Dyadic Imitation Modeling With Lag-Aware Dual-Stream Framework for Autism Classification

Wenqi Ji, Tong Liu, Xue Yang, Ying Guo, Xin Chen, Pei Yang, Kaiyun Li, and Yong-Jin Liu, *Senior Member, IEEE*

***Abstract*—This paper introduces a novel lag-aware dual-stream (LADS) framework and a carefully curated dual-view video dataset for automatic Autism Spectrum Disorder (ASD) classification through imitation tasks. In contrast to prior single-view approaches that overlook the interactive dynamics of imitation, our dataset is the first to capture synchronized experimenter-child interactions with rich pose and motion features. Building on this data, the LADS framework explicitly learns the temporal alignment between the experimenter's demonstration and the child's imitative response. A Lag-Aware Alignment module uses constrained cross-attention to compute an adaptive time warping and extract per-frame lag feature, revealing delays in the child's imitation. Additionally, a lightweight diffusion-based regularizer enforces representation consistency by denoising perturbed child features conditioned on the experimenter's motion, improving generalization. We then achieve the classification of ASD versus Typical Development (TD) behavior by integrating the aligned dual-stream features, imitation lag, and action discrepancy within an attention-pooling classifier. Experiments on our dual-view imitation dataset show that LADS significantly outperforms conventional single-stream models and a recent dyadic transformer baseline, achieving state-of-the-art classification accuracy. The results demonstrate the importance of modeling interpersonal timing in social behavior analysis. Our work provides a new, public dataset and a computational tool for interdisciplinary research, bridging computer vision and psychological studies of autism. Both dataset and code will be made publicly available.**

***Index Terms*—ASD classification, imitative deficits, dyadic interaction modeling, deep learning.**

I. Introduction

AUTISM Spectrum Disorder (ASD) is a prevalent neurodevelopmental disorder characterized by difficulties in social communication, restricted interests and stereotyped behavior [1]. Among various manifestations of ASD in children, deficits in imitative abilities stand out as significant features. These deficits can adversely impact social learning and development of children, highlighting the importance of reliable, scalable assessment of ASD for early intervention and treatment planning [2], [3].

Much prior work on video-based ASD analysis relies on single-view recordings or observer notes [4], which offer valuable but inherently limited perspectives on interactive behaviors. Single-view pipelines often struggle to capture the dyadic structure of imitation, where the child's behavior must be interpreted in relation to the experimenter's demonstration [5]. Consequently, many existing datasets do not expose sufficient information about temporal coupling between participants, constraining fine-grained analysis of imitation quality [6], [7]. In contrast, dual-view recordings can preserve the interaction context by simultaneously capturing the experimenter's action and the child's response. This provides a more comprehensive description of imitation that enables more reliable, multi-dimensional behavior analysis than single-view approaches.

Recent advances in machine learning have offered new opportunities for automated behavioral analysis in healthcare applications. Some works have employed machine learning algorithms to accelerate ASD diagnosis based on the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS), respectively [8], [9]. However, these questionnaire and observation-based methods still rely on manual assessments and can be time-consuming. Video-based ASD classification has thus gained popularity due to its non-invasive nature and scalability for large-scale screenings [10], [11]. Nonetheless, most existing video-based methods for ASD classification process only a single stream [12], analyzing specific types of behaviors, such as facial expressions or hand gestures [7], [13]. In contrast, our investigation utilizes synchronized dual-view recordings of experimenters and children, enabling effective classification of ASD. To our knowledge, such dyadic, dual-stream analysis for ASD identification remains largely unexplored and thereby our work offers a complementary advance beyond single-person approaches.

This paper contributes to both dataset and method for dyadic imitation analysis in ASD classification. We introduce a novel

Received 7 October 2025; revised 27 November 2025; accepted 6 December 2025. Date of publication 11 December 2025; date of current version 7 May 2026. This work was supported by the Natural Science Foundation of China under Grant U2336214. This article was recommended by Associate Editor C. Chen. *(Corresponding authors: Yong-Jin Liu; Kaiyun Li.)*

Wenqi Ji, Tong Liu, Xue Yang, Xin Chen, Pei Yang, and Yong-Jin Liu are with the Department of Computer Science and Technology and the MOE-Key Laboratory of Pervasive Computing, Tsinghua University, Beijing 100084, China (e-mail: jwq21@mails.tsinghua.edu.cn; tongliu@tsinghua.edu.cn; yang_shirley@tsinghua.edu.cn; yangpei20@mails.tsinghua.edu.cn; liuyongjin@tsinghua.edu.cn).

Ying Guo is with the School of Computer Science, North China University of Technology, Beijing 100144, China (e-mail: guoying@ncut.edu.cn).

Kaiyun Li is with the School of Education and Psychology, University of Jinan, Jinan, Shandong 250022, China (e-mail: Sep_liky@ujn.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2025.3642987

dual-view dataset that captures synchronized experimenter action and child imitation, accompanied by multiple feature modalities (e.g., skeleton, optical flow, and heatmaps). On the method side, we propose a lag-aware dual-stream (LADS) architecture that explicitly models the temporal delay and behavioral difference between the experimenter and the child. Our approach is built upon two key components. First, a **Lag-Aware Alignment (LAA) module** that applies constrained cross-attention within a local temporal window between the two encoded sequences from the experimenter and the child. This module produces an attention map, a time-varying lag curve, and a warped trace of the experimenter sequence. Second, we introduce a **lightweight diffusion style feature regularizer**, conditioned on the experimenter's encoded sequence. The model learns to denoise a noisy embedding of the child sequence, acting as a representation-level regularizer that improves generalization without pixel-level generation. The aligned pair and the explicit lag then feed a difference encoder and an attention-guided pooling head for ASD/TD classification.

The main contributions of this work include:

- We present a comprehensive **dual-view dataset** capturing a wide range of imitative actions from children with ASD and TD, including meaningful gestures and object placement tasks, to enable the analysis of social communication behaviors.
- We establish a **comprehensive benchmark** with implementation and evaluation of baseline methods of ASD classification on our proposed dataset, providing a foundation for future research.
- We propose a **lag-aware dual-stream (LADS) architecture** that integrates LAA and a diffusion-style feature regularizer for analyzing dyadic interaction data. Experiment results show that incorporating dual-view information improves ASD classification performance compared to traditional single-view approaches.

II. Related Works

A. Imitation Tasks for Autism Classification

Imitation tasks have been widely used to investigate the social and cognitive development of children with ASD. This experimental paradigm requires the child to map observed actions onto their own movements and to coordinate timing and actions with the experimenter. By design, the experimental paradigms include both meaningful and meaningless gestures and object manipulations, enabling researchers to distinguish pure motor reproduction from action understanding of participated children [14], [15].

A substantial body of literature shows that in imitation tasks, the imitation of meaningless gestures is one of the most rigorous paradigms for testing imitation deficits. This approach effectively distinguishes children with ASD from those with TD, as it eliminates reliance on familiarity with the goals and methods of the actions. Studies using this paradigm have reported impairments in meaningless object manipulation tasks, such as placing bricks without a specific order, in children with ASD across various age groups and developmental levels [16], [17]. Furthermore, research has found that the deficit of imitating meaningless gestures persist through adolescence and adulthood [18]. Further analysis reveals specific patterns of impairment: While children with ASD show preserved capacity for imitating meaningful object manipulation (e.g., activating mechanical devices) [19], they exhibit difficulties in both meaningless object manipulation tasks (e.g., arranging blocks in non-functional configurations) and facial expression imitation [20], [21]. This dissociation suggests domain-specific deficits in processing non-social or structurally ambiguous stimuli.

The behavioral findings have facilitated clinical applications of imitation assessment. For instance, Mussey and Klinger [15] suggested that the impairment of imitation could be a distinguishing characteristic, which may be used as an early marker for autism diagnosis. Similarly, Du te al. [22] showed that ASD children exhibit atypical patterns of neural synchrony while performing imitation tasks, further supporting the use of imitation-based behavioral and neural markers for autism classification. Building on these insights, prior work motivates datasets and models that not only score how closely a child reproduces a experimenter's movement but also quantify temporal alignment between the child and the experimenter. Accordingly, our study focuses on imitation-based evaluation and releases a representation-level dataset featuring synchronized dual streams of the experimenter's demonstration and the child's imitation. Rather than distributing raw videos, we provide preprocessed feature sequences as skeletons, optical flow, and heatmap, which faithfully encode pose and motion, facilitate privacy-preserving sharing, and enable reproducible benchmarking for dyadic interaction modeling.

B. Autism Classification Methods

Traditional ASD diagnostic approaches rely on time consuming clinical assessments and subjective observations that is labor intensive [23]. Video-based methods, in contrast, provide a non-invasive and scalable alternative [11], [12]. By analyzing video recordings, researchers and clinicians can now identify ASD-related behaviors in a more systematic and automated manner, enabling accurate pre-clinical ASD classification [12].

