discussion, are provided in the supplementary material.

Implementation Details

Our image-specific model pipeline (Figure 2) uses ViT-G/14 from EVA-CLIP (Sun et al. 2023) as the encoder, with Image Q-Former initialized from InstructBLIP (Dai et al. 2023) (Frame Q-Former is the same in a video-specific pipeline from supplementary material, where Video Q-Former is initialized from Video-LLaMA (Zhang, Li, and Bing 2023)). Input images and video frames are resized to 224×224 and normalized. The number of query tokens is set equal to the number of annotators, and each annotator's classifier model uses an MLP. We train the model using the AdamW optimizer with an initial learning rate of 1e-4, a weight decay of 0.01, and gradient clipping with a maximum norm of 1.0. A linear warmup strategy is applied for the first 20% steps followed by cosine learning rate decay. The model is trained for up to 200 epochs with early stopping (patience = 25) to avoid overfitting. Training is conducted using distributed data parallelism (DDP) on four NVIDIA V100 GPUs.

Evaluation Metrics

To evaluate the performance of individual annotator modeling and consensus prediction (majority-votes the predictions of multi-annotators), we use accuracy (a standard metric in the multi-annotator learning) and F_1 score (Zhan et al. 2019) balancing precision and recall, suitable for potential class imbalance from uneven annotation densities, as in AMER (1,040 vs. 5,195 labels for annotators 1–10 vs. 11–13).

Quantitative Results

We analyze evaluation results on modeling individual annotators' behavior patterns, their utility for consensus application, and applicability under sparse annotation scenarios.

Individual Annotator Modeling. Individual annotator modeling aims to capture the behavior patterns of different annotators. Results in Tables 2 and 3 show that our method consistently outperforms all baselines in both accuracy and F_1 score (See supplementary material for F_1 results on STREET dataset) across individual annotators on the STREET and AMER datasets. This validates the superiority of our approach in capturing individual annotator behavior patterns.

Consensus Application Benefits. Real-world applications often seek a single consensus label despite subjectivity among annotators. We evaluate consensus prediction to validate whether modeling individual annotators preserves valuable information for practical needs. As no definitive ground truth exists, we use majority vote over raw annotations as a proxy, acknowledging its potential biases. Existing baselines adopt different aggregation strategies: D-LEMA learns weighted fusion; PADL applies meta-learning; and MaDL jointly optimizes consensus and annotator classifiers. They may average annotator perspectives during training, potentially diminishing individual nuances. For a fair comparison, we apply unified majority voting over annotator-specific predictions from all methods rather than using their original aggregated outputs. Results (CoPr) in Tables 2, 3, and F_1 results show that our method achieves superior consensus performance, suggesting that modeling individual annotators helps retain valuable information, potentially benefiting real-world consensus applications. *Note:* This experiment serves to validate practical utility of individual annotator modeling rather than to assert overall superiority.

Metric	Methods	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	Avg	CoPr
ACC	Base	0.30	0.29	0.43	0.25	0.24	0.20	0.26	0.19	0.34	0.10	0.41	0.48	0.19	0.28	0.35
	D-LEMA	0.86	0.88	0.85	0.87	0.89	0.86	0.88	0.87	0.85	0.86	0.45	0.51	0.33	0.78	0.55
	PADL	0.89	0.90	0.88	0.93	0.87	0.91	0.86	0.94	0.89	0.88	0.47	0.54	0.35	0.79	0.52
	MaDL	0.93	0.91	0.90	0.89	0.90	0.88	0.90	0.89	0.87	0.92	0.50	0.53	0.37	0.80	0.57
	Ours	**0.94**	**0.93**	**0.93**	**0.94**	**0.94**	**0.92**	**0.93**	**0.95**	**0.93**	**0.93**	**0.59**	**0.61**	**0.40**	**0.84**	**0.60**
F_1	Base	0.26	0.27	0.40	0.22	0.23	0.17	0.24	0.18	0.31	0.08	0.35	0.41	0.14	0.25	0.32
	D-LEMA	0.84	0.87	0.81	0.84	0.86	0.85	0.86	0.82	0.83	0.84	0.38	0.44	0.27	0.73	0.52
	PADL	0.86	0.88	0.85	0.91	0.83	0.89	0.82	0.92	0.85	0.86	0.41	0.50	0.29	0.76	0.49
	MaDL	0.90	0.85	0.87	0.86	0.87	0.85	0.87	0.86	0.84	0.92	0.45	0.48	0.33	0.77	0.54
	Ours	**0.91**	**0.91**	**0.90**	**0.92**	**0.91**	**0.90**	**0.89**	**0.93**	**0.91**	**0.93**	**0.54**	**0.55**	**0.34**	**0.81**	**0.57**

Table 2: The accuracy (ACC) and F_1 score evaluate results on the AMER dataset. We assess performance for individual annotator modeling (each annotator A_k, $k = 1, \ldots, 13$), the average (Avg), and consensus prediction (CoPr). Higher is better.

