QuMAB: Query-based Multi-Annotator Behavior Modeling with Reliability under Sparse Labels

Liyun Zhang
D3 Center, Osaka University
Japan

Zheng Lian
Institute of automation, Chinese academy of science
China

Hong Liu
Xiamen University
China

Takanori Takebe
Cincinnati Children's Hospital Medical Center
Japan

Yuta Nakashima
D3 Center, Osaka University
Japan

ABSTRACT

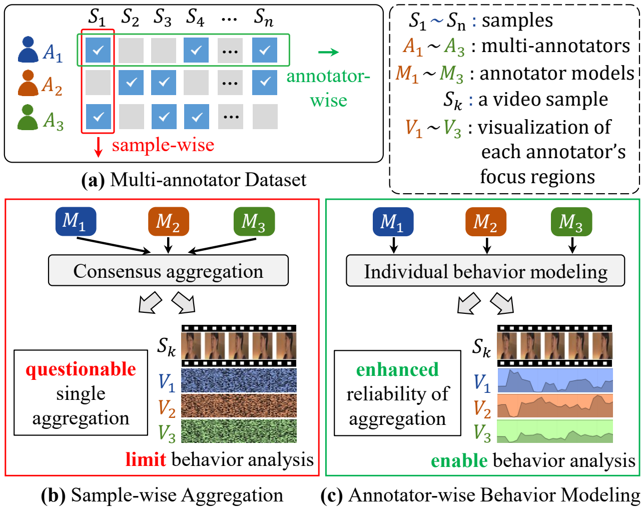
Multi-annotator learning traditionally aggregates diverse annotations to approximate a single "ground truth", treating disagreements as noise. However, this paradigm faces fundamental challenges: subjective tasks often lack absolute ground truth, and sparse annotation coverage makes aggregation statistically 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.

CCS CONCEPTS

• **Computing methodologies → Multi-task learning**.

KEYWORDS

Multi-annotator learning, Annotator tendency, Behavior patterns, Multi-annotator datasets


S_1 S_2 S_3 S_4 ⋯ S_n
A_1
A_2
A_3
→ annotator-wise
↓ sample-wise
$S_1 \sim S_n$: samples
$A_1 \sim A_3$: multi-annotators
$M_1 \sim M_3$: annotator models
S_k : a video sample
$V_1 \sim V_3$: visualization of each annotator's focus regions
M_1
M_2
M_3
Consensus aggregation
S_k
V_1
V_2
V_3
questionable single aggregation
limit behavior analysis
Individual behavior modeling
enhanced reliability of aggregation
enable behavior analysis

(a) Multi-annotator Dataset

(b) Sample-wise Aggregation

(c) Annotator-wise Behavior Modeling

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.

1 INTRODUCTION

In real-world multi-annotation scenarios, such as medical image analysis [17], sentiment analysis [15], and visual perception [50], different annotators often provide different labels to the same sample [24] due to different personal backgrounds, subjective interpretations, and preferences. Traditional multi-annotator learning focuses on learning different characteristics (e.g., confusion mode [33], agreement [36], expertise level [10]) from multiple annotators, then treating these discrepancies as bias or noise to elimant for achieving aggregation to approximate a single "ground truth" prediction [12, 39].

, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 [24, 31]. (2) In real-world crowdsourcing, each annotator labels only a small fraction of the data, with most samples receiving annotations from different, often disjoint annotator subsets. This sparse and fragmented coverage makes aggregation statistically unreliable, as there is insufficient overlap to establish robust consensus patterns [20].

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 propose modeling individual annotator behavior patterns as annotator-wise valuable information. These patterns capture consistent differences in judgment arising from expertise, preference, or perspective. By longitudinally learning reliable annotator-specific models—tracking each individual's behavior patterns across consecutive sample-label pairs per annotator rather than aggregating multiple annotators per sample—we unlock three compelling advantages: (1) **Cost reduction**: reconstructing unlabeled data enables comprehensive annotation coverage; (2) **Enhanced reliability**: aggregating over reconstructed sufficient coverage of sample-label pairs yields statistically more robust consensus than sparse, fragmented annotations; (3) **Behavioral insights**: understanding the behavior patterns underlying each annotator's decisions and explaining sources of disagreement.

Currently, existing research focusing on multi-annotator behavior pattern modeling and presenting explainable analysis of behavior understanding is sparse. Some works similar to it have attempted to model individual annotators through various techniques (e.g., MaDL's confusion matrices [9] or PADL's Gaussian distribution fitting [17]) to understand more about individual annotator patterns. However, their aggregation-oriented mechanism (e.g., PADL's meta-learning, and MaDL's jointly optimizing consensus and annotator classifiers) may average annotator perspectives to lose annotator-specific information, influencing individual behavior modeling. Otherwise, if completely independent modeling individual annotator, it preserves sufficient individual information but easily suffers from overfitting during the annotator model is trained on its small set of labels as described previously. Moreover, existing models [4] lack explainable analysis of behavior understanding, or they only implicitly reveal trends where certain annotators play a larger role in the prediction [6].

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.

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 per annotator) and AMER (average 3,118 labels per 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.

2 RELATED WORK

2.1 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 [8], EM algorithms [38], Gaussian models [28], and biased estimation [37]. Tanno et al. [34] proposed modeling annotator confusion matrices as learnable parameters in neural networks. Cao et al. [3] introduced max-MIG to learn from multiple annotators. SimLabel [44] addresses the practical challenge

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.

Dataset	Dataset description	Modality	# samples per annotator
QUBIQ-kidney [18]	kidney image	image	24
QUBIQ-tumor [18]	brain tumor image	image	32
QUBIQ-growth [18]	brain growth image	image	39
QUBIQ-prostate [18]	prostate image	image	55
CIFAR-10H [22]	object recognition	image	200
MUSIC [26]	music genre classification	audio	2~368
MURA [23]	radiographic image	image	556
RIGA [1]	retinal cup and disc segmentation	image	750
LIDC-IDRI [2]	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

of missing labels in multi-annotator scenarios. NEAL [5] 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 [33], agreement distribution [36], and Gaussian distributions [17]. This sample-wise aggregation paradigm often treats annotator disagreements as noise to be averaged away rather than valuable information [12, 39]. 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.