Early video-based ASD classification methods analyzed gaze patterns associated to ASD from eye-tracking data with hand-crafted frequency distribution features and support vector machine (SVM) [24], [25]. While eye-tracking provides valuable insights into visual attention patterns, researchers have expanded their focus to encompass other behavioral markers. Facial expression dynamics constitute an important behavioral domain, with substantial evidence suggesting atypical expression patterns in ASD individuals compared to TD individuals [26]. Grossard et al. [27] computed geometric features by measuring distances between facial landmarks and derived appearance features from histograms of oriented gradients (HOG), subsequently employing Random Forest classifiers for ASD classification. In parallel work, Krishnappababu et al. [28] used facial landmark data to calculate multi-scale entropy (MSE) for quantifying the dynamic complexity in the eyebrow and mouth regions, feeding the MSE

features into a decision tree classifier to effectively differentiate between ASD and TD children.

The advent of deep learning has transformed facial video analysis, enabling researchers to capture both spatial and temporal patterns more comprehensively. Zhang et al. [29] employed CNNs [30] to encode spatial features from clinical interview recordings of adult ASD patients, incorporating few-shot learning techniques to improve model generalizability across limited training samples.

Beyond facial analysis, gesture and pose information has emerged as critical modalities for ASD classification, as facial expression analysis typically requires high-quality video captured in controlled laboratory environments and fails to capture the full spectrum of dynamic social interactions. Importantly, autism is characterized not only through emotional states, which is primarily reflected in facial expressions, but also through distinctive patterns in social interaction, communication, and repetitive behaviors [31]. These complex behavioral patterns are more effectively observed through body movements and gestures during interactive contexts. Early deep learning approaches for gesture-based ASD classification focused on developing end-to-end models that automatically extract features from videos and perform the classification. Zunino et al. [7] pioneered a hybrid architecture combining CNNs with long short-term memory networks (LSTM) [32] to capture spatio-temporal patterns from hand action videos during object manipulation tasks. Subsequent work explored more sophisticated architectures. Sun et al. [13] integrated 3D-CNNs with spatial attention mechanisms through bilinear pooling, thereby enhancing the model's ability to focus on diagnostically relevant spatial features.

Recent studies have also explored various learning paradigms to improve model robustness or address data scarcity for ASD classification. Rani and Verma [12] introduced a contrastive learning framework that uses ResNet [33] to extract video features and incorporates multiple datasets, to learn discriminative representations for distinguishing ASD from TD children. In parallel, as gesture-based ASD classification relies on capturing motion and pose patterns, researchers have investigated whether established action recognition architectures can effectively transfer to ASD classification tasks. Studies have employed various pre-trained models including I3D [34] and PoseC3D [35] as motion feature extractors. For instance, Pandey et al. [6] investigated the use of both I3D and TSN [36] as backbone models to extract features from videos and assess the skills of children with autism.Beyond ASD-specific applications, skeleton-based action recognition has also witnessed rapid progress. Li et al. [37] proposed the U-FEFP framework, which combines BYOL-style contrastive learning with reversed-sequence reconstruction to learn rich spatio-temporal features from large-scale 3D skeleton benchmarks, serving as a generic pre-training scheme.

Recent advances in transformer-based architectures have further expanded the methodological toolkit. Zahan et al. [38] combined vision transformers (ViT) [39] with its GCN and TCN based model to compute an aggregated embeddings for derivation of euclidean distance loss, which offered a notably performance improvements. Nevertheless, despite the continuous development of methods in action recognition and temporal modeling [40], [41], these advances remain largely unexplored in autism research, presenting potential for improving automated ASD classification systems.Related progress has also been made in group activity recognition (GAR), where models are designed to capture complex multi-person interactions in sports or surveillance scenarios. For instance, Pei et al. [42] introduced a Key Role Guided Transformer that performs coarse-to-fine reasoning to highlight key individuals while maintaining global spatio-temporal context, and Wang et al. [43] further proposed a knowledge-augmented relation inference framework that injects class–class and class–position priors into transformer-based relation modeling. Unlike these GAR approaches that infer categorical group activity labels from raw RGB videos with multi-person bounding boxes, our LADS framework focuses on dyadic experimenter–child imitation for ASD classification and explicitly models temporal alignment between two interaction partners.

A critical limitation persists across most existing approaches: they predominantly focus on analyzing individual behaviors in isolation, overlooking the inherently dyadic nature of social interactions that are fundamental to autism assessment. Recognizing this gap, growing researches have begun to explore social interaction scenarios, acknowledging that ASD characteristics often emerge most clearly during interpersonal communication and coordinated activities. Zhao et al. [11] captured videos of both child and examiner interactions, employing motion energy analysis (MEA) [44] to extract full-body motion time series. They subsequently applied cross wavelet analysis (CWA) to quantify interpersonal movement coordination (IMC) as behavioral markers, utilizing SVM for final classification. Similarly, Cheng et al. [45] developed a computer-aided ASD diagnostic system for social interaction scenarios, employing ResNet-50 to extract gesture and facial expression features from children, although their system did not account for multi-person behavior patterns.

Beyond autism-specific research, recent advances in other domains have demonstrated sophisticated approaches for processing synchronized dual-stream data. Notably, the Dyadic Interaction Modeling (DIM) [46] framework illustrates this capability by jointly modeling speaker and listener behaviors in conversational settings, using self-supervised pretraining to learn unified, context-aware representations of bidirectional nonverbal interactions. While originally designed for generating 3D facial animations in human-computer interaction scenarios, DIM's dual-stream architecture, which captures the temporal coordination between interaction partners, offers valuable insights for modeling similar dyadic dynamics in ASD classification contexts.

C. Video-Based Autism Classification Datasets

Despite promising advances in video-based ASD classification, the field faces significant data availability challenges that most existing methods rely on privately collected datasets, while publicly accessible resources remain severely limited. Among the few available datasets, Zunino et al. [7] introduced

a collection featuring 20 ASD and 10 TD children performing structured manipulation tasks, which including right-hand grasping, bottle placement, and water pouring, with recordings restricted to hand movements from fixed camera positions. While this dataset has facilitated several studies [12], [13], its modest scale and highly controlled setup constrain model generalizability. Similarly, Pandey et al. [6] presented a dataset containing video recordings of 37 ASD children interacting with clinicians in a laboratory-controlled environment, covering eight action types. However, this dataset captures only the children's behaviors, omitting the experimenters' corresponding actions that would provide essential interaction context.

Critically, the field lacks publicly available datasets specifically designed for imitation-based ASD assessment. Existing datasets either focus on isolated actions or unidirectional recordings, failing to capture the synchronized, bidirectional nature of imitation tasks. To our knowledge, no public dataset simultaneously records both the experimenter's actions and the child's imitative responses during structured imitation protocols. This absence represents a fundamental gap, as imitation tasks require analyzing not just whether a child performs an action, but how their movements temporally and spatially align with the experimenter. A dual-view imitation dataset would enable models to extract richer behavioral signatures, including timing delays, movement synchrony, and response accuracy. Our work addresses this critical need by introducing a public dual-stream imitation dataset for ASD classification, capturing synchronized recordings of both experimenter demonstrations and child responses across diverse imitation tasks.

III. Dual-View Imitation Dataset for Autism Classification

Recognizing the scarcity in publicly available imitation-based ASD assessment data, we present the first dual-view dataset specifically designed for autism classification through imitation tasks. Unlike existing datasets that capture only isolated actions or single-perspective recordings, our dataset simultaneously records both the view of experimenters and that of the participating children, enabling comprehensive analysis of imitation behavior patterns. To address privacy concerns while maintaining research utility, we release preprocessed feature representations rather than raw videos, preserving participant confidentiality while providing rich behavioral information for computational model development.

A. Ethical Declarations

All experiments conducted in this study were approved by the Ethics Committee of University of Jinan. Verbal consent was obtained from the parents of the children who participated in this study. All the children's parents consented to video recordings during the experimental sessions and understood the research purposes of the data collection.

B. Participants

We recruited 84 preschool children from special education centers, rehabilitation facilities, and kindergartens in Jinan, China. Among the children, 45 were diagnosed with ASD and 39 were TD children. The gender and age distributions of the TD group were matched to those of the ASD group. Participation requirements for the ASD group included: (1) formal clinical diagnosis based on DSM-V criteria [47], (2) ability to follow basic verbal instructions (e.g., "perform this action"), and (3) absence of severe behavioral issues, uncorrected vision, or hearing impairments. The TD group consisted of children with no known developmental disorders or neurological conditions. Parents provided informed consent and assisted during data collection when necessary. The role of the experimenter was performed by a trained 26-year-old female researcher.

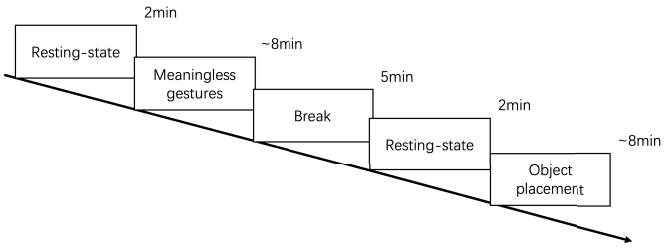

Fig. 1. The trial procedure of our dual-view imitation video data collection.

C. Experimental Setup and Procedure

1) Task Selection: Based on established psychological paradigms [48], [49], we selected two categories of imitation tasks that effectively differentiate ASD-related behavioral patterns:

- Meaningless gestures: Non-goal-directed movements including finger spreading, hand rotation near the mouth, and touching left shoulder with right hand. These tasks isolate action imitation abilities from semantic understanding.
- Object manipulation: Structured placement of blocks in specific configurations on a table surface, testing both movement planning and spatial awareness.

