Computerized Assessment of Motor Imitation for Distinguishing Autism in Video (CAMI-2DNet)

Kaleab A. Kinfu, *Student Member, IEEE*, Carolina Pacheco, *Student Member, IEEE*, Alice D. Sperry, Deana Crocetti, Bahar Tunçgenç, Stewart H. Mostofsky, René Vidal, *Fellow, IEEE*

***Abstract*— Motor imitation impairments are commonly reported in individuals with autism spectrum conditions (ASCs), suggesting that motor imitation could be used as a phenotype for addressing autism heterogeneity. Traditional methods for assessing motor imitation are subjective and labor-intensive, and require extensive human training. Modern Computerized Assessment of Motor Imitation (CAMI) methods, such as CAMI-3D for motion capture data and CAMI-2D for video data, are less subjective. However, they rely on labor-intensive data normalization and cleaning techniques, and human annotations for algorithm training. To address these challenges, we propose CAMI-2DNet, a scalable and interpretable deep learning-based approach to motor imitation assessment in video data, which eliminates the need for ad hoc normalization, cleaning and annotation. CAMI-2DNet uses an encoder-decoder architecture to map a video to a motion representation that is disentangled from nuisance factors such as body shape and camera views. To learn a disentangled representation, we employ synthetic data generated by motion retargeting of virtual characters through the reshuffling of motion, body shape, and camera views, as well as real participant data. To automatically assess how well an individual imitates an actor, we compute a similarity score between their motion encodings, and use it to discriminate individuals with ASCs from neurotypical (NT) individuals. Our comparative analysis demonstrates that CAMI-2DNet has a strong correlation with human scores while outperforming CAMI-2D in discriminating ASC vs NT children. Moreover, CAMI-2DNet performs comparably to CAMI-3D while offering greater practicality by operating directly on video data and without the need for ad hoc normalization and human annotations.**

***Index Terms*— Autism Spectrum Conditions, Behavior Analysis, Motor Imitation Assessment, Motion Analysis in Video Data, Disentangled Motion Representation Learning.**

I. Introduction

AUTISM spectrum conditions (ASCs) are defined by core symptoms of social-communicative difficulties, restricted interests and repetitive behaviors. However, the considerable heterogeneity in ASCs creates challenges for efficient diagnosis and treatment options [1]. Based on a growing amount of neuroscience and behavioral research, one promising way of addressing this heterogeneity is through precise quantification of motor imitation impairments, which are highly prevalent in autism [2]. Motor imitation skills are fundamental for socialization, communication, and acquiring essential skills, particularly during early development and in social interactions [3]. Motor imitation impairments seem specific to ASCs, and not shared with other highly co-occurring conditions, such as Attention-Deficit/Hyperactivity Disorder (ADHD) [4], which are typically challenging to distinguish in clinical settings. Importantly, impairments in motor imitation are associated with core autism symptoms and brain mechanisms involved in social communication and learning [5], [6].

Human Observation Coding (HOC) has been the standard method for assessing motor imitation. HOC relies on trained human coders to directly observe and evaluate individuals' imitation skills [7]. Although it provides valuable insights into specific imitation challenges indicative of ASCs, it has several limitations. First, it is inherently subjective, since it depends on human judgment, potentially introducing biases and inconsistencies. In addition, HOC is labor-intensive, requiring significant time, effort, and trained personnel for accurate analysis and behavior coding. These limitations significantly hinder its scalability and practicality in clinical and home settings.

With the increasing prevalence of ASCs [8], [9] and the growing demand for early and accurate assessments, there is a need for automated and objective assessment tools that address the challenges associated with HOC. The development of tools that are effective and widely applicable offers several potential advantages, including efficiency, objectivity, and scalability. However, the development of such tools faces several challenges, including (i) the diverse range and complex nature of human actions involved in imitation, (ii) the trade-off between sensitivity and specificity in recognizing atypical imitation patterns, and (iii) the need to ensure the adaptability of the tool across diverse settings.