2.2 Multi-annotator Behavior Modeling Architecture

Previous studies have attempted to model individual annotators through various techniques: D-LEMA [19] trains annotator models on non-contradictory subsets with spatial weights for noise handling; PADL [17] models annotator preferences via Gaussian assumptions in its Human Preference Module (HPM) and employs a Sample Embedding Module (SEM) for meta-classification; MaDL [9] jointly optimizes ground truth classifiers and annotator models through weighted embeddings. However, their aggregation-oriented mechanisms compromise individual behavior modeling by averaging annotator perspectives: D-LEMA sacrifices annotator-specific patterns for fusion objectives, PADL forces individual distributions to converge at the cost of behavioral details, and MaDL smooths individual characteristics to minimize consensus loss. Meanwhile, existing efforts on explainable analysis of annotator behavior understanding remain limited. Some works provide insights, e.g., TAX [6] associates convolutional kernels with prototype libraries for pixel-level annotation decisions, MAGI [51] leverages annotator explanations to address noisy annotations, and Schaekermann et al. [29] analyze factors contributing to disagreements. However, they only reveal annotators' trends in aggregation or analyze isolated factors without behavioral analysis. In comparison, our architecture models individual annotators via lightweight queries, leveraging inter-annotator correlations as regularization against overfitting while preserving individualization, with attention visualization analyzing behavioral patterns.

2.3 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 [21], based on the CIFAR-10 [13] dataset, includes 10,000 test samples labeled by 2,571 annotators, but each annotator ID has on average only about 200 consecutive labels. LabelMe [25] includes an average of approximately 42.4 consecutive labels per annotator ID. Audio dataset Music [27] 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 [11], 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.

3 DATASET CONSTRUCTION

We contribute two new large-scale datasets[1]: 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 [1], each annotator labels 750 samples, while in the CIFAR-10H dataset [22], each annotator labels 200 samples.

[1] We are negotiating to publish both datasets after acceptance.

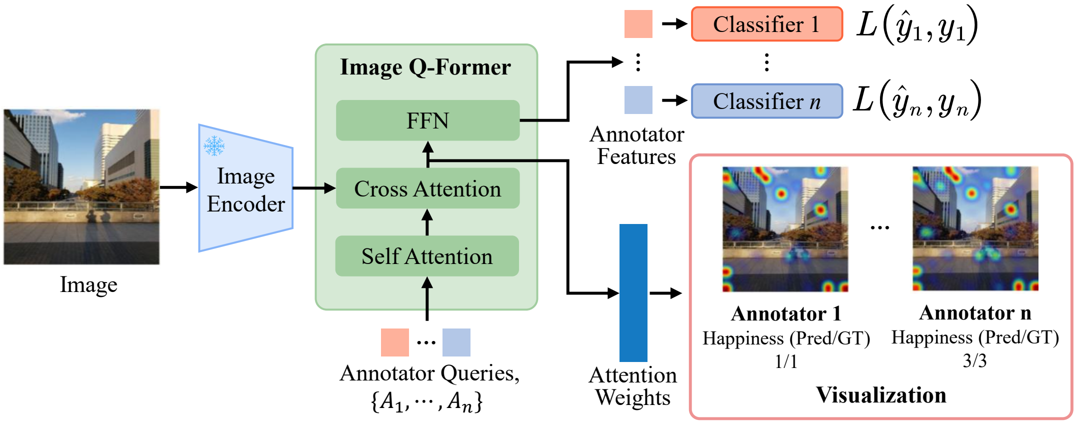


Figure 2: QuMAB for image. A frozen pre-trained image encoder extracts features, which interact with annotator-specific learnable queries in the Image Q-Former through cross-attention, producing annotator-specific features for classification. Different annotators' cross-attention weights reflect the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

(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., 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 [16], 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 [35]. 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.

4 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.

Given an image input $I \in \mathbb{R}^{H \times W \times 3}$, a frozen pre-trained image encoder [32] first extracts image features, which are then fed into the Image Q-Former [14]. 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.

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 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. 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.

4.1 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.

5 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 [19], an ensemble-based multi-annotator learner; PADL [17], which fits Gaussian distributions for each annotator; and MaDL [9], 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. We evaluate on two multi-annotator datasets: AMER (video-based emotion recognition with 13 annotators) and STREET (urban image impression assessment with 10 annotators across five dimensions: Happiness, Healthiness, Safety, Liveliness, and Orderliness). The AMER dataset's complexity in capturing temporal emotion dynamics aligns with recent advances in time-sensitive emotion recognition [43, 46]. These datasets provide dense, diverse annotations crucial for modeling annotator-specific behavior patterns. Accuracy and F_1 score [40] 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.

5.1 Implementation Details

Our image-specific model pipeline (Figure 2) uses ViT-G/14 from EVA-CLIP [32] as the encoder, with Image Q-Former initialized from InstructBLIP [7] (Frame Q-Former is the same in a video-specific pipeline from supplementary material, where Video Q-Former is initialized from Video-LLaMA [41]). 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.

5.2 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 [40] 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).

5.3 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 [2] 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

[2] See supplementary material for F_1 results on STREET dataset.

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.

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 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.

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**

all methods rather than using their original aggregated outputs. Results (CoPr) in Tables 2, 3, and F_1 results [2] 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.

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[3], our model's average performance drops by 20.4%, whereas the best baseline PADL experiences a larger drop of 27.4%.

[3] See supplementary material for results of more sparse rates.