2) Experimental Protocol: The experiment was divided into two phases: a practice phase and the formal experimental phase.

During the practice phase, the child and experimenter were asked to sit facing each other. The experimenter demonstrated two target actions to the child, and instructed them to imitate the actions. Each action was repeated three times, and verbal encouragement was provided throughout. If the child failed to imitate the actions correctly during the practice phase, the experimenter used verbal or physical guidance to ensure the child understood the task before proceeding to the formal experiment phase.

The formal experiment process, shown in Figure 1, consisted of two tasks: a meaningless gesture imitation task and an object placement imitation task. A five-minute break separated the tasks. At the beginning of each task, the child was asked to sit quietly for two minutes of rest. After the rest period, the experimenter demonstrated several specific actions. Each action was demonstrated three times in a randomized order.

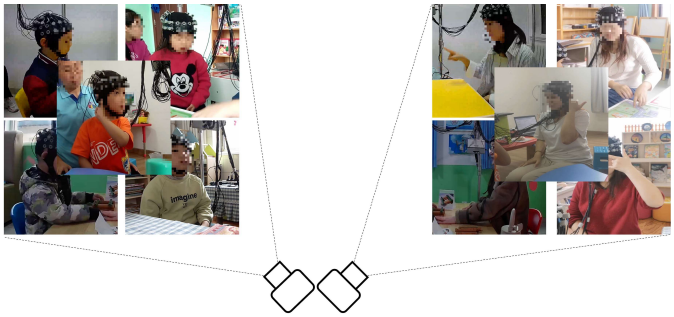

Fig. 2. Example of a meaningless gesture imitation task from the perspectives of the experimenter and the subject.

The child was then instructed to imitate the actions. This phase lasted approximately eight minutes. Throughout the imitation tasks, each attempt by the child was met with both verbal and non-verbal encouragement, but no verbal or physical assistance was provided during the imitation process.

3) Data Acquisition Setup: To capture the dyadic nature of imitation tasks, we employed a dual-camera configuration in a naturalistic classroom environment as illustrated in Figure 2. As shown in the figure, we also recorded Electroencephalogram (EEG) during the sessions, but it falls outside the scope of this paper. We focus on video-derived features because they are more deployable in real-world settings, whereas EEG acquisition is impractical. This setting, compared with strictly controlled laboratory conditions, encourages natural behavioral expression while maintaining sufficient control for standardized assessment and provides more relevant data for the early detection of ASD in daily scenario. Specifically, we set up two synchronized cameras positioned to capture the actions of both participants from different angles. One camera was directed at the experimenter, while the other was aimed at the child participant, forming the dual-view setup illustrated in Figure 2. The video data was captured using webcams operating at 30 frames per second.

D. Data Preprocessing

To capture and provide the rich behavioral information present in imitation tasks, we extract multiple complementary features directly from the raw video recordings, as illustrated in Figure 3. Rather than focusing on multi-modal fusion, our approach emphasizes comprehensive video representation through diverse feature extraction methods. Each feature type captures distinct aspects of movement and pose, collectively providing a thorough characterization of imitation behaviors.

In our setup, the two webcams are mounted at slightly oblique angles with respect to the participants as shown in Figure 2. BlazePose outputs the joint coordinates and a per-joint visibility score, which are robust to moderate viewpoint changes. Empirically, visual inspection of random samples from both views shows that the recovered skeletons maintain consistent poses across time.

The extracted features naturally exhibit different temporal characteristics: skeleton and heatmap features are computed per-frame (yielding T features for T frames), while optical flow features capture inter-frame transitions (producing $T-1$ features for T frames). We preserve these inherent temporal properties without artificial synchronization, allowing downstream models to utilize the natural timing relationships within the data.

1) Skeleton Features: We employ MediaPipe's BlazePose model [50] to extract 33 anatomical landmarks per frame, each represented by normalized 3D coordinates (x, y, z) and visibility confidence. This 132-dimensional (33×4) per-frame representation captures body configuration independent of camera perspective or subject distance.

2) Motion Features via Optical Flow: We compute two types of optical flow to capture different scales of movement. *Sparse optical flow*, calculated using the pyramidal Lucas-Kanade algorithm [51], tracks the displacement of 33 skeletal keypoints between consecutive frames, yielding 66-dimensional (33×2) motion vectors that effectively isolate human movement from background.*Dense optical flow*, computed via the Farneback algorithm [52], provides global motion patterns by generating pixel-wise flow fields that are spatially pooled into an 8×8 grid, resulting in 128-dimensional $(8 \times 8 \times 2)$ features that capture overall scene dynamics and contextual movement information.

3) Spatial Pose Representations: Following the PoseC3D methodology [35], we generate heatmaps for 17 key joints by rendering 2D Gaussian distributions ($\sigma = 2$) at detected joint locations. These 56×56 heatmaps undergo downsampling through average pooling, producing 7×7 spatial maps that maintain joint location information while reducing computational demands. The concatenated 833-dimensional (17×49) feature serves as the pose representation.

E. Dataset Structure

The dataset comprises 1,054 paired video recordings from 84 participants, organized as synchronized experimenter-subject pairs. Each sample contains extracted features from both participants performing imitation tasks as following structure:

All features are stored as NumPy arrays using Python's pickle format for efficient storage and loading, with metadata preserved as Python dictionaries for flexible querying.

IV. Methods

A. Lag-Aware Dual-Stream Framework

We propose LADS (Lag-Aware Dual-Stream), a novel framework for autism classification through imitation task analysis. Our approach explicitly models the temporal dynamics between the experimenter demonstrations and the subject responses by learning adaptive alignment patterns and quantifying response delays inherent in imitation behaviors. As shown in Figure 5, the framework consists of four key components: dual-stream temporal encoders, lag-aware alignment module, discrepancy-based classification, and diffusion regularizer.

1) Dual-Stream Temporal Encoding: Given the synchronized video features from an experimenter $Feat_{exp} \in \mathbb{R}^{T \times F}$ and a subject $Feat_{sub} \in \mathbb{R}^{T \times F}$, where T denotes sequence length

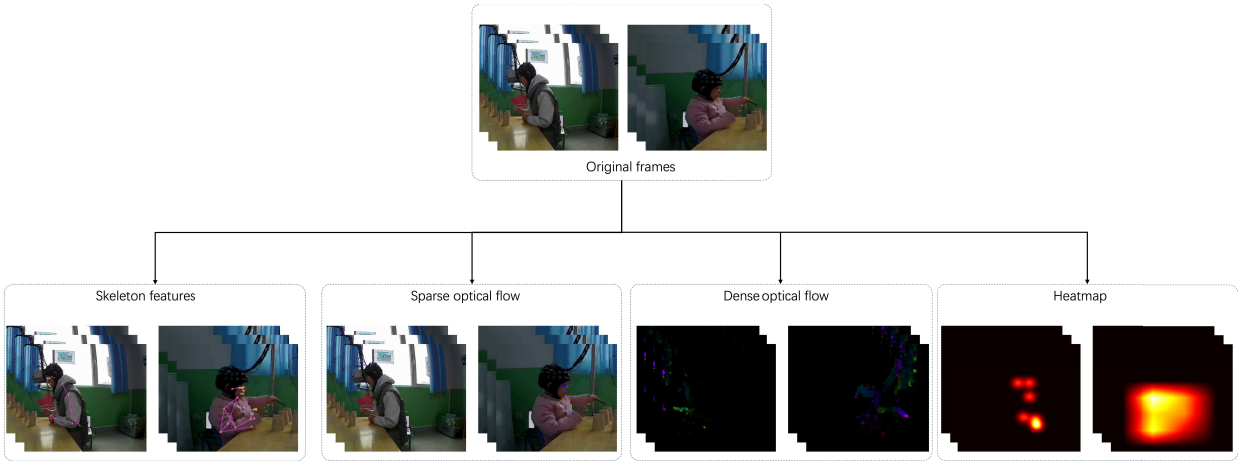

Fig. 3. The dual-view imitation dataset provide video features including skeleton sequence, sparse optical flow, dense optical flow and heatmap sequence.

```
Dataset Entry {Dict}
├─'exp': Experimenter features {Dict}
│  ├─'skeleton': (T, 132)
│  ├─'sparse_flow': (T-1, 66)
│  ├─'dense_flow': (T-1, 128)
│  └─'heatmap': (T, 833)
├─'sub': Subject features {Dict}
│  ├─'skeleton': (T, 132)
│  ├─'sparse_flow': (T-1, 66)
│  ├─'dense_flow': (T-1, 128)
│  └─'heatmap': (T, 833)
└─'metadata': Sample information {Dict}
   ├─'label': 'ASD' or 'TD'
   ├─'subject_id': subject identifier
   ├─'task_type': 'gesture' or 'object'
   └─'clip_id': Unique sequence
     identifier
```

Fig. 4. Hierarchical structure of the dual-view imitation dataset.

and F represents feature dimension, we employ two independent temporal encoders to capture stream-specific dynamics. Each encoder first linearly projects the input features to a d_{enc}-dimensional latent space, then adds a positional encoding to embed temporal order information. Then a transformer-based architecture with multi-head attention layers per stream is introduced. The encoding process yields context-aware representations as $E = Encoder_{\theta_E}(Feat_{exp}) \in \mathbb{R}^{T\times d_{enc}}$ and $S = Encoder_{\theta_S}(Feat_{sub}) \in \mathbb{R}^{T\times d_{enc}}$. The dual-stream encoders thereby produce two synchronized latent sequences which capture the abstracted representations of experimenter actions and subject movements independently, which is subsequently used for cross-attention based alignment. Both encoders also apply masking for any padded time steps to handle varying sequence lengths.