Perspectives	Methods	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	Avg	CoPr
Happiness	Base	0.80	0.12	0.27	0.38	0.56	0.35	0.44	0.44	0.31	0.55	0.42	0.45
	D-LEMA	0.85	0.71	0.44	0.36	**0.70**	0.43	0.46	0.54	0.41	0.47	0.54	0.57
	PADL	0.93	0.74	0.48	0.53	0.57	0.47	0.42	0.51	0.50	0.60	0.58	0.55
	MaDL	0.91	0.77	0.44	0.38	**0.70**	0.47	0.46	**0.54**	0.48	0.47	0.56	0.58
	Ours	**0.94**	**0.80**	**0.54**	**0.55**	0.69	**0.51**	**0.53**	**0.54**	**0.52**	**0.64**	**0.63**	**0.62**
Healthiness	Base	0.77	0.19	0.14	0.47	0.87	0.36	0.43	0.44	0.51	0.54	0.47	0.56
	D-LEMA	0.83	**0.77**	0.44	0.43	0.87	0.41	0.44	0.46	0.55	0.46	0.57	0.54
	PADL	**0.92**	0.72	0.55	0.44	0.84	0.47	0.46	0.44	0.52	0.55	0.59	0.50
	MaDL	0.89	**0.77**	0.44	0.44	**0.90**	0.41	0.44	0.46	0.55	0.46	0.58	0.55
	Ours	**0.92**	0.75	**0.56**	**0.52**	**0.90**	**0.49**	**0.48**	**0.54**	**0.64**	**0.61**	**0.64**	**0.58**
Safety	Base	0.58	0.65	0.36	0.36	0.61	0.37	0.56	0.40	0.32	0.53	0.47	0.51
	D-LEMA	0.62	0.69	0.27	0.41	0.50	0.46	0.48	0.40	0.35	0.50	0.47	0.49
	PADL	**0.72**	0.78	0.24	0.44	0.69	0.44	**0.53**	0.42	0.46	0.48	0.52	0.54
	MaDL	0.63	0.63	0.27	0.32	0.61	0.38	0.46	0.42	0.36	0.52	0.46	0.56
	Ours	**0.72**	**0.80**	**0.38**	**0.48**	**0.71**	**0.54**	**0.53**	**0.50**	**0.52**	**0.58**	**0.58**	**0.61**
Liveliness	Base	0.79	0.60	0.53	0.46	0.76	0.30	0.35	0.36	0.53	0.57	0.53	0.55
	D-LEMA	0.79	0.58	0.37	0.42	0.74	0.38	0.44	0.41	0.50	0.46	0.51	0.53
	PADL	0.85	0.66	0.56	0.46	0.75	0.44	0.43	0.47	0.56	0.57	0.58	0.54
	MaDL	0.78	0.56	0.35	0.40	0.76	0.34	**0.48**	0.42	0.47	0.47	0.50	0.56
	Ours	**0.87**	**0.68**	**0.57**	**0.53**	**0.80**	**0.49**	**0.48**	**0.51**	**0.62**	**0.61**	**0.62**	**0.59**
Orderliness	Base	0.50	0.64	0.39	0.45	0.86	0.31	0.39	0.34	0.31	0.49	0.47	0.52
	D-LEMA	0.55	0.60	0.32	0.36	0.82	0.39	0.42	0.36	0.37	0.47	0.47	0.57
	PADL	0.73	0.65	0.44	0.45	0.93	0.45	0.45	0.36	0.42	**0.63**	0.55	0.54
	MaDL	0.61	0.60	0.34	0.36	0.86	0.37	0.47	0.36	0.37	0.49	0.48	0.58
	Ours	**0.74**	**0.71**	**0.52**	**0.55**	**0.94**	**0.47**	**0.54**	**0.44**	**0.56**	0.62	**0.61**	**0.62**

Table 3: The accuracy metric is to evaluate results on the STREET dataset. We assess performance for individual annotator modeling (each annotator A_k, $k = 1, \ldots, 13$), the average (Avg), and consensus prediction (CoPr). Higher is better.