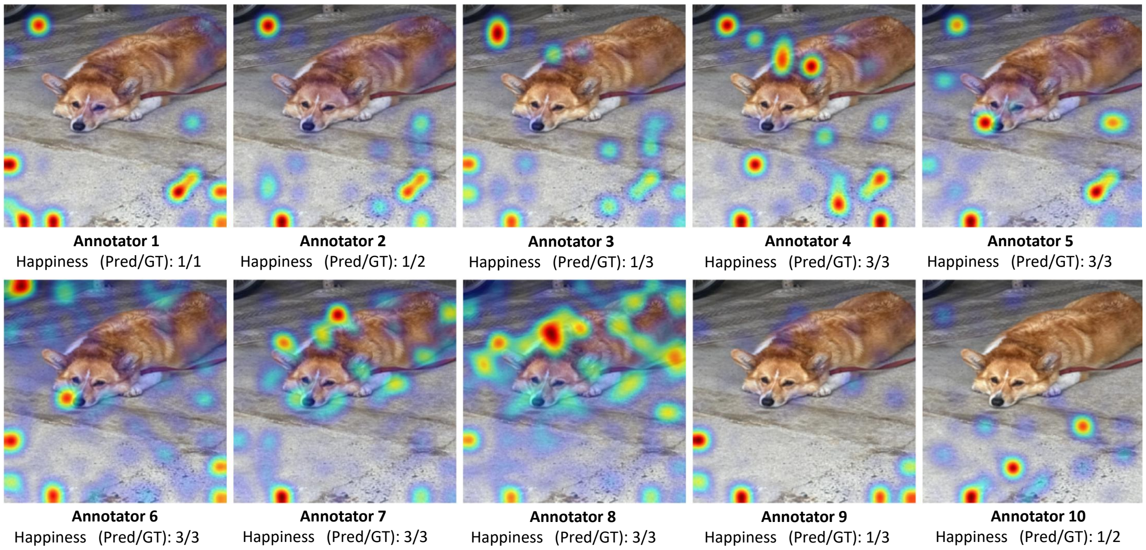


Figure 3: A visualization analysis of the different image patches that annotators focused on in the STREET dataset. The *annotators 4, 5, 6, 7, and 8* exhibit centralized focuses on a cute dog compared to other annotators.

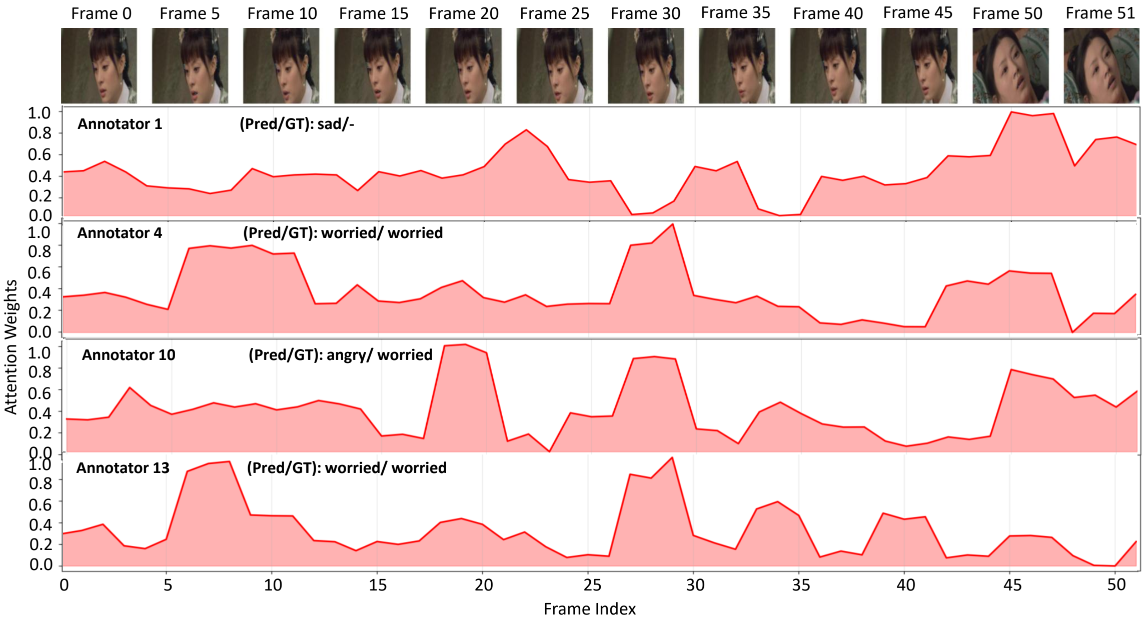


Figure 4: A visualization analysis of the different video frames that annotators focused on in the AMER dataset. The *annotator 1* exhibits focus on final frames (the frames contain different person), while *annotators 4 and 13* focus on the middle frames.

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 robustness and generalization under sparse annotations.

5.4 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, on the STREET dataset (city impression classification), annotators exhibit distinct spatial focus: *annotators 4,*

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.

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 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.

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

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 "happiness" dimension, suggesting that variations in focus on specific semantics may contribute to their differing judgments.

Figure 4 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 the people or dialogues in the corresponding frame segments they focused on may underlie decision diversity.

5.5 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.

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.

Table 6: Aggregation accuracy under simulated missing labels in the test set. Modeling annotators before majority voting (S-Ours-MV) outperforms direct voting (S-MV) and the best-performing baseline (S-PADL-MV), validating the benefit of annotator-wise modeling for reliable aggregation.