2) Lag-Aware Alignment Module: To align subject motions with experimenter demonstrations, we introduce a LAA module based on a constrained cross-attention mechanism. The LAA treats S as the query stream and E as the key–value stream. For a subject frame u, we only attend to experimenter frames t where $0 \le u - t \le w$, enforcing the temporal causality in action-imitation pattern with maximum lag window w. This constraint is implemented by a binary mask $Mask_{u,t}$ which is 1 only if t lies in the allowable range for u, and $-\infty$ otherwise. Subsequently the learned projections of S and E are denoted as $Q_u = W_Q \cdot s_u$, $K_t = W_K \cdot e_t$ and $V_t = W_V \cdot e_t$. Then the masked attention score and the attention weights are given by:

$$\text{logits: } e_{u,t} = \frac{1}{\sqrt{d}} Q_u K_t, \text{ with mask } M_{u,t}$$
$$\text{weights: } A_{u,t} = \frac{\exp(e_{u,t}/\tau)}{\sum_{t'=0}^{T-1} \exp(e_{u,t'}/\tau)}, \tag{1}$$

where τ is a temperature parameter.

Given the final alignment matrix $A \in \mathbb{R}^{T\times T}$, the warped experimenter feature sequence is produced by applying these attention weights A to the value vectors as:

$$E_{warp} = \sum_t A_{u,t} V_t \in \mathbb{R}^{T\times d_{enc}} \tag{2}$$

which is now time-aligned to the subject's timeline. In parallel, to quantify the temporal delays, we compute the expected alignment position for each subject frame u. The expected matched time $t^*(u)$ represents the weighted average of experimenter frame indices:

$$t^*(u) = \sum_{t=0}^{T-1} A_{u,t} \cdot t \tag{3}$$

which gives the expected value of the experimenter frame index that subject frame u is aligned with. The lag at frame u is then computed as:

$$\ell_u = \max(u - t^*(u), 0) \tag{4}$$

This represents the temporal delay. While the window constraint $0 \le u - t \le w$ restricts the alignment search space to positive delays up to w frames, the max operation further ensures non-negative lags by filtering any remaining violations. These ensure the prior constraint in imitation tasks that the imitator's actions should lag behind rather than lead the experimenter's demonstrations. The lag curve $\ell = [\ell_0, \ldots, \ell_{T-1}] \in \mathbb{R}^T$ serves as one of the discrepancy

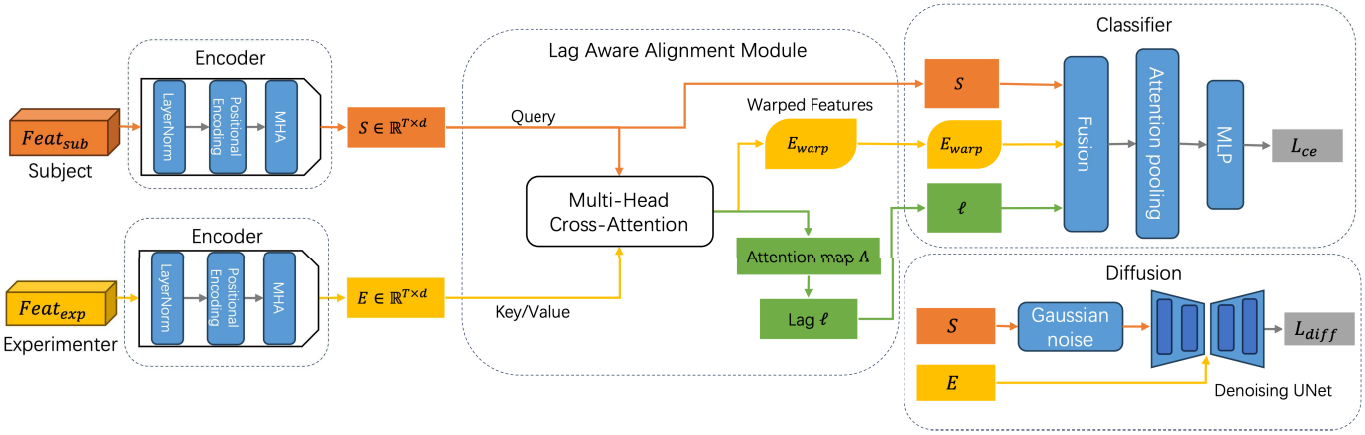

Fig. 5. Our proposed lag-aware dual-stream (LADS) framework that consists of four key components, dual-stream temporal encoders, lag-aware alignment module, discrepancy-based classification, and diffusion regularizer. Dual-stream temporal encoders first extract subject representations S and experimenter representations E. The Lag-Aware Cross-Attention module produces a warped experimenter sequence E_{warp} and a per-frame lag curve ℓ. These signals are fused in the discrepancy classifier via attention pooling for ASD classification. During training only, a diffusion-style regularizer injects noise into S and learns to denoise it conditioned on E, encouraging consistent interaction-aware representations.

features. We normalize the lag values by the window size for scale consistency as

$$\tilde{\ell}_u = \min(\ell_u / w, 1.0) \tag{5}$$

and this normalized lag is concatenated with other discrepancy features for classification. Furthermore, the lag pattern across frames reveals distinctive imitation characteristics, providing valuable insights into ASD-related imitation abilities.

In practice, the maximum lag window w is grounded in the expected response delay in structured imitation tasks. As clinical observations suggest that children's imitative responses typically follow experimenter demonstrations within several hundred milliseconds. At 30 fps this corresponds to a delay of roughly 6–18 frames. We therefore treat w as a hyperparameter and evaluate $w \in \{4, 8, 12, 16, 20\}$ on the validation folds. As reported in Section VI-C, the performance is relatively stable across this range and peaks at $w = 12$, which is adopted as the default setting throughout our experiments.

3) Diffusion-Style Feature Regularization: While the LAA module addresses temporal misalignment, we further enhance generalization through a diffusion-based regularization inspired by the Denoising Diffusion Probabilistic Models (DDPM) [53] and a recent work on applying generative models for discriminative training [54]. Unlike traditional approaches that generate synthetic data samples, we employ a diffusion process in the latent feature space as a regularizer, improving the generalization of the encoded representation without requiring additional data generation or manual parameter tuning.

Compared with generic regularizers such as weight decay or dropout that impose parameter or activation-wise reduction, our diffusion regularizer is data-dependent and operates directly in the representation space, encouraging semantically meaningful invariances rather than purely norm-based constraints.

Our diffusion-based regularization approach operates on S, and treats E as conditioning information. During training, we implement a forward diffusion process that progressively adds Gaussian noise to S across K discrete timesteps. Following the simplification adopted in DDPM, we do not introduce learnable parameters in forward diffusion process. The noise schedule $\{\beta_k\}_{k=1}^{K}$ is set as linear with $\beta_1 = 10^{-4}$ and $\beta_K = 0.02$. Then the cumulative noise level $\bar{\alpha}_k$ can be computed as:

$$\bar{\alpha}_k = \prod_{i=1}^{k} (1 - \beta_i) \tag{6}$$

For each mini-batch we sample $k \sim \text{Uniform}(1, \ldots, K)$ and draw Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ as:

$$\mathbf{S}_k = \sqrt{\bar{\alpha}_k}\, \mathbf{S} + \sqrt{1 - \bar{\alpha}_k}\, \boldsymbol{\varepsilon} \tag{7}$$

The conditioning signal E is mean-pooled as $\bar{\mathbf{e}} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{e}_t$ and is concatenated with a time step embedding $\phi(k)$ and mean-pooled subject feature $\bar{\mathbf{s}}$ to the conditioning vector:

$$\mathbf{c} = \mathbf{W}_c(\bar{\mathbf{s}} \oplus \bar{\mathbf{e}} \oplus \phi(k)) \tag{8}$$

As the denoiser, a lightweight temporal U-Net UNet_θ is implemented and used to predict the injected noise with $S_k^\top$ and $\mathbf{c}$:

$$\hat{\epsilon} = \text{UNet}_\theta(S_k, \mathbf{c}), \tag{9}$$

where the conditioning is injected with an AdaGN [55] style as $\mathbf{c}$ is mapped to scale and shift parameters, and applied to each intermediate feature map. This channel-wise modulation forces the network to restore the subject sequence while being continuously guided by the experimenter context.

Finally, the diffusion regularization loss is derived as $L_{\text{diff}} = \text{MSE}(\hat{\epsilon}, \epsilon)$. Minimising L_{diff} drives the encoders to yield representations in which the subject's motion can be recovered from the experimenter demonstration and remains correlated to the experimenter's context.