Method	S-Ha	S-He	S-Sa	S-Li	S-Or	AMER
Full-PADL	0.58	0.59	0.52	0.58	0.55	0.79
Full-Ours	0.63	0.64	0.58	0.62	0.61	0.84
Sparse-PADL	0.43	0.42	0.38	0.41	0.40	0.58
Sparse-Ours	**0.52**	**0.51**	**0.46**	**0.49**	**0.48**	**0.66**

Table 4: Evaluation by accuracy for sparse scenarios (40% of annotations are randomly removed). S-Ha, S-He, S-Sa, S-Li, and S-Or represent five perspectives of STREET dataset: happiness, healthiness, safety, liveliness, and orderliness.

Applicability under Sparse Annotations. To evaluate our model's applicability under sparse annotation scenarios, we simulated real-world conditions by randomly removing annotations at various rates. As shown in Table 4, when 40% of annotations are removed (See supplementary material for results of more sparse rates), our model's average performance drops by 20.4%, whereas the best baseline PADL experiences a larger drop of 27.4%. Results suggest that our superiority stems from modeling inter-annotator correlations, which regularizes individual annotator representations, preventing overfitting to sparse labels and promoting consistency with shared patterns across annotators, to enhance ro-

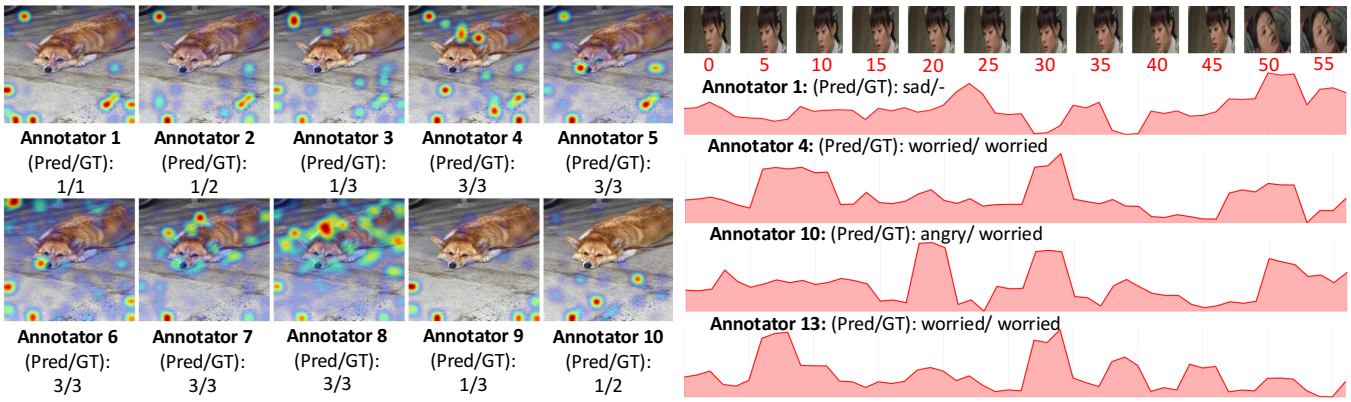

Figure 3: Visualization Analysis: **(Left)** Different image patches annotators focused in STREET dataset (healthiness perspective), *annotators 4, 5, 6, 7, and 8* exhibit centralized focuses on a cute dog compared to other annotators; **(Right)** Different video frames annotators focused in AMER dataset. The *annotator 1* exhibits focus on final frames containing a different person, while *annotators 4 and 13* focus on middle frames. Red number indicates frame ID; Pred/GT denotes prediction/ground-truth.

bustness and generalization under sparse annotations.

Qualitative Results

We qualitatively analyze how the learned attention reflects annotators' behavior patterns by visualizing cross-attention weights from Q-Former. These weights highlight the image patches or video frames that different annotators may focus on when making predictions.

As shown in Figure 3 (Left), on the STREET dataset (city impression classification: healthiness perspective), annotators exhibit distinct spatial focus: *annotators 4, 5, 6, 7, and 8* focus more centrally on a dog in the image, while others do not. Correspondingly, these annotators also provided higher scores for the healthiness perspective, e.g., *annotator 8* predicts a high score (Pred) 3 while *annotator 9* predicts a low score 1, suggesting that variations in focus on specific semantics may contribute to their differing judgments.

Figure 3 (Right) illustrates results on the AMER dataset (video emotion classification). Annotators differ in temporal focus: *annotator 1* focuses on the final frames (45–50), while *annotator 4 and 13* concentrate early frames (5–10). These patterns align with their predictions: "sad" for *annotator 1*, "worried" for *annotator 4 and 13*, suggesting that differences in people or dialogues in the corresponding frame segments they focused on may underlie decision diversity.