Several automated methods have been proposed to address these challenges [10]–[14]. Among these, motion-capture-based methods, which rely on precise 3-dimensional (3D) motion data, have been proven to be effective. For instance, Computerized Assessment of Motor Imitation (CAMI-3D) [15] uses a Kinect Xbox cameras to collect 3D motion data and applies Dynamic Time Warping (DTW) [16] and linear

K. Kinfu and R. Vidal are with the Center for Innovation in Data Engineering and Science at the University of Pennsylvania. C. Pacheco is with the Department of Biomedical Engineering at Johns Hopkins University. A. Sperry, D. Crocetti and S. Mostofsky are with the Center for Neurodevelopmental and Imaging Research at the Kennedy Krieger Institute. S. Mostofsky is also affiliated with the Department of Neurology and the Department of Psychiatry and Behavioral Sciences at the Johns Hopkins University School of Medicine. B. Tunçgenç is with the Department of Psychology at the Nottingham Trent University. This work was supported by NSF grant 2124277 and 2430816.

regression to produce a similarity score that takes into account variations in both motion trajectories and timing discrepancies.

CAMI-3D has shown promising results in autism, demonstrating high test-retest reliability and surpassing the performance of HOC in effectively discriminating children with ASCs from neurotypical (NT) children as well as from children with ADHD, a highly prevalent condition that is both a differential diagnosis of ASC and a frequent co-occurring diagnosis [4]. However, despite its promising results, CAMI-3D has several limitations. First, it relies on specialized hardware, such as Kinect or similar 3D cameras, which may restrict its scalability and applicability in settings like homes or clinics. Second, obtaining accurate motion coordinates requires extensive manual frame-by-frame data cleaning to obtain accurate motion coordinates. CAMI-3D also depends on hand-crafted normalization techniques to handle variations in body shape (*e.g.*, height, limb length) and slight differences in camera angles. While these techniques may be effective in controlled studies, they may not generalize well to diverse anatomical and pose variations encountered in real-world scenarios. Third, HOC annotations are needed for training, thus requiring continued human input for new action sequences.

To address the dependence on costly 3D motion capture devices and enhance scalability, advances in 2D pose estimation techniques [17]–[19] can be utilized. These methods accurately detect and track skeletal joints from video data captured by widely accessible 2D cameras. Motivated by these advances, the CAMI-2D method [20] employs an off-the-shelf pose estimation network, OpenPose [17], to extract 2D joint trajectories from videos. Exactly as CAMI-3D, CAMI-2D compares joint trajectories using DTW and computes imitation scores through linear regression. However, CAMI-2D inherits CAMI-3D limitations, such as the need for hand-crafted normalization and ongoing human annotations (HOC). Additionally, 2D joint locations are more heavily affected by camera viewpoint due to perspective projection and occlusions. Therefore, comparing 2D trajectories can be misleading, as these trajectories are affected by nuisance factors such as variations in body shape and camera viewpoint.

Recent advances in deep learning offer a more robust and efficient approach to comparing human movements in video data [21], [22]. The key idea is to use a neural network to map the video to a compressed *motion representation* that captures the essence of the movements. This is achieved by using large-scale video data or pose sequences to learn *disentangled motion representations*, *i.e.*, motion representations that are invariant to nuisance factors such as body shape and viewpoint.

For example, [23] proposes a novel approach for decomposing motion data into dynamic and static representations. Originally developed for motion retargeting, this technique utilizes an encoder-decoder network to separate motion data into skeleton-independent dynamic features and skeleton-dependent static features. The model in [24] further decomposes a pose sequence into individual body parts, generating representations for each part separately. This results in motion representations that are suitable for measuring the similarity between different motions of each part. The network is trained with a motion variation loss, enhancing its ability to distinguish even subtly different motions. However, these methods are not directly applicable for distinguishing an individual with ASC as they need very large training datasets to be able to distinguish fine-grained differences in motion, *e.g.*, when an individual is trying to imitate precise movements. A detailed review of related works is provided in Appendix I.