Method	S-Ha	S-He	S-Sa	S-Li	S-Or	AMER
20% missing rate						
S-MV	0.97	0.97	0.96	0.96	0.96	0.85
S-PADL-MV	0.97	**0.98**	**0.97**	0.96	**0.97**	0.86
S-Ours-MV	**0.98**	**0.98**	**0.97**	**0.98**	**0.97**	**0.89**
30% missing rate						
S-MV	0.94	0.93	0.92	0.93	0.92	0.85
S-PADL-MV	0.95	0.95	**0.95**	0.94	0.94	0.87
S-Ours-MV	**0.97**	**0.96**	**0.95**	**0.96**	**0.95**	**0.92**
40% missing rate						
S-MV	0.92	0.92	0.89	0.91	0.90	0.81
S-PADL-MV	0.93	0.93	0.91	0.92	0.91	0.83
S-Ours-MV	**0.95**	**0.95**	**0.93**	**0.94**	**0.93**	**0.88**

6 CONCLUSION

This paper introduced a paradigm shift in multi-annotator learning from sample-wise aggregation to annotator-wise behavior modeling, and proposed a lightweight query-based architecture (QuMAB) to effectively model individual annotator behavior patterns. We contributed the STREET and AMER datasets with dense per-annotator labels for longitudinal annotator behavior understanding. Experiments demonstrate the superiority of our method in modeling individual annotators' behavior patterns, their utility for consensus prediction, and applicability under sparse annotations. This work presents a novel view on recognizing and understanding the challenges in the multi-annotator learning field.

7 SUPPLEMENTARY MATERIAL

7.1 Overview

First, we provide an empirical analysis, elaborating on our first contribution in the paper, i.e., practical value of our paradigm shift from sample-wise aggregation to annotator-wise modeling, further clarifying our motivation and practical contribution. Subsequently, we illustrate the detailed video-specific architecture. Additionally, we present supplementary experimental analysis. The analysis of model efficiency and the discussion are given. Finally, the discussion of limitation and future work is provided.

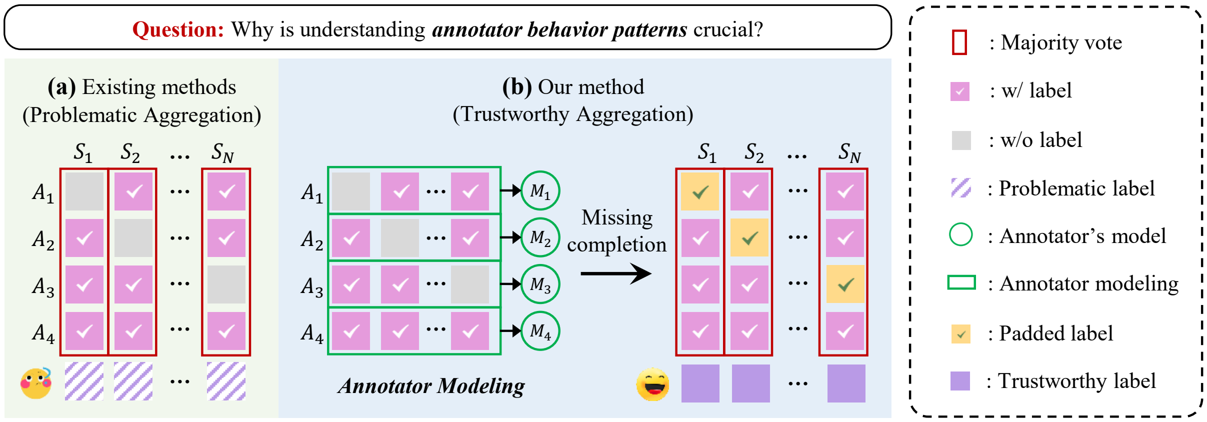


Figure 5: An illustration of paradigm shift from sample-wise aggregation to annotator-wise behavior modeling. (a) Existing methods overlook individual annotator information, potentially leading to suboptimal or biased consensus aggregation on real-world annotation matrices, where each annotator labels only a small subset of samples with disjoint coverage. (b) Our method captures each annotator's longitudinal behavior patterns across their labeled samples, improving consensus aggregation reliability by leveraging reconstructed unlabeled data.

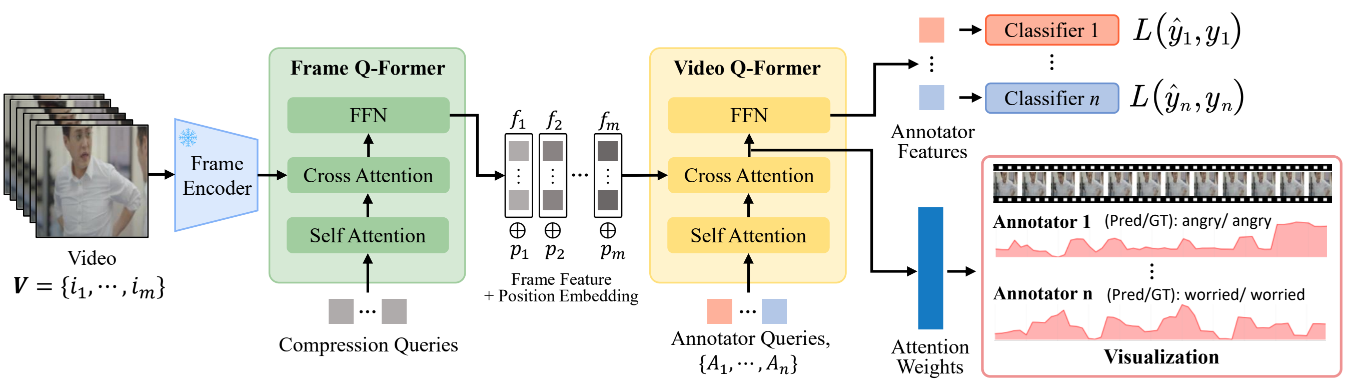


Figure 6: QuMAB architecture for video-specific pipeline. A frozen pre-trained encoder extracts frame features, which are compressed through the Frame Q-Former. With added frame position embedding, these interact with $1, \ldots, n$ annotator-specific learnable queries in the Video Q-Former through cross-attention, producing annotator-specific features for classification. Different annotators' cross-attention weights represent different annotator tendencies and provide visualization for analysis.