4) Discrepancy-Based Classification: With aligned sequences and lag information, we construct comprehensive discrepancy features that capture multiple aspects of imitation quality. We fuse the features of the encoded subject sequence S, the warped experimenter sequence E_{warp} and the lag features ℓ to represent the behavioral discrepancies between

the children and the experimenters. For each time step u, the discrepancy vector D_u is computed with the original features of (1) subjects and experimenters S and E_{warp} that captures actual poses and motions, (2) the normalized lag $\tilde{\ell}$ that serves as the imitation lag feature, (3) the cosine similarity $\cos(S, E_{warp}) = \frac{S E_{warp}}{\|S\| \cdot \|E_{warp}\|}$ that represents the difference of action between subjects and experimenters. The components are concatenated as:

$$D_u = S \oplus E_{warp} \oplus \cos(S, E_{warp}) \oplus \tilde{\ell} \tag{10}$$

The D_u is then passed through a small projector network f with two fully connected layers and nonlinear activations to derive a transformed discrepancy embedding $H \in \mathbb{R}^{d_{model} \times T}$.

To aggregate the discrepancy features into an overall representation, we employ an attention-guided pooling mechanism that a small attention head computes a scalar weight α for each frame as:

$$a_u = \text{softmax}(\mathbf{w}^\top \phi(\mathbf{W}_{\text{attn}} \mathbf{h}_u)), \tag{11}$$

where ϕ is a nonlinear activation GELU, W_{attn}, $\mathbf{w}$ are learnable parameters, and $h_u \in \mathbb{R}^{d_{\text{model}}}$ denotes the embedded discrepancy vector from H at frame u. We then form the attentive pooled feature:

$$\mathbf{z}_{\text{attn}} = \sum_{u=0}^{T-1} a_u \cdot \mathbf{h}_u \tag{12}$$

The pooled discrepancy representation $\mathbf{z}$ is fed into the final classifier to predict the binary outcome (ASD or TD). The classifier is a simple MLP, trained via the standard binary cross-entropy loss L_{ce}. Then L_{ce} is combined with L_{diff} to obtain the overall training objective:

$$L = L_{\text{ce}} + \lambda L_{\text{diff}}, \tag{13}$$

where λ is the combination weight. The gradients from L_{diff} affect only the encoders and diffusion UNet which is not used in the inference, while the L_{ce} is used to update all parameters of the encoder, the LAA module and the classifier.

5) Practical Considerations: From a practical perspective, the design of our framework keeps the inference graph lightweight. All models in our benchmark, including LADS and the baselines operate on the pre-extracted skeleton and motion features, so the most computationally expensive video decoding and feature extraction stages are shared and not specific to our method. The additional cost of LADS relative to a vanilla dual-stream encoder is mainly from the lag-aware cross-attention and the diffusion regularizer. The Lag-Aware Alignment module restricts cross-attention to a local temporal window of size w, which reduces the alignment complexity from $O(T^2)$ to $O(Tw)$, with $w \ll T$ in all our experiments, thereby avoiding the quadratic cost of unconstrained attention. In addition, the diffusion-based regularizer is only activated during training and is completely removed at inference, which only involves the dual-stream encoders, the LAA module and the discrepancy-based classifier. Consequently, LADS adds only a slight computational burden compared with standard dual-stream transformers, while delivering consistently higher accuracy and interpretable lag profiles, which is acceptable for offline ASD screening scenarios that value diagnostic robustness.

V. Experimental Setup

A. Training Schemes

We employ a stratified 5-fold cross-validation training scheme to evaluate performances of the models. The entire dataset was split into 5 folds, ensuring subject independence as no subject's data appeared in more than one fold, which was strictly enforced by checking and avoiding any overlap of subject IDs between training, validation, and test sets of each fold. In each fold, the data was further divided into training, validation, and test subsets, with 20% of the total data used as the test set and 15% of the training data used as a validation set for model selection. The model was trained from a fixed number of epochs (60 epochs per fold without early stopping) and using the AdamW optimizer with learning rate = 3×10^{-4}. During training, the best model checkpoint was determined by the highest validation F1 score. Finally, we report the model's average test accuracy and F1-score across the 5 folds along with the standard deviation, providing a robust estimate of the model's generalization performance. We ensured all models of baseline classifiers presented in the next section use the same subject-independent 5-fold cross-validation as our LADS framework for evaluation.

B. Baseline Classifiers

To evaluate the dataset and validate our LADS advantages, we compare against several baseline classifiers. Since most existing deep learning methods are designed for classification on single-view data, we train and validate these baselines using the data from only the subject's perspective.

1) Traditional Machine Learning Methods: To verify that extracted feature sequences retain sufficient discriminative information compared to naive pixel-level statistics, and to incorporate machine learning approaches commonly used in classical ASD analysis, we adopted the methodology from [11], which combines hand-crafted features with classical machine learning techniques. The work [11] extracts motion energy and cross-wavelet coherence features from raw frame pixels to quantify interpersonal motor coordination (IMC) as synchronization features for dual-view videos. An SVM classifier with RBF kernel is trained on these hand-crafted synchronization features.

All subsequent baselines are trained on the extracted feature sequences released with our dual-view imitation dataset, focusing exclusively on sequence-level models. The motion analysis approach [11], which operates on raw pixel-level statistics, serves as a representative baseline from traditional ASD assessment practices used in psychology researches. We do not include additional video-based classification baselines, as our focus is on modeling imitation-specific temporal dynamics rather than benchmarking video analysis architectures.

2) Deep Learning Methods: In addition to comparisons with traditional ASD analysis methods [11], we also explored the application of classical deep learning models — known for their generalizability and performance in feature extraction and classification across various tasks — to ASD classification. Specifically, we implemented a VGG-style [56] 1D CNN that adapts the classical VGG design through stacked convolutional

blocks with progressively increasing filter sizes from 64 to 256 channels, where each block contains paired convolutions followed by pooling operations. This establishes a baseline for local temporal pattern recognition without any cross-stream interaction. Moreover, we also evaluate a ResNet-34 model [33], another widely-used CNN architecture known for its effectiveness in image classification. This model includes an initial convolutional layer (64 channels) and pooling, then residual blocks at increasing widths (64, 128, 256, 512 channels) and shortcut connections. Global average pooling and a fully-connected layer yield the final ASD/TD prediction.

Based on the single-stream data from only the subject, we also explore whether advanced temporal modeling strategies can compensate for the absence of dual-stream information. PatchTST [40], originally proposed for time-series forecasting, segments the input sequence into non-overlapping patches of 16 frames each, treating each patch as a discrete token. This channel-independent architecture then applies Transformer self-attention over the patch embeddings to capture long-range temporal dependencies, followed by a global average pooling for aggregation. The learned representations were originally introduced for long-term forecasting and we adapt it for ASD classification by adding a linear layer. The patching strategy allows the model to operate at multiple temporal scales simultaneously, making it a strong baseline for assessing the sophisticated single-stream modeling.

We also include BlockGCN [41] — a state-of-the-art graph convolutional network that treats poses from each frame as a spatial graph where joints form nodes connected by anatomical constraints — as a specialized baseline for skeleton feature data. For our ASD classification task, we feed 33-joint skeleton sequence ((T,33,4) with 3D coordinates and confidence) from subjects into the BlockGCN. It performs temporal graph convolutions and outputs a representation, which we feed to a linear classifier for ASD classification.

For dual-stream processing, we include the Dyadic Interaction Model (DIM) [46] — which serves as a comparative framework for dual-stream fusion — as a baseline. DIM is originally designed for dyadic facial animation and we adapt an open implementation for our task. It consists of separate transformer encoders for the experimenter's and subject's sequences, each producing a downsampled latent sequence. These latent features are then passed through a vector quantization layer for each stream, effectively discretizing the representations. The quantized EXP and SUB sequences are concatenated and fed into a cross-modal transformer that fuses the two streams. A classification head uses pooled output from the transformer to predict ASD versus TD. We train this model with a composite loss — cross-entropy plus the VQ commitment losses. Essentially, this baseline learns a joint embedding of the dual streams using a transformer, providing a relevant comparative baseline for our LAA module.

VI. Experimental Results

A. Benchmarking Results

We evaluate our LADS framework with diffusion regularization against established baseline methods using the four feature modalities (skeleton, dense flow, sparse flow, and heatmap) provided in our proposed dual-view imitation dataset. The comparative results, summarized in Table I, demonstrate the effectiveness of our approach for ASD classification.

1) Overall Performance Comparison: LASD consistently outperforms all baselines across all feature modalities, achieving accuracy improvements ranging from 2.6% to 7.7% over the strongest competing method across the four features. Most notably, the skeleton-based implementation of LADS achieves 85.6% accuracy (±4.9%), establishing a new state-of-the-art performance on this dataset and demonstrating a 5.7% improvement over the best-performing baseline (DIM at 79.9%). This substantial gain validates our core hypothesis that explicitly modeling temporal lag patterns while enforcing representation consistency through diffusion regularization can effectively capture critical aspects of imitative behavior that conventional approaches overlook.

These benchmarking results position LADS as the new state-of-the-art method for ASD classification on dual-view imitation data, validating both the importance of modeling temporal correspondence in imitation tasks and the value of diffusion-based regularization for learning discriminative behavioral representations.