In this work, we propose CAMI-2DNet, a novel deep learning-based method to assess motor imitation for distinguishing individuals with ASC in video data. CAMI-2DNet uses an encoder-decoder architecture to learn a motion representation or encoding that is disentangled from nuisance factors such as skeletal shape and camera views. Such a disentangled representation is crucial for ensuring that the model accurately captures the essence of the movements themselves, without being influenced by irrelevant variations. For example, people with different body shapes may perform the same movement in slightly different ways due to variations in limb length or body shape. If the model does not disentangle these factors, it might incorrectly attribute these differences to the quality of the imitation rather than as natural variations due to the person's physical characteristics. Similarly, variations in camera angles or distances could make identical motions look different. By isolating the motion characteristics from these factors, CAMI-2DNet can consistently evaluate the quality of motor imitation, regardless of an individual's body shape, or whether the video was recorded from a different angle, thereby eliminating the need for labor-intensive tasks like manual frame-by-frame data cleaning and ad hoc normalization, which are needed in methods like CAMI-3D and CAMI-2D.

To effectively learn these disentangled representations, we employ large-scale synthetic data generated by motion retargeting of virtual characters through the reshuffling of motion, body, and camera views, along with participant data from individuals with ASCs and neurotypical individuals. CAMI-2DNet automatically assesses a person's imitation performance by computing a similarity score between motion encodings, which can then be used for autism diagnosis. CAMI-2DNet addresses critical limitations of existing manual methods (HOC) and automated systems (CAMI-3D and CAMI-2D) by providing a quick, reliable, and easy-to-use tool for assessing imitation.

The ability of CAMI-2DNet to precisely quantify motor imitation performance offers a scalable approach for addressing autism heterogeneity in ways that account for a skill, motor imitation, that is fundamental for development of socialization, communication, and other essential skills central to diagnosis of, and targeted intervention for, autism. As such, CAMI-2DNet has the potential to support more frequent, accessible, and detailed assessments, facilitating earlier and more accurate diagnoses, and more personalized treatment planning.

Specifically, the contributions of this paper are as follows:

- **A deep-learning-based approach to motor imitation assessment.** CAMI-2DNet uses an encoder-decoder architecture trained on (a) large-scale synthetic videos generated through motion retargeting and (b) participant videos from individuals with ASCs as well as neurotypical individuals to effectively learn a motion representation that is disentangled from skeletal shape and camera

viewpoint, providing a more robust and objective representation for quantitative assessment of motor imitation.

- **Interpretability through localized scores.** By segmenting the motion representation into different body parts and movement types, CAMI-2DNet offers localized imitation scores, which not only improves the interpretability of the results but also has the potential to enable tailored interventions based on the specific imitation deficits identified in individuals with ASCs.
- **Scalability and practicality.** Unlike CAMI-3D, which relies on specialized 3D cameras, hand-crafted normalization, and HOC annotations for training, and CAMI-2D, which also requires hand-crafted normalization and HOC annotations, CAMI-2DNet operates directly on standard video input and requires neither ad hoc normalization nor HOC annotations. This significantly enhances the scalability and practicality of CAMI-2DNet making it suitable for use in varied settings, including clinics and home environments.
- **Empirical validation.** Our comparative analysis demonstrates that CAMI-2DNet strongly correlates with HOC, matches the performance of CAMI-3D, and outperforms both CAMI-2D and HOC in classifying children into diagnostic groups. These results validate CAMI-2DNet as an effective, practical, and scalable tool for assessing motor imitation in autism diagnosis.

II. Overview of CAMI-2DNet

In this section, we summarize our CAMI-2DNet method for distinguishing individuals with autism in video. Given a video of an actor performing a sequence of movements and a video of a person imitating these movements, the goal is to produce a score that quantifies how closely the person's movements match those of the actor. This imitation score is then used to help discriminate individuals with ASCs from NT individuals.

A. Stages of CAMI-2DNet

An overview of CAMI-2DNet is illustrated in Figure 1a. The method comprises three main stages: estimating body pose, learning a disentangled motion representation, and computing the imitation score. Here is a summary of each stage, with further details provided in the corresponding sections.