Table 7: Label statistics and missing rates of the AMER dataset with 13 annotators $A_i, i \in \{1, \cdots, 5\}$. For each annotator, the number of labeled samples, the corresponding missing rate (%), as well as the average data (Average), are provided.

Perspective	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}	Average
Sample size	1096	1031	1022	1036	1012	970	1064	1049	1060	1062	5187	5197	5202	1999.1
Missing rates (%)	79.0	80.2	80.4	80.1	80.6	81.4	79.6	79.9	79.6	79.6	0.4	0.2	0.1	69.6

7.2 Empirical Validation: Annotator Modeling Before Aggregation for Reliability

We present empirical validation and analysis to concretize our first contribution—the paradigm shift from sample-wise aggregation to annotator-wise behavior modeling—further clarifying our motivation and practical significance.

Importance of Annotator Modeling for Aggregation Reliability. As illustrated in Figure 5, traditional aggregation strategies operate along the sample axis without modeling longitudinal annotator behavior. In real-world sparse and non-overlapping scenarios, this leads to suboptimal or biased consensus aggregation: sample S_1 may be aggregated from annotators A_2, A_3, A_4, while sample S_2

from A_1, A_3, A_4—sometimes even entirely disjoint groups. Therefore, we hypothesize that modeling annotators individually allows for better reconstruction of their missing labels, thereby enabling more consistent and trustworthy aggregation across all samples (e.g., all S_1–S_N aggregated from A_1, A_2, A_3, A_4) given well-trained annotator models.

Validation: Direct Aggregation vs. Aggregation after Annotator-wise Modeling. To validate this, as shown in Table 6, we simulate missingness on the test set by masking {20%, 30%, 40%} of annotations and compare: (1) **S-MV**: Direct sample-wise majority vote on remaining labels; (2) **S-Ours-MV**: Predict missing labels using annotator-specific models trained on the full training set, then aggregate all (remaining + padded) labels; Similarly, **S-PADL-MV** represents the best-performing baseline PADL [17] results. The results in Table 6 show that our method consistently outperforms baselines, validating that longitudinal annotator-wise modeling enhances aggregation reliability.

This section provides empirical evidence for our paradigm shift, further clarifying the motivation of this paper while depicting its practical utility value for the real world.

7.3 Video-specific Architecture

Different from the image input pipeline in the main paper, Figure 6 presents the video input pipeline specifically designed for the AMER, etc. video datasets. Specifically, given a video input $V \in \mathbb{R}^{T\times H\times W\times 3}$, a frozen pre-trained image encoder [32] first extracts frame features, which are then further compressed by the image Q-Former [14] using a specific number of compression queries, typically 32, to alleviate subsequent computational costs [14]. The frame position embedding is then added and input to the video Q-Former. Subsequently, we assign a learnable query token in a Q-Former for modeling each annotator A_k ($k = 1, \ldots, n$), then all annotator-specific queries in cross-attention simultaneously interact with input features to capture their diverse tendencies and obtain the annotator-specific feature. Finally, the specific annotator features are mapped through a fully connected layer to an appropriate feature dimension and then connected to each annotator's corresponding classifier to output classification results.

For qualitative understanding, the focus regions of the video input are based on all frames. Different annotators' varying levels of focus on different frames indicate their behavior pattern differences: *annotator 1* shows more focus on the final frames, while *annotator n* focuses more on the middle frames. This dataset is for video emotion classification. Analyzing the annotation and visualization results, we see that annotators labeled emotions "anger" and "worried". The differences in the frame segments they focused on might partially contribute to the variations in their judgments.

7.4 Model Efficiency Analysis

This section evaluates the model efficiency from two aspects: model complexity and inference time. For model complexity, we use the number of parameters as the evaluation metric; for inference time, we compute the average processing time per sample [30, 42, 47–49]. To ensure a fair comparison, Table 8 compares the performance of different methods in the image input pipeline (see Figure 2 in the main paper). Specifically, we evaluate the efficiency of different methods on one perspective of the STREET dataset with 10 annotators. It is worth noting that different methods use different backbone networks. To ensure fairness, we standardize the backbone network to ResNet-34 for all methods. In Table 8, we observe that our model has fewer parameters than the other models, while also maintaining competitive average processing time. This validates our claim in the paper that, compared to the baselines that create separate conventional models for each annotator, our architecture demonstrates significant advantages in model efficiency.

Table 8: Model efficiency analysis. In this table, we report the number of parameters and the average inference time per sample. Lower values for both metrics indicate higher model efficiency.

Models	Parameters (M)	Average Processing Time (s)
D-LEMA	214.18	5.64
PADL	168.94	4.59
MaDL	201.37	5.06
Ours	**106.02**	**4.28**

7.5 Discussion

We supplement some extended discussions about our work here. Our architecture design superiorly balances effectiveness and complexity, provides an accompanying visualization analysis of annotator tendencies through cross-attention weights. This has potential value and interest for understanding different annotator judgments as described in the main paper. Currently, it is introduced as an accompanying function rather than a main contribution of our paper, but in the future, we hope it can be developed into a mature interpretability solution. We plan to explore to enhance it, which may need pixel-level annotations indicating specific regions that annotators focus on during the annotation process, although this could be expensive. We could also explore quantitative evaluation ways to this further interpretability, such as feature importance ranking consistency, attention-based fidelity metrics, and human intuition consistency assessments through user studies, etc, to further enhance its value and contribution to the community.

In the ablation study of our main paper, we validated an insight that modeling multiple annotator tendencies can obtain more annotator information and demonstrated its effectiveness through experimental results based on consensus aggregation using majority voting. In future work, we plan to extend more similar experimental scenarios and validations to enhance the explanation of how multi-annotator tendency learning helps and adds benefits to specific applications [45].

7.6 Additional Results

We provide additional experimental results about visualization analysis of annotator tendencies on both AMER and STREET datasets.