2) Single-Stream Versus Dual-Stream Analysis: Experimental results demonstrate that dual-stream models (DIM and LADS) — which take data from both subjects and experiments — consistently outperform single-stream classifiers across all feature modalities. Among single-stream approaches, ResNet achieves the highest accuracy on skeleton features (78.9%). However, this performance still lags behind dual-stream approaches, falling short of DIM by approximately 1% and LADS by 6.6% on skeleton features, showing that explicitly incorporating both experimenter and subject views yields gains, confirming the value of our dual-view dataset for capturing interaction cues. This performance gap widens further for motion-based features (dense and sparse optical flow), where single-stream models struggle to effectively distinguish subject movements without the contextual reference of experimenter demonstrations.

On top of this, our LADS framework further improves accuracy to 85.6%, achieving a 5.7% absolute gain over DIM on skeleton features, and 2.6–7.7% gains on the optical flow and heatmap modalities. Since all models share the dyadic input features and training protocol, these improvements can be attributed to the proposed lag-aware alignment and diffusion-based regularization rather than to differences in data or supervision.

3) Sophisticated Single-Stream Models: Notably, more sophisticated single-stream architectures do not yield better performance. PatchTST, despite its advanced patch-based temporal modeling and self-attention mechanisms, underperforms simpler CNN architectures on most feature modalities, achieving only 70.6% compared to ResNet's 78.9% on skeleton features. Similarly, BlockGCN — a state-of-the-art graph convolution network specifically designed for skeletal action recognition — achieves only 70.4% accuracy. These counterintuitive results suggests that the inductive biases of these specialized architectures may not align well with the

TABLE I

COMPARISON OF MODEL PERFORMANCE ACROSS DIFFERENT FEATURE TYPES AND DATA PERSPECTIVES. BOLD VALUES INDICATE THE BEST ACCURACY AND F1-SCORE FOR EACH FEATURE TYPE

Model	Data Perspective	Feature Type	Accuracy	F1-Score
MEA [11]	SUB only	Hand crafted IMC	0.540 ± 0.073	0.552 ± 0.067
CNN (VGG-Style) [56]	SUB only	Skeleton	0.749 ± 0.064	0.748 ± 0.064
		Dense flow	0.730 ± 0.051	0.729 ± 0.050
		Sparse flow	0.699 ± 0.049	0.698 ± 0.048
		Heatmap	0.728 ± 0.035	0.730 ± 0.033
ResNet [33]	SUB only	Skeleton	0.789 ± 0.022	0.785 ± 0.023
		Dense flow	0.617 ± 0.030	0.622 ± 0.026
		Sparse flow	0.627 ± 0.026	0.625 ± 0.026
		Heatmap	0.734 ± 0.057	0.736 ± 0.056
PatchTST [40]	SUB only	Skeleton	0.706 ± 0.033	0.709 ± 0.031
		Dense flow	0.621 ± 0.031	0.621 ± 0.032
		Sparse flow	0.549 ± 0.025	0.552 ± 0.025
		Heatmap	0.690 ± 0.054	0.690 ± 0.053
BlockGCN [41]	SUB only	Skeleton	0.704 ± 0.108	0.705 ± 0.106
DIM [46]	EXP + SUB	Skeleton	0.799 ± 0.054	0.799 ± 0.055
		Dense flow	0.792 ± 0.105	0.792 ± 0.105
		Sparse flow	0.693 ± 0.139	0.679 ± 0.207
		Heatmap	0.764 ± 0.024	0.759 ± 0.031
LADS (Ours)	EXP + SUB	Skeleton	**0.856 ± 0.049**	**0.850 ± 0.048**
		Dense flow	**0.818 ± 0.070**	**0.810 ± 0.075**
		Sparse flow	**0.770 ± 0.055**	**0.761 ± 0.056**
		Heatmap	**0.790 ± 0.035**	**0.777 ± 0.041**

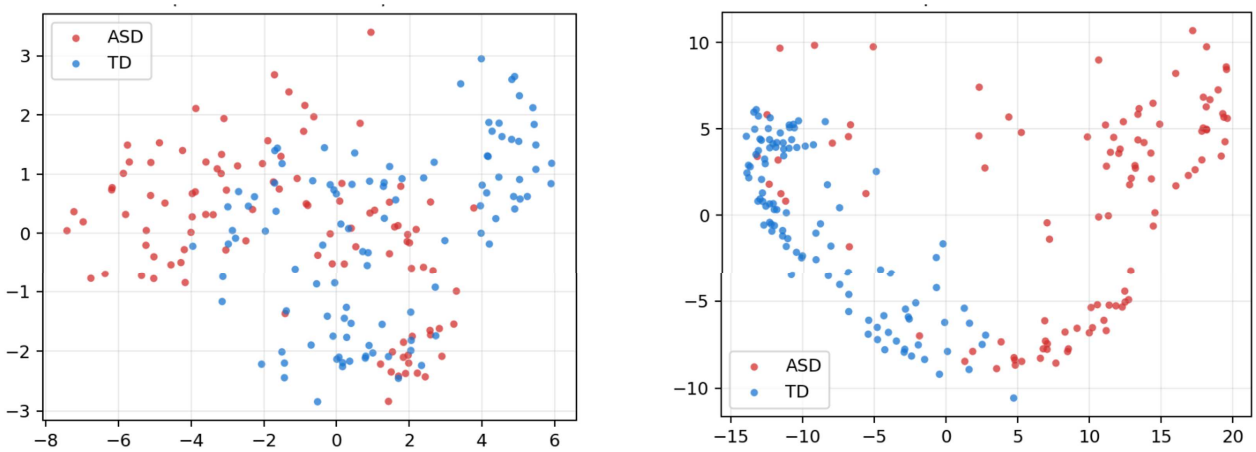

Fig. 6. Feature-space discriminability via PCA. Each point corresponds to one test sample; colors denote classes (ASD in red, TD in blue). *Left*: PCA of the raw SUB features. *Right*: PCA of the learned discrepancy representation **z**.

specific challenges of imitation assessment. PatchTST's patch-based tokenization, while effective for long-term forecasting tasks, may average out brief hesitations and delay in frame level that are diagnostically relevant for ASD classification. BlockGCN's graph convolutions, optimized for recognizing distinct action categories, may over-emphasize spatial joint configurations while under-weighting temporal dynamics.

4) Variance and Stability Analysis: As discussed above, the BlockGCN may be sensitive to skeleton tracking quality variations in our proposed dataset that was collected in a naturalistic classroom environment, reflected by its high variance (±10.8%). Moreover, DIM also shows high variance on sparse flow features (±13.9%), suggesting instability when processing noisy motion estimates.

B. Discrimination of the Learned Discrepancy Representation

We analyze feature-space separability using Principal Component Analysis (PCA). As shown in Figure 6, when projecting raw SUB features (we only run this examination on skeleton features), ASD and TD largely overlap, suggesting limited discriminability without dyadic reasoning. In contrast, projecting the learned discrepancy representation **z** built in the discrepancy classifier produces clearly separated ASD and TD clusters. This confirms that our complete processing framework — including the encoding, alignment, discrepancy construction, attentive pooling and the diffusion regularizer — effectively transforms raw imitation data into a discriminative representation.

C. Ablation Study

All ablations are performed using skeleton features, since it demonstrated the strongest performance in our benchmark evaluation, making it the most appropriate setting for isolating architectural contributions.

1) Component-Wise Ablation of LADS: We conduct ablation studies on three key components of our framework using the same subject-independent 5-fold cross-validation protocol and training schedule: (1) the diffusion regularizer (L_{diff}), (2) the lag-aware alignment (LAA) module, switching to a direct, non-lag alignment, and (3) the temporal pooling strategy (replacing attention pooling with simple mean pooling). Specifically, to evaluate LAA while maintaining interaction analysis capability, we replace it with a hard Dynamic Time Warping (DWT) approach [57]. DTW finds a warping path with cosine similarity cost matrices and computes a per-frame lag $\ell = \max(u - t_{\text{DTW}}(u), 0)$. This variant removes the learnable lag modeling. In each ablation experiment, only one component is modified, while all others are kept identical to the full LADS configuration.

Removing the diffusion regularization loss ($-L_{\text{diff}}$) results in a 2.4% accuracy decrease, demonstrating its substantial contribution to model performance. The diffusion process serves as an effectively consistency constraint, ensuring that encoding subject features remain recoverable given experimenter context, thereby enhancing the encoder's capacity to capture the interaction patterns.

Replacing our LAA module with direct frame-wise measurement of ℓ (–LAA) reduces accuracy by 1.9% to 83.7%. Although this performance drop is smaller than without diffusion or pooling, it is consistent across all cross-validation folds, confirming that explicit temporal delay modeling improves the discriminative ability of learned representations.

Replacing attention-weighted pooling with simple mean pooling ($-\mathbf{z}_{\text{attn}}$) causes the largest performance degradation (2.8%). In this variant, we compute the final representation by simply averaging all temporal features rather than learning importance weights through the attention mechanism. This substantial drop highlights that the attention mechanism successfully identifies diagnostically relevant components from the lag and encoded embedding features that are fed into the classifier.

These ablation results validate our architectural design, indicating that each component makes distinct contributions to achieving state-of-the-art performance in ASD classification on dual-view imitation data.