- **Estimate 2D Body Pose:** Extract a sequence of 2D body joints (e.g., elbows, knees) from each video, converting visual data to 2D trajectories suitable for motion analysis.
- **Learn a Disentangled Motion Representation:** Map the sequence of 2D body joints to a learnable motion representation that is disentangled from nuisance factors such as variations in body shape or camera viewpoint.
- **Compute Motion Imitation Score:** Use the disentangled motion representation to compute a score that quantifies how well a person imitates the movements of the actor.

These stages allow CAMI-2DNet to yield a more accurate and robust motor imitation assessment by disentangling the dynamics of motion and eliminating distortions caused by irrelevant variables such as body shape and camera viewpoint.

B. Estimating 2D Body Pose

The first step in CAMI-2DNet is to extract the 2D coordinates of the human body joints (*e.g.*, elbows, knees, shoulders) in each video frame, a.k.a. 2D pose estimation. Given a video $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$, where T represents the number of frames and (H, W) denotes the height and width of each frame, a pose estimation model predicts the 2D coordinates of key body joints, which we represent as $\mathbf{J} \in \mathbb{R}^{T\times J\times 2}$, where J is the number of body joints. These joint positions form a time series that tracks the subject's body movements throughout the video, providing suitable data for understanding their motion in subsequent stages of motor imitation assessment.

In this work, we utilize a Vision Transformer-based pose estimation model, EViTPose [19], due to its ability to capture long-range dependencies between body parts while maintaining computational efficiency. This allows the model to generate accurate joint positions even in complex scenarios, such as when the subject is partially occluded or in varying postures. We use EViTPose as a pre-trained, off-the-shelf pose estimator without any additional fine-tuning.

To further enhance the specificity and interpretability of motor imitation assessments, we divide the overall pose sequence $\mathbf{J}$ into S segments, each one corresponding to a specific body part (*e.g.*, arms, legs, torso), and denote the trajectories of the joints in body part segment $\mathcal{S}$ by $\mathbf{J}_{\mathcal{S}}$. This localization into body parts allows for more detailed insights into which specific areas may be contributing to any observed differences in motor imitation. For more details, please refer to Appendix IV-A.

C. Learning Disentangled Representations

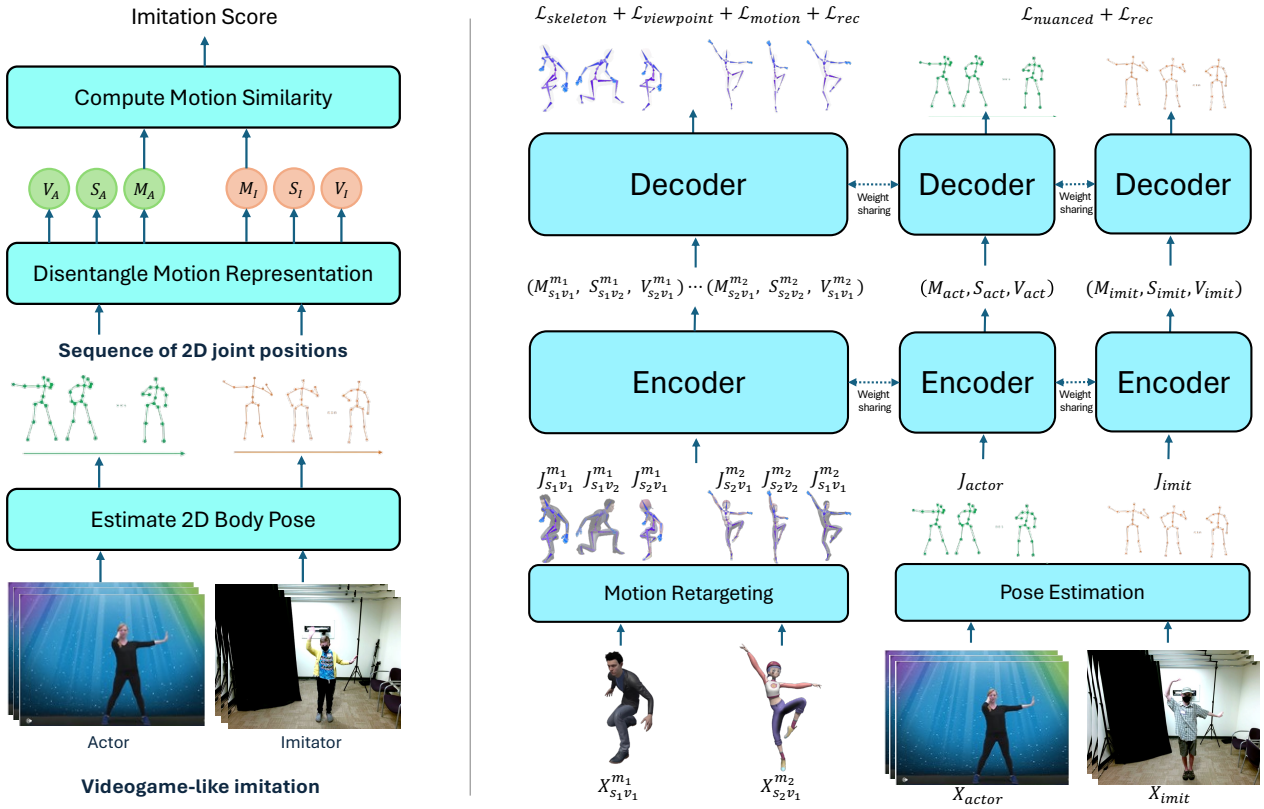
Assessing motor imitation by comparing raw joint trajectories obtained via pose estimation can be misleading because anatomical and viewpoint variations in these trajectories can make it difficult to evaluate accurately how well an individual is imitating a target action. For example, two people performing the same action, such as raising an arm, may appear different due to variations in body posture, like limb length. Similarly, identical movements can look significantly different when captured from a side view versus a front view.