As shown in Figure 7, on the AMER dataset, annotators demonstrate different tendencies through varying focus on different frames. Sample 1 reveals the differences: *annotators 5, 10, and 11* show similarly higher focus on the final frames (45-54), while *annotator 12* focuses more on middle frames (30-35). They all predict the correct label "happy", but the different focus positions indicate that *annotator 12*'s preference pattern for happiness in sample 1 differs from most annotators. Sample 2 shows: *annotators 2, 8, and 13* exhibit similarly higher focus on the final frames (45-55), while *annotator 12* focuses more on early frames (5-10). They all predict "sad" but *annotator 12* demonstrates a different preference pattern.

As shown in Figure 8, on the STREET dataset, annotators exhibit different tendencies through varying focus on different semantic elements within the same input image. In sample 1 from the orderliness perspective, the focus regions show differences: *annotator 5* focuses more intensively on graffiti in the image compared with other annotators. The results show that prediction and ground truth are highly consistent in the emotion reflected in the semantics, corresponding to this annotator's expected lowest score. In sample 2 of the healthiness perspective, most annotators focus on uncleaned garbage and leaves, and only *annotator 6* shows more focus on the surrounding environment, like cars and buildings, which might influence the result of not giving a low score.

REFERENCES

[1] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. 2017. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images. *International ophthalmology* 37 (2017), 701–717.

[2] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, 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 (2011), 915–931.

[3] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. 2019. Max-mig: an information theoretic approach for joint learning from crowds. *arXiv preprint arXiv:1905.13436* (2019).

[4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* 32 (2019).

[5] Junfan Chen, Richong Zhang, Jie Xu, Chunming Hu, and Yongyi Mao. 2023. A Neural Expectation-Maximization Framework for Noisy Multi-Label Text Classification. *IEEE Transactions on Knowledge and Data Engineering* 35, 11 (2023), 10992–11003. https://doi.org/10.1109/TKDE.2022.3223067

[6] Yuan-Chia Cheng, Zu-Yun Shiau, Fu-En Yang, and Yu-Chiang Frank Wang. 2023. TAX: Tendency-and-Assignment Explainer for Semantic Segmentation with Multi-Annotators. *arXiv preprint arXiv:2302.09561* (2023).

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems.* https://openreview.net/forum?id=vvoWPYqZJA

[8] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[9] Marek Herde, Denis Huseljic, and Bernhard Sick. 2023. Multi-annotator Deep Learning: A Probabilistic Framework for Classification. *arXiv preprint arXiv:2304.02539* (2023).

[10] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. 2021. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 12336–12346. https://doi.org/10.1109/CVPR46437.2021.01216

[11] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. 2021. 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.

[12] Uthman Jinadu, Jesse Annan, Shanshan Wen, and Yi Ding. 2023. Loss Modeling for Multi-Annotator Datasets. *arXiv preprint arXiv:2311.00619* (2023).

[13] Alex Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. (Jan 2009).

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning.* PMLR, 19730–19742.

[15] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, 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.

[16] Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, 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.

[17] Zehui Liao, Shishuai Hu, Yutong Xie, and Yong Xia. 2024. Modeling annotator preference and stochastic annotation error for medical image segmentation. *Medical Image Analysis* 92 (2024), 103028.

[18] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, and et al. 2020. Quantification of uncertainties in biomedical image quantification challenge. https://qubiq.grand-challenge.org/.

[19] Zahra Mirikharaji, Kumar Abhishek, Saeed Izadi, and Ghassan Hamarneh. 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.

[20] Jeppe Nørregaard and Leon Derczynski. 2022. Sparse Probability of Agreement. *arXiv preprint arXiv:2208.06161* (2022).

[21] Joshua Peterson, Ruairidh Battleday, Thomas Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* https://doi.org/10.1109/iccv.2019.00971

[22] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision.* 9617–9626.

[23] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957* (2017).

[24] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning From Crowds. *Journal of Machine Learning Research* 11, 43 (2010), 1297–1322. http://jmlr.org/papers/v11/raykar10a.html

[25] Filipe Rodrigues and Francisco Pereira. 2018. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[26] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* 34, 12 (2013), 1428–1436.

[27] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2013. Learning from multiple annotators: distinguishing good from random labelers. *Pattern Recognition Letters* 34, 12 (2013), 1428–1436.

[28] Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Gaussian Process Classification and Active Learning with Multiple Annotators. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 433–441. https://proceedings.mlr.press/v32/rodrigues14.html

[29] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. 2019. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 76 (Nov. 2019), 23 pages. https://doi.org/10.1145/3359178

[30] Xuanmeng Sha, Liyun Zhang, Tomohiro Mashita, and Yuki Uranishi. 2024. 3DFacePolicy: Speech-Driven 3D Facial Animation with Diffusion Policy. *arXiv preprint arXiv:2409.10848* (2024).