We further visualized the lag feature via normalizing the ℓ and averaging per class. The resulting curves in Figure 7 reveal clear qualitative differences across ablation conditions. Using the full model, the TD curve remains consistently lower with a narrow band, while ASD is shifted upward with a broader band, which illustrates that the model learns a stable, class-specific delay pattern in which TD exhibits smaller and more consistent response latencies than ASD. Removing LAA, leading to insufficient interaction contextual information, produces a pronounced jitter and reduced class separation. Turning off diffusion module yields a counter

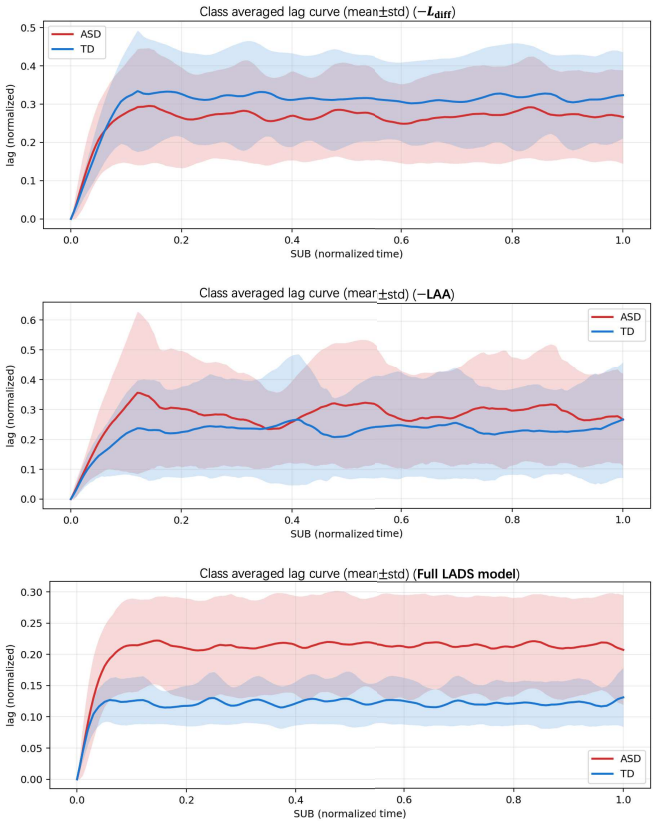

Fig. 7. Class-averaged lag curves ℓ under three settings. Each curve is resampled on the normalized subject timeline and averaged per class (ASD in red, TD in blue). *Top:* no diffusion. *Middle:* no LAA. *Bottom:* full LADS model.

intuitive result where the TD curve rides slightly above that of ASD with a large variation, suggesting that the model may overfit local patterns in noisy TD sequences, which can be suppressed by diffusion regularization. This also explains the overall performance drop in the ($-L_{\text{diff}}$) ablation. Taken together, these curves indicate that LAA effectively capture the discriminative lag features, while diffusion regularization stabilizes the underlying features. We do not include the attention pooling ablation in the lag curve figure because these curves visualize the alignment-level signal ℓ, produced before discrepancy features aggregation.

2) Effect of the Lag Window Size: The local window size w controls the maximum allowed temporal delay between experimenter and subject frames in the LAA module, and is therefore a key hyperparameter for modeling imitation latency. To study its effect, we fix all other settings and vary $w \in \{4, 8, 12, 16, 20\}$ frames. For each choice of w, we re-train LADS from scratch using the same 5-fold protocol.

Table III reports the performance under different window sizes. We observe that a moderate window size (e.g., $w = 12$) achieves the best accuracy and F1-score. When w is too small, the model cannot capture most genuine delays in the child's responses and tends to force the synchronous alignments, which underestimates atypical imitation latencies in ASD. Conversely, a large window size allow the attention to match subject frames with distant experimenter actions, making the

TABLE II
ABLATION RESULTS

Variant	Accuracy	F1-Score
Baseline (LAA + $\mathbf{z}_{\text{attn}}$ + L_{diff})	0.856	0.850
Remove the diffusion regularizer ($-L_{\text{diff}}$)	0.832	0.823
DTW alignment (–LAA)	0.837	0.831
Mean pooling ($-\mathbf{z}_{\text{attn}}$)	0.828	0.822

TABLE III
EFFECT OF THE LAG WINDOW SIZE w IN THE LAA MODULE (SKELETON FEATURES)

Window size w	Accuracy	F1-Score
4	0.835	0.821
8	0.847	0.839
12	**0.856**	**0.850**
16	0.846	0.835
20	0.848	0.841

TABLE IV
THE MEAN TEST ACCURACY AND F1-SCORE OVER FIVE FOLDS FOR COMPARISON OF REGULARIZATION STRATEGIES

Regularization strategy	Accuracy	F1-Score
None (cross-entropy only)	0.832	0.823
Weight decay, $\lambda = 10^{-2}$	0.848	0.838
Dropout, $p = 0.5$	0.842	0.833
Diffusion regularizer L_{diff}	**0.856**	**0.850**

estimated lag curve noisier and less discriminative. Based on this study, we set $w = 12$ as the default configuration in all other experiments.

3) Comparison of Regularization Strategies: To further clarify the role of the proposed diffusion-based regularizer, we compare it against more conventional regularization schemes applied to the same LADS architecture. With setting of (cross-entropy only), the model is trained with cross-entropy loss without any explicit regularizer. This configuration corresponds to the ($-L_{\text{diff}}$) variant in Table II. As comparisons, we add ℓ_2 weight decay with coefficient $\lambda = 10^{-2}$ to all trainable parameters and insert a dropout layer with rate $p = 0.5$ to the temporal encoding module.

From Table IV we observe that both conventional strategies, which are weight decay and dropout, indeed provide noticeable improvements over using cross-entropy alone (accuracy increases from 0.832 to 0.848 and 0.842, respectively). This confirms that the LADS classifier benefits from standard regularization. However, the diffusion-based regularizer still achieves the best performance, improving accuracy by 0.8% and F1-score by 1.4% points over the second-best configuration with weight decay, indicating that the proposed diffusion formulation is particularly well aligned with the dyadic imitation modeling task. Unlike norm-based penalties or random feature dropping, our diffusion module injects structured Gaussian noise into the subject stream and trains a denoising network conditioned on the experimenter features. This latent-space denoising objective explicitly enforces cross-stream consistency and encourages the encoder to discard noise while retaining diagnostically relevant interaction patterns.

VII. CONCLUSION

In this work, we introduced a novel dual-view imitation dataset and proposed the lag-aware dual-stream (LADS) framework to advance automated ASD classification through imitation task analysis. Our dataset, which is the first of its kind, provides synchronized recordings of both experimenter demonstrations and child responses during structured imitation tasks, enabling analysis of when and how children with ASD imitate actions compared to typically developing peers. Building on this resource, our LADS framework leverages cross-attention to align the dual-view streams and explicitly quantify the child's response delay. We further incorporate a diffusion-style feature regularization to guide the encoder with coherence between the child's representation and the context from the experimenter.

Experimental results demonstrate that integrating dual-perspective interactions and temporal dynamics yields substantial improvements over traditional single-view methods. LADS achieved state-of-the-art performance in distinguishing ASD from TD participants, outperforming both single-stream deep learning models and an existing dual-stream method. These findings underscore the value of capturing interpersonal coordination cues, such as timing alignment and pose discrepancy, for identifying ASD-related behavioral patterns. The learned lag profiles are consistent with clinical observations of delayed and atypical imitation in ASD, indicating that our computational approach captures meaningful behavioral markers.

*Limitation and Future Work:*The current release contains 1054 dyadic clips from 84 participants collected at a single clinical center. While this provides a unique resource for dual-view imitation analysis, the geographic scope and sample size are inherently limited. To mitigate demographic and appearance biases, we do not distribute raw RGB video clips, instead, we provide only mid-level features such as skeletons, optical flow, and pose heatmaps. These representations abstract away background, clothing, and camera-specific details. Nevertheless, we acknowledge that fully assessing the robustness of multi-group of participants will require additional patients group or independent datasets. Creating and introducing such multi-group dyadic imitation datasets is an important direction for future work and will be greatly facilitated by the public release of our current benchmark.

REFERENCES

[1] D. Fein, L. Green, and L. Waterhouse, "- autism and pervasive developmental disorders," in *The Disorders*, H. S. Friedman, Ed., New York, NY, USA: Academic, 2001, pp. 97–106. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780122678059500110

[2] J. H. G. Williams, A. Whiten, and T. Singh, "A systematic review of action imitation in autistic spectrum disorder," *J. Autism Develop. Disorders*, vol. 34, no. 3, pp. 285–299, Jun. 2004.

[3] B. Ingersoll, "The social role of imitation in autism: Implications for the treatment of imitation deficits," *Infants Young Children*, vol. 21, no. 2, pp. 107–119, 2008.

[4] J. M. Rehg et al., "Decoding children's social behavior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3414–3421.

[5] R. K. Greene, I. Vasile, K. R. Bradbury, A. Olsen, and S. W. Duvall, "Autism diagnostic observation schedule (ADOS-2) elevations in a clinical sample of children and adolescents who do not have autism: Phenotypic profiles of false positives," *Clin. Neuropsychologist*, vol. 36, no. 5, pp. 943–959, Jul. 2022.

[6] P. Pandey, P. Ap, M. Kohli, and J. Pritchard, "Guided weak supervision for action recognition with scarce data to assess skills of children with autism," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 1, pp. 463–470.

[7] A. Zunino et al., "Video gesture analysis for autism spectrum disorder detection," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3421–3426.