As discussed in the introduction, methods like CAMI-3D [15] attempt to address these variations using hand-crafted techniques, such as adjusting and reorienting joint coordinates to account for changes in body proportions and camera angles, respectively. While effective in controlled settings, these rule-based techniques often fail to generalize to real-world variability. In contrast, CAMI-2DNet learns disentangled representations of motion, shape, and viewpoint, i.e., it learns motion features that are invariant to body shape or viewpoint variations, as detailed in Section III.

D. Computing Motion Similarity

After disentangling motion from skeletal and viewpoint variations, CAMI-2DNet computes an imitation score that quantifies how closely an individual mimics the actor's movements. By comparing the motion representations alone, ignoring irrelevant factors related to body shape and camera viewpoint, CAMI-2DNet provides a robust and objective measure

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2025.3637089



(a) Overview of the CAMI-2DNet Architecture

(b) Overview of the CAMI-2DNet Training Process

Fig. 1: **Overview of CAMI-2DNet.** (a) Given videos of an actor performing a target action and an individual imitating it, CAMI-2DNet extracts 2D joint positions using a pose estimation network and encodes the joint trajectories into disentangled motion, shape, and viewpoint components. The imitation score is computed by calculating the cosine similarity of the motion representations (M_a for the actor and M_i for the individual). (b) During training, the model learns these disentangled representations from synthetic data generated via motion retargeting (varying motion, shape, and viewpoint) and real participant data from neurotypical individuals and individuals with ASCs. The encoder-decoder architecture is optimized using reconstruction and disentanglement losses, ensuring effective encoding and disentanglement of motion, shape, and viewpoint.

of motor imitation performance. This score is then used to distinguish typical imitation from potential impairments, such as those seen in ASCs [25]. The details of how we compute the motion imitation score are discussed in Section IV.

III. Learning Disentangled Representations

As discussed in the previous section, simply relying on raw pose sequences for motor imitation assessment is insufficient due to the entanglement of motion with irrelevant factors like body shape and camera viewpoint. To address these challenges, CAMI-2DNet automatically learns a motion representation from the raw pose sequences that is disentangled from these nuisance factors. In this section, we discuss how CAMI-2DNet achieves this disentanglement. First, we describe the encoder-decoder model architecture that processes pose sequences to produce disentangled motion, shape, and viewpoint encodings. Next, we discuss the role of training data, specifically motion retargeting and the integration of synthetic and participant data. Finally, we outline the training objectives that guide the model to learn robust and disentangled representations. The overall training process is illustrated in Figure 1b, which provides a visual summary of how CAMI-2DNet leverages both synthetic and real data to achieve effective disentanglement of motion, shape, and viewpoint.