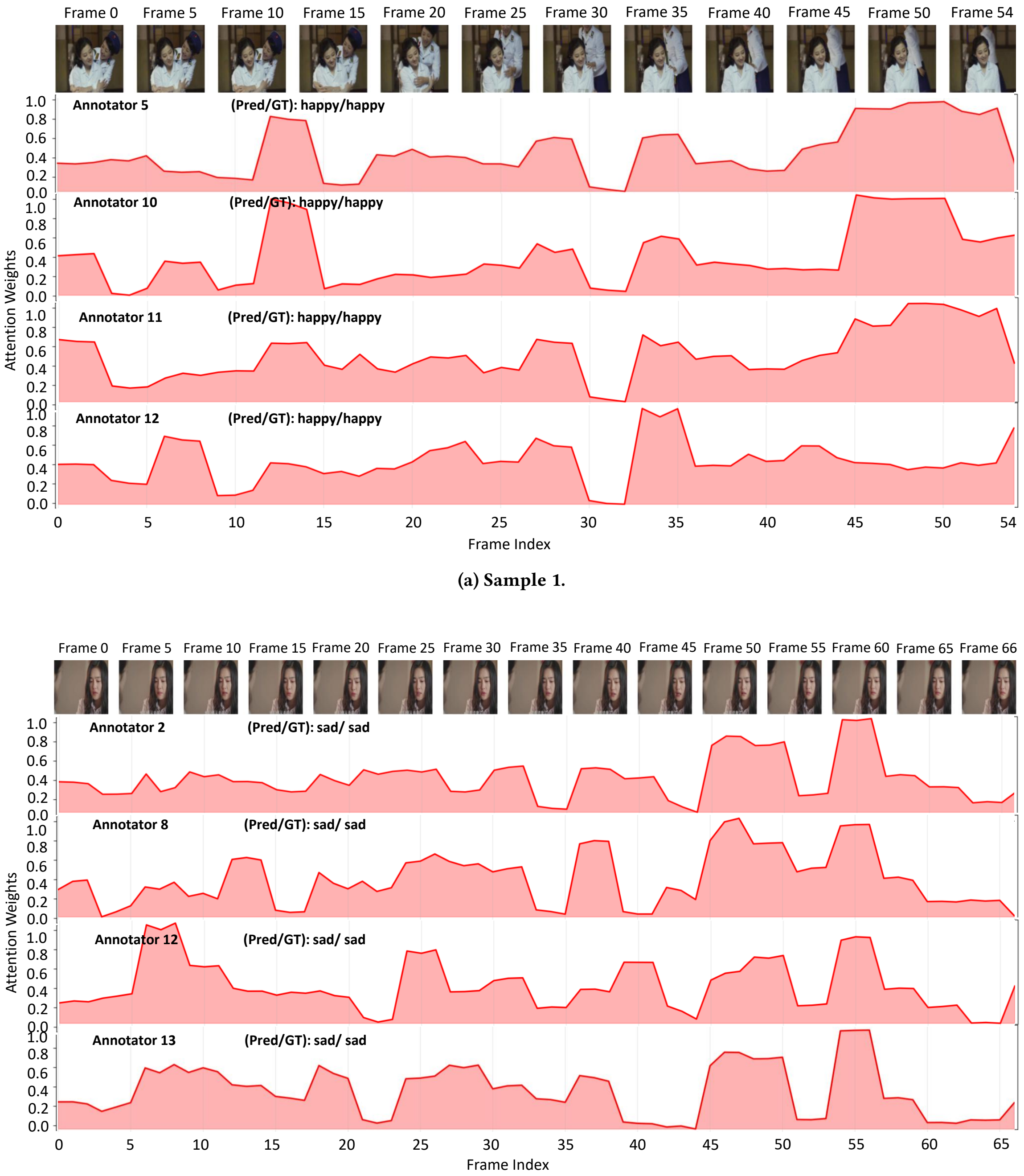
[31] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Mirella Lapata and Hwee Tou Ng (Eds.). Association for Computational Linguistics, Honolulu, Hawaii, 254–263. https://aclanthology.org/D08-1027

[32] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).

[33] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. 2019. 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.

[34] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. 2019. Learning From Noisy Labels by Regularized Estimation of Annotator Confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[35] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. https://github.com/HumanSignal/label-studio Open source software available from https://github.com/HumanSignal/label-studio.



(a) Sample 1.

(b) Sample 2.

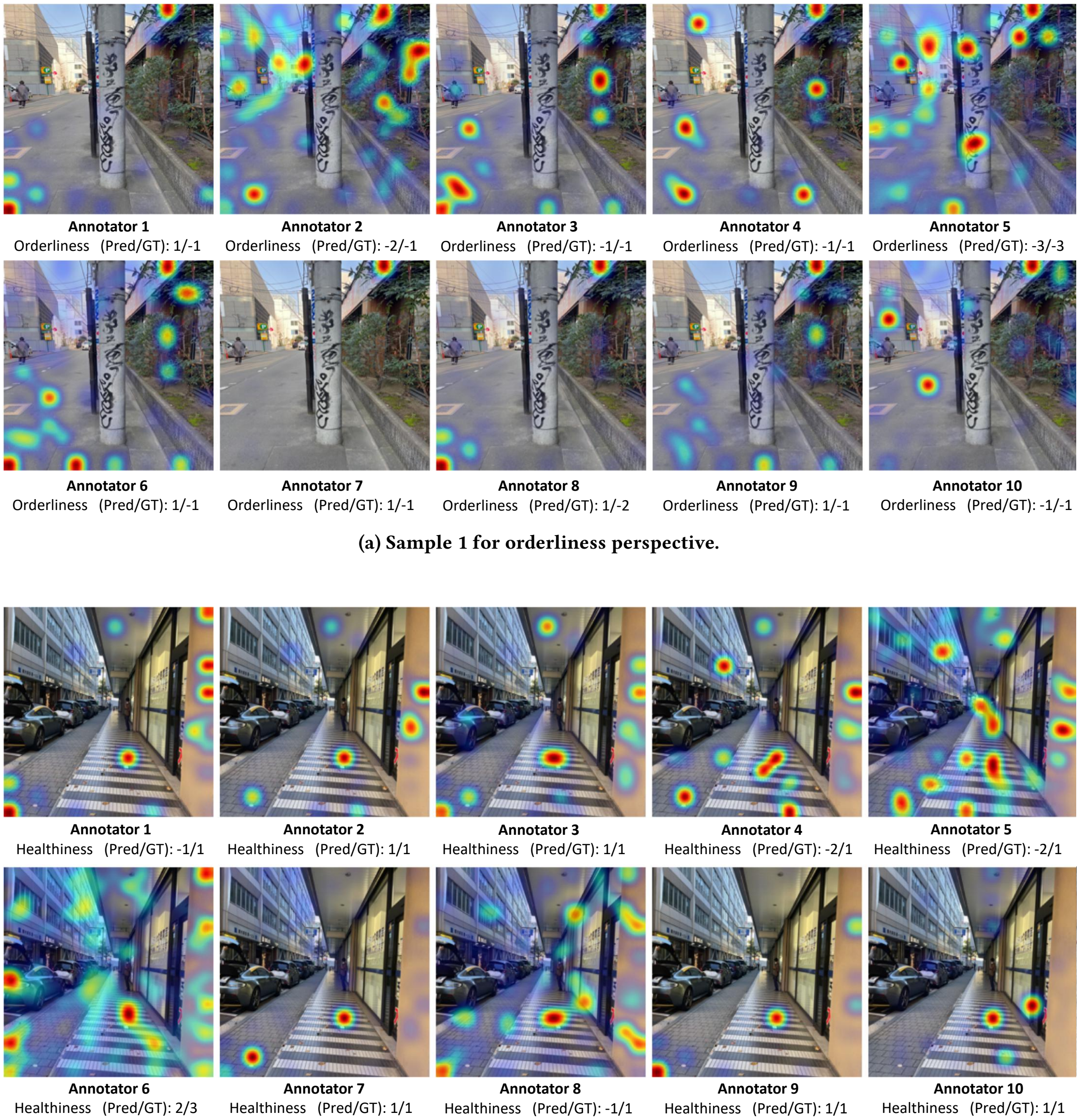
Figure 7: Additional sample experimental results visualize the annotator tendencies on the AMER video dataset. The tendencies of multi-annotators reveal distinct preferences.