[8] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, "Use of artificial intelligence to shorten the behavioral diagnosis of autism," *PLoS ONE*, vol. 7, no. 8, Aug. 2012, Art. no. e43855.

[9] M. Duda, J. A. Kosmicki, and D. P. Wall, "Testing the accuracy of an observation-based classifier for rapid detection of autism risk," *Translational Psychiatry*, vol. 4, no. 8, p. e424, Aug. 2014.

[10] C. Liu, T. Tang, K. Lv, and M. Wang, "Multi-feature based emotion recognition for video clips," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 630–634.

[11] Z. Zhao, X. Zhang, X. Zhang, X. Qu, X. Hu, and J. Lu, "Interpersonal motor coordination in children with autism and the establishment of machine learning models to objectively classify children with autism and typical development," *IRBM*, vol. 45, no. 5, Oct. 2024, Art. no. 100838.

[12] A. Rani and Y. Verma, "Activity-based early autism diagnosis using a multi-dataset supervised contrastive learning approach," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 7773–7782.

[13] K. Sun, L. Li, L. Li, N. He, and J. Zhu, "Spatial attentional bilinear 3D convolutional network for video-based autism spectrum disorder detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3387–3391.

[14] M. Sevlever and J. M. Gillis, "An examination of the state of imitation research in children with autism: Issues of definition and methodology," *Res. Develop. Disabilities*, vol. 31, no. 5, pp. 976–984, Sep. 2010.

[15] J. L. Mussey and L. G. Klinger, "Imitation in ASD: Performance on an imitation choice task," *Res. Autism Spectr. Disorders*, vol. 73, May 2020, Art. no. 101530.

[16] I. M. Smith and S. E. Bryson, "Gesture imitation in autism: II. symbolic gestures and pantomimed object use," *Cognit. Neuropsychol.*, vol. 24, no. 7, pp. 679–700, Oct. 2007.

[17] I. Akin-Bulbul and S. Ozdemir, "Imitation performance in children with autism and the role of visual attention in imitation," *J. Autism Develop. Disorders*, vol. 53, no. 12, pp. 4604–4617, Dec. 2023.

[18] H. Stieglitz Ham, A. Bartolo, M. Corley, G. Rajendran, A. Szabo, and S. Swanson, "Exploring the relationship between gestural recognition and imitation: Evidence of dyspraxia in autism spectrum disorders," *J. Autism Develop. Disorders*, vol. 41, no. 1, pp. 1–12, Jan. 2011.

[19] J. W. Du Bois, R. P. Hobson, and J. A. Hobson, "Dialogic resonance and intersubjective engagement in autism," *Cognit. Linguistics*, vol. 25, no. 3, pp. 411–441, Aug. 2014.

[20] R. Bernier, G. Dawson, S. Webb, and M. Murias, "EEG μrhythm and imitation impairments in individuals with autism spectrum disorder," *Brain Cognition*, vol. 64, no. 3, pp. 228–237, Aug. 2007.

[21] M. Sano, K. Yamaguchi, R. Fukatsu, and M. Hoshiyama, "Action performance in children with autism spectrum disorder at preschool age: A pilot study," *Int. J. Develop. Disabilities*, vol. 66, no. 4, pp. 289–295, Aug. 2020.

[22] B. Du et al., "Higher or lower? Interpersonal behavioral and neural synchronization of movement imitation in autistic children," *Autism Res.*, vol. 17, no. 9, pp. 1876–1901, Sep. 2024.

[23] E. Möricke, J. K. Buitelaar, and N. N. J. Rommelse, "Do we need multiple informants when assessing autistic traits? The degree of report bias on offspring, self, and spouse ratings," *J. Autism Develop. Disorders*, vol. 46, no. 1, pp. 164–175, Jan. 2016.

[24] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Res.*, vol. 9, no. 8, pp. 888–898, Aug. 2016.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.

[26] G. Dawson and G. Sapiro, "Potential for digital behavioral measurement tools to transform the detection and diagnosis of autism spectrum disorder," *JAMA pediatrics*, vol. 173, no. 4, p. 305, 2019.

[27] C. Grossard et al., "Children with autism spectrum disorder produce more ambiguous and less socially meaningful facial expressions: An experimental study using random forest classifiers," *Mol. Autism*, vol. 11, no. 1, pp. 1–14, Dec. 2020.

[28] P. R. K. Babu et al., "Exploring complexity of facial dynamics in autism spectrum disorder," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 919–930, Apr. 2023.

[29] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, "Discriminative few shot learning of facial dynamics in interview videos for autism trait classification," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1110–1124, Apr. 2023.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *IEEE Pervasive Comput.*, vol. 13, no. 2, pp. 84–87, Apr. 2014.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[35] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2969–2978.

[36] L. Wang et al., "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[37] C. Li et al., "Unsupervised feature enrichment and fidelity preservation learning framework for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 7116–7128, Jul. 2025.

[38] S. Zahan, Z. Gilani, G. M. Hassan, and A. Mian, "Human gesture and gait analysis for autism detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 3328–3337.

[39] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[40] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–24.

[41] Y. Zhou, X. Yan, Z.-Q. Cheng, Y. Yan, Q. Dai, and X.-S. Hua, "BlockGCN: Redefine topology awareness for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 2049–2058.

[42] D. Pei, D. Huang, L. Kong, and Y. Wang, "Key role guided transformer for group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7803–7818, Dec. 2023.

[43] Z. Wang et al., "Knowledge augmented relation inference for group activity recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 11644–11656, Nov. 2024.

[44] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome," *J. Consulting Clin. Psychol.*, vol. 79, no. 3, pp. 284–295, 2011.

[45] M. Cheng et al., "Computer-aided autism spectrum disorder diagnosis with behavior signal processing," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2982–3000, Oct. 2023.

[46] M. Tran, D. Chang, M. Siniukov, and M. Soleymani, "Dim: Dyadic interaction modeling for social behavior generation," in *Proc. Eur. Conf. Comput. Vis.*Cham, Switzerland: Springer, 2024, pp. 484–503.

[47] B. Roehr, "American psychiatric association explains DSM-5," *BMJ*, vol. 346, no. 6, p. 3591, Jun. 2013.

[48] B. Ingersoll, "Brief report: Pilot randomized controlled trial of reciprocal imitation training for teaching elicited and spontaneous imitation to children with autism," *J. Autism Develop. Disorders*, vol. 40, no. 9, pp. 1154–1160, Sep. 2010.

[49] G. Goldenberg, "Matching and imitation of hand and finger posturesin patients with damage in the left or right hemispheres," *Neuropsychologia*, vol. 37, no. 5, pp. 559–566, Feb. 1999.

[50] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," 2020, *arXiv:2006.10204*.

[51] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Int. joint Conf. Artif. Intell.*, Vancouver, BC, Canada, Aug. 1981, pp. 674–679. [Online]. Available: https://hal.science/hal-03697340

[52] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds., Berlin, Germany: Springer, 2003, pp. 363–370.
[53] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2024, pp. 6840–6851.
[54] T. Asakura, N. Inoue, and K. Shinoda, "Diffusion-based generative regularization for supervised discriminative learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 8915–8926.
[55] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NIPS*, vol. 34, 2021, pp. 8780–8794.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[57] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. 3rd Int. Conf. Knowl. discovery data mining*, 1994, pp. 359–370.

Wenqi Ji received the B.Sc. degree in computer science and technology from Beihang University, Beijing, China, in 2018, and the M.Sc. degree in computer science from Queen Mary University of London, London, U.K., in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include affective computing and deep learning.

Tong Liu received the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2023. She is currently a Post-Doctoral Fellow with the Department of Computer Science and Technology, Tsinghua University, Beijing, China.

Xue Yang received the Ph.D. degree in management science from Fuzhou University in 2023. She is a Post-Doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University. Her research interests include cognitive neuroscience and computational psychiatry.

Ying Guo received the B.S. degree from the University of Electronic Science and Technology of China in 2014 and the Ph.D. degree from Shanghai Jiao Tong University in 2019. She is currently a Lecture with the School of Computer Science, North China University of Technology. Her research interests include machine learning, reinforcement learning, learning automata and their applications, computer communication networks, and optimization problems.

Xin Chen received the B.Sc. degree in computer science and technology from Beihang University, Beijing, China. He performed this work as a Research Intern with the Department of Computer Science and Technology, Tsinghua University. His research interests include cognitive computation and deep learning.

Pei Yang received the bachelor's degree from Inner Mongolia University, China, in 2009, and the master's degree from the Minzu University of China in 2016. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Technology, Tsinghua University. His research interests include emotion recognition and machine learning.

Kaiyun Li is currently a Professor with the School Education and Psychology, University of Jinan. Her research interests include social interaction deficit of ASD children and machine learning methods using for ASD detection.

Yong-Jin Liu (Senior Member, IEEE) received the B.Eng. degree from Tianjin University, China, in 1998, and the M.Phil. and Ph.D. degrees from The Hong Kong University of Science and Technology, Hong Kong, China, in 2000 and 2004, respectively. He is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include machine emotional intelligence and pattern analysis. He is also an Associate Editor of IEEE TRANSACTIONS ON AFFECTIVE COMPUTING. For more information, visit http://cg.cs.tsinghua.edu.cn/people/Yongjin/Yongjin.htm