Ablation Study

We conduct an ablation study to validate our architectural design choices and assess the effect of individual behavior pattern modeling on consensus prediction (Table 5).

Architecture Choices. Removing Q-Former yields the Base model, and replacing individual classifiers with a unified classifier (w/ U-cls) both lead to significant performance drops, validating their effectiveness in our architecture.

Inter-Annotator Correlation Analysis. Disabling self-attention (w/o S-Attn) leads to clear performance degradation, underscoring the role of inter-annotator correlations as an implicit structural regularizer that improves generalization and robustness of individual annotator modeling.

Method	S-Ha	S-He	S-Sa	S-Li	S-Or	AMER
Base	0.42	0.47	0.47	0.53	0.47	0.28
w/ U-cls	0.50	0.56	0.49	0.56	0.49	0.61
w/o S-Attn	0.52	0.54	0.46	0.53	0.56	0.73
Pre-mv	0.58	0.57	0.56	0.58	0.60	0.43
Post-mv	0.62	0.58	0.61	0.59	0.62	0.60
Full (avg)	0.63	0.64	0.58	0.62	0.61	0.84

Table 5: Ablation study. The average performance (accuracy) of replacing modules, removing inter-annotator correlations, and a consensus prediction comparison with and without individual annotators' behavior pattern modeling.

Individual Behavior Modeling for Consensus. To investigate the role of individual behavior modeling in consensus prediction, we compare two strategies: (1) Pre-mv performs majority voting before modeling,i.e., applies global pooling over all Q-Former queries for a single prediction; (2) Post-mv first models each annotator individually and then aggregates their predictions. Post-mv consistently outperforms Pre-mv, especially on AMER. This demonstrates that modeling individual annotators' behavior patterns preserves valuable information otherwise lost in early aggregation, potentially benefiting real-world consensus prediction or other practical applications.

Conclusion

This paper introduced a paradigm focus shift in multi-annotator learning from sample-wise aggregation to annotator-wise behavior modeling, proposing a lightweight query-based architecture to model individual annotator patterns. We contributed the STREET and AMER datasets with dense per-annotator labels and demonstrate superior performance in behavior modeling, consensus prediction, and sparse annotation scenarios, offering a novel perspective on multi-annotator learning challenges.

Acknowledgments

This work was supported by JST Grant Number JPMJPF2115 and the Future Social Value Co-Creation Project, the University of Osaka. This work is also supported by the Excellent Youth Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (No. MAIS2024311) and Youth Science Fund Project of National Natural Science Foundation of China (No. 62201572).

References

Almazroa, A.; Alodhayb, S.; Osman, E.; Ramadan, E.; Hummadi, M.; Dlaim, M.; Alkatee, M.; Raahemifar, K.; and Lakshminarayanan, V. 2017. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. *International ophthalmology*, 37: 701–717.

Armato III, S. G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M. F.; Meyer, C. R.; Reeves, A. P.; Zhao, B.; Aberle, D. R.; Henschke, C. I.; Hoffman, E. A.; et al. 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2): 915–931.

Cao, P.; Xu, Y.; Kong, Y.; and Wang, Y. 2019. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436*.

Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Chen, J.; Zhang, R.; Xu, J.; Hu, C.; and Mao, Y. 2023. A Neural Expectation-Maximization Framework for Noisy Multi-Label Text Classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 10992–11003.

Cheng, Y.-C.; Shiau, Z.-Y.; Yang, F.-E.; and Wang, Y.-C. F. 2023. TAX: Tendency-and-Assignment Explainer for Semantic Segmentation with Multi-Annotators. *arXiv preprint arXiv:2302.09561*.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Dawid, A. P.; and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 20–28.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Herde, M.; Huseljic, D.; and Sick, B. 2023. Multi-annotator Deep Learning: A Probabilistic Framework for Classification. *arXiv preprint arXiv:2304.02539*.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021a. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12336–12346.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021b. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12341–12351.

Jinadu, U.; Annan, J.; Wen, S.; and Ding, Y. 2023. Loss Modeling for Multi-Annotator Datasets. *arXiv preprint arXiv:2311.00619*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Lian, Z.; Sun, H.; Sun, L.; Chen, K.; Xu, M.; Wang, K.; Xu, K.; He, Y.; Li, Y.; Zhao, J.; et al. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9610–9614.