[36] Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lum Lam, Nicholas D Lane, and Nadia Bianchi-Berthouze. 2023. Learn2agree: Fitting with multiple annotators without objective ground truth. In *International Workshop on Trustworthy Machine Learning for Healthcare*. Springer, 147–162.

[37] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems* 23 (2010).

[38] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems* 22 (2009).

[39] Yan Yan, Rómer Rosales, Glenn Fung, Ramanathan Subramanian, and Jennifer Dy. 2014. Learning from multiple annotators with varying expertise. *Machine learning* 95 (2014), 291–327.


Annotator 1
Orderliness (Pred/GT): 1/-1
Annotator 2
Orderliness (Pred/GT): -2/-1
Annotator 3
Orderliness (Pred/GT): -1/-1
Annotator 4
Orderliness (Pred/GT): -1/-1
Annotator 5
Orderliness (Pred/GT): -3/-3
Annotator 6
Orderliness (Pred/GT): 1/-1
Annotator 7
Orderliness (Pred/GT): 1/-1
Annotator 8
Orderliness (Pred/GT): 1/-2
Annotator 9
Orderliness (Pred/GT): 1/-1
Annotator 10
Orderliness (Pred/GT): -1/-1
Annotator 1
Healthiness (Pred/GT): -1/1
Annotator 2
Healthiness (Pred/GT): 1/1
Annotator 3
Healthiness (Pred/GT): 1/1
Annotator 4
Healthiness (Pred/GT): -2/1
Annotator 5
Healthiness (Pred/GT): -2/1
Annotator 6
Healthiness (Pred/GT): 2/3
Annotator 7
Healthiness (Pred/GT): 1/1
Annotator 8
Healthiness (Pred/GT): -1/1
Annotator 9
Healthiness (Pred/GT): 1/1
Annotator 10
Healthiness (Pred/GT): 1/1

(a) Sample 1 for orderliness perspective.

(b) Sample 2 for healthiness perspective.

Figure 8: Additional sample experimental results visualize the annotator tendencies on the STREET image dataset. The tendencies of 10 annotators reveal distinct preferences.

[40] Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. 2019. Learning from multi-annotator data: A noise-aware classification framework. *ACM Transactions on Information Systems (TOIS)* 37, 2 (2019), 1–28.

[41] Hang Zhang, Xin Li, and Lidong Bing. 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.

[42] Liyun Zhang. 2024. *Integrating Panoptic-Level to Image Translation.* Ph. D. Dissertation. PhD Dissertation.

[43] Liyun Zhang. 2024. MicroEmo: Time-Sensitive Multimodal Emotion Recognition with Micro-Expression Dynamics in Video Dialogues. *arXiv preprint arXiv:2407.16552* (2024).

[44] Liyun Zhang, Zheng Lian, Hong Liu, Takanori Takebe, and Yuta Nakashima. 2025. SimLabel: Similarity-Weighted Semi-supervision for Multi-annotator Learning with Missing Labels. *arXiv preprint arXiv:2504.09525* (2025).

[45] Liyun Zhang, Nanyan Liu, Yuanbin Hou, and Xiaojian Liu. 2014. Uneven illumination image segmentation based on multi-threshold S-F. *Opto-Electronic Engineering* 41, 7 (2014), 81–87.

[46] Liyun Zhang, Zhaojie Luo, Shuqiong Wu, and Yuta Nakashima. 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.

[47] Liyun Zhang, Photchara Ratsamee, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2023. Panoptic-level image-to-image translation for object recognition and visual odometry enhancement. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 2 (2023), 938–954.

[48] Liyun Zhang, Photchara Ratsamee, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2022. Thermal-to-Color Image Translation for Enhancing Visual

Odometry of Thermal Vision. In *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 33–40.

[49] Liyun Zhang, Photchara Ratsamee, Bowen Wang, Zhaojie Luo, Yuki Uranishi, Manabu Higashida, and Haruo Takemura. 2023. Panoptic-aware image-to-image translation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 259–268.

[50] Le Zhang, Ryutaro Tanno, Moucheng Xu, Yawen Huang, Kevin Bronik, Chen Jin, Joseph Jacob, Yefeng Zheng, Ling Shao, Olga Ciccarelli, et al. 2023. Learning from multiple annotators for medical image segmentation. *Pattern Recognition* 138 (2023), 109400.

[51] Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. 2023. Magi: Multi-annotated explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1977–1987.