A. Model Architecture

To effectively learn a representation disentangled from nuisance factors, we employ an encoder-decoder architecture. The encoder compresses the input pose sequences into latent representations that focus on different action components, while the decoder reconstructs the original pose sequence to validate the quality of the learned representation.

1) Encoding: The encoder is designed to isolate the core characteristics of motion so that the learned representation accurately reflects the subject's motor abilities, free from distortions caused by body shape and camera perspectives. The encoding process is formally defined as:

$$(\mathbf{M}, \mathbf{S}, \mathbf{V}) = f_{\text{enc}}(\mathbf{J}; \theta_{\text{enc}}), \tag{1}$$

where f_{enc} is the encoding network, parameterized by weights θ_{enc}, which transforms the raw pose sequence $\mathbf{J}$ into three

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2025.3637089

disentangled components: (i) $\mathbf{M}$ is a motion representation that captures the essence of the subject's movement, (ii) $\mathbf{S}$ is a shape representation that models the body shape of the subject, which can vary across individuals, and (iii) $\mathbf{V}$ is a viewpoint representation that accounts for the camera perspective from which the movement is captured.

By disentangling these components, the encoder allows CAMI-2DNet to focus solely on the core aspects of the motion $\mathbf{M}$, independent of irrelevant factors like body shape $\mathbf{S}$ or camera viewpoint $\mathbf{V}$. This is crucial for enabling accurate motion comparisons across subjects and environments.

2) Decoding: The decoder plays a vital role in ensuring that the learned latent representation not only captures the essential characteristics of the motion but also retains sufficient information for accurate reconstruction of the original pose sequence. The decoder's objective is to reconstruct the pose sequence $\hat{\mathbf{J}}$ from the disentangled representations, $\mathbf{M}$ (motion), $\mathbf{S}$ (shape), and $\mathbf{V}$ (viewpoint). This reconstruction ensures that the latent space adequately represents all the necessary details to model the original movement accurately. Formally, the decoding process is defined as:

$$\hat{\mathbf{J}} = f_{\text{dec}}(\mathbf{M}, \mathbf{S}, \mathbf{V}; \theta_{\text{dec}}), \tag{2}$$

where f_{dec} is the decoding network, parameterized by weights θ_{dec}, which maps the disentagled components $\mathbf{M}, \mathbf{S}$, and $\mathbf{V}$ to a reconstructed pose sequence $\hat{\mathbf{J}}$ that matches the original pose sequence $\mathbf{J}$ as closely as possible. Please refer to Appendix II-A for further details about the encoder-decoder architecture.

B. Training Data

1) Motion Retargeting: Directly learning disentangled representations from real-world data is inherently challenging due to the absence of explicit information about the underlying motion, body shape, or camera viewpoint. In natural scenarios, the variations in motion are often entangled with differences in body shapes (e.g., height, limb length) and camera perspectives (e.g., angle, distance), making it difficult to isolate the core movement characteristics from irrelevant factors.

As illustrated in Figure 1b, we address this challenge by leveraging *motion retargeting*, which allows us to systematically reshuffle motion, body shape, and camera viewpoint by applying identical movements to different virtual characters with varying body shapes and capturing these motions from multiple viewpoints. For example, as depicted in the bottom left, the same underlying movement is performed by two distinct virtual characters ($X^{m_1}_{s_1v_1}$ and $X^{m_1}_{s_2v_1}$) or viewed from different camera angles ($X^{m_1}_{s_1v_1}$ and $X^{m_1}_{s_1v_2}$). This produces a diverse and rich synthetic dataset, allowing the encoder-decoder to effectively disentangle the core motion dynamics from irrelevant factors such as body shape and viewpoint.