Lian, Z.; Sun, H.; Sun, L.; Wen, Z.; Zhang, S.; Chen, S.; Gu, H.; Zhao, J.; Ma, Z.; Chen, X.; et al. 2024. MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary Multimodal Emotion Recognition. In *MRAC'24: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 41–48.

Liao, Z.; Hu, S.; Xie, Y.; and Xia, Y. 2024. Modeling annotator preference and stochastic annotation error for medical image segmentation. *Medical Image Analysis*, 92: 103028.

Menze, B.; Joskowicz, L.; Bakas, S.; Jakab, A.; Konukoglu, E.; Becker, A.; and et al. 2020. Quantification of uncertainties in biomedical image quantification challenge. https://qubiq.grand-challenge.org/.

Mirikharaji, Z.; Abhishek, K.; Izadi, S.; and Hamarneh, G. 2021. D-lema: Deep learning ensembles from multiple annotations-application to skin lesion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1837–1846.

Nørregaard, J.; and Derczynski, L. 2022. Sparse Probability of Agreement. *arXiv preprint arXiv:2208.06161*.

Peterson, J.; Battleday, R.; Griffiths, T.; and Russakovsky, O. 2019a. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019b. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9617–9626.

Rajpurkar, P.; Irvin, J.; Bagul, A.; Ding, D.; Duan, T.; Mehta, H.; Yang, B.; Zhu, K.; Laird, D.; Ball, R. L.; et al. 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning From Crowds. *Journal of Machine Learning Research*, 11(43): 1297–1322.

Rodrigues, F.; and Pereira, F. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013a. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12): 1428–1436.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2013b. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters*, 34(12): 1428–1436.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian Process Classification and Active Learning with Multiple Annotators. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 433–441. Bejing, China: PMLR.

Schaekermann, M.; Beaton, G.; Habib, M.; Lim, A.; Larson, K.; and Law, E. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Lapata, M.; and Ng, H. T., eds., *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Honolulu, Hawaii: Association for Computational Linguistics.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019a. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11244–11253.

Tanno, R.; Saeedi, A.; Sankaranarayanan, S.; Alexander, D. C.; and Silberman, N. 2019b. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Wang, B.; Chang, J.; Qian, Y.; Chen, G.; Chen, J.; Jiang, Z.; Zhang, J.; Nakashima, Y.; and Nagahara, H. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. *Advances in neural information processing systems*, 37: 74999–75011.

Wang, C.; Gao, Y.; Fan, C.; Hu, J.; Lam, T. L.; Lane, N. D.; and Bianchi-Berthouze, N. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *International Workshop on Trustworthy Machine Learning for Healthcare*, 147–162. Springer.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J.; and Ruvolo, P. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22.

Yan, Y.; Rosales, R.; Fung, G.; Subramanian, R.; and Dy, J. 2014. Learning from multiple annotators with varying expertise. *Machine learning*, 95: 291–327.

Zhan, X.; Wang, Y.; Rao, Y.; and Li, Q. 2019. Learning from multi-annotator data: A noise-aware classification framework. *ACM Transactions on Information Systems (TOIS)*, 37(2): 1–28.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 543–553.

Zhang, L.; Ke, J.; Fan, S.; Sha, X.; and Lian, Z. 2025a. A Unified Evaluation Framework for Multi-Annotator Tendency Learning. *arXiv preprint arXiv:2508.10393*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025b. QuMAB: Query-based Multi-Annotator Behavior Modeling with Reliability under Sparse Labels. *arXiv preprint arXiv:2507.17653*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025c. QuMATL: Query-based Multi-annotator Tendency Learning. *arXiv preprint arXiv:2503.15237*.

Zhang, L.; Lian, Z.; Liu, H.; Takebe, T.; and Nakashima, Y. 2025d. SimLabel: Similarity-Weighted Semi-supervision for Multi-annotator Learning with Missing Labels. *arXiv preprint arXiv:2504.09525*.

Zhang, L.; Luo, Z.; Wu, S.; and Nakashima, Y. 2024. MicroEmo: Time-Sensitive Multimodal Emotion Recognition with Subtle Clue Dynamics in Video Dialogues. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 110–115.

Zhang, L.; Tanno, R.; Xu, M.; Huang, Y.; Bronik, K.; Jin, C.; Jacob, J.; Zheng, Y.; Shao, L.; Ciccarelli, O.; et al. 2023a. Learning from multiple annotators for medical image segmentation. *Pattern Recognition*, 138: 109400.

Zhang, Y.; Gu, S.; Gao, Y.; Pan, B.; Yang, X.; and Zhao, L. 2023b. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1977–1987.