We leverage the Synthetic Actors and Real Actions Dataset [26], a synthetic dataset generated via motion retargeting. In this dataset, virtual characters from a set $\mathcal{B}$ perform motions from a set $\mathcal{M}$, and these motions are observed from different viewpoints in a set $\mathcal{V}$. Each training sample includes pairs of virtual characters $s_1, s_2 \subseteq \mathcal{B}$ performing a triplet of motions $m_1, m_2, m_3 \subseteq \mathcal{M}$, captured from two distinct viewpoints $v_1, v_2 \subseteq \mathcal{V}$. The motions m_1 and m_2 are variations within the same motion class – such as a low jump and a high jump – while m_3 is a distinctly different motion, such as sitting. By reshuffling these components, we ensure that the model learns robust, disentangled motion representations that are invariant to skeletal and viewpoint differences, ultimately leading to more accurate and fair motion comparisons.

2) Integrating Synthetic and Real Participant Data: While synthetic data is essential for learning disentangled representations, it introduces a domain gap: synthetic motions are generic and do not fully reflect the specific types of motions we target in motor imitation assessment for distinguishing ASCs. To address this gap, we employ a balanced, integrated training strategy, where each training batch includes an equal mix of samples from both synthetic and real participant data. The model is optimized jointly using the corresponding training objectives from both data sources. Synthetic data lays the foundation for learning robust motion representations by disentangling motion from skeletal and viewpoint variations, while the participant data refines the model's ability to adapt to the target motion types and captures nuanced differences present in practical settings. This combined training approach ensures that the model benefits from the controlled variability of large-scale synthetic data while simultaneously adapting to the complexity and subtleties of real-world scenarios.

C. Training Objectives

During training, we employ a combination of loss functions tailored to both synthetic and participant data. These loss functions are crucial in guiding the model to effectively separate the core motion from irrelevant factors such as body shape and camera viewpoint, while also capturing the nuanced variations in motor imitation tasks.

For the synthetic data, the model is trained with losses that enforce the disentanglement of motion, shape, and viewpoint, in addition to the reconstruction loss. In contrast, the participant data training utilizes the reconstruction loss alongside a nuanced motion loss, which helps the model capture the subtleties of the motor imitation task.

1) Disentanglement Losses: Here, we describe the disentanglement loss functions, beginning with a common triplet loss formulation that is applied across all action components.

Triplet Loss: The triplet loss is designed to ensure that encodings of the same type (motion, shape, or viewpoint) are closer to each other than encodings of different types. The triplet consists of an anchor, a positive example (similar to the anchor), and a negative example (dissimilar to the anchor). The objective of the triplet loss is to minimize the distance between the anchor and the positive example while maximizing the distance between the anchor and the negative example, encouraging separation between distinct factors. Formally, the triplet loss for an anchor encoding $\mathbf{E}_{\text{A}}$ with its corresponding positive and negative example encodings $\mathbf{E}_{\text{P}}, \mathbf{E}_{\text{N}}$, is defined as:

$$\mathcal{L}_{\text{triplet}}(\mathbf{E}_{\text{A}}, \mathbf{E}_{\text{P}}, \mathbf{E}_{\text{N}}) = \left[\|\mathbf{E}_{\text{A}} - \mathbf{E}_{\text{P}}\|_2^2 - \|\mathbf{E}_{\text{A}} - \mathbf{E}_{\text{N}}\|_2^2 + \alpha\right]_+ \tag{3}$$

where $\mathbf{E}$ can represent motion, shape, or viewpoint representations, $[\cdot]_+$ denotes $\max(0, \cdot)$ and α is a margin parameter

that ensures the distance between $\mathbf{E}_{\text{A}}$ and $\mathbf{E}_{\text{P}}$ is smaller than the distance between $\mathbf{E}_{\text{A}}$ and $\mathbf{E}_{\text{N}}$ by at least the margin α.

Shape Disentanglement Loss: To ensure that the shape encoding $\mathbf{S}$ is invariant to variations in motion and viewpoint but still captures differences in body shape, we apply the triplet loss to shape encodings. In each training sample from the synthetic dataset, virtual characters $s_1, s_2 \subseteq \mathcal{B}$ perform motions $m_1, m_2, m_3 \subseteq \mathcal{M}$ from two distinct viewpoints $v_1, v_2 \subseteq \mathcal{V}$. This allows us to create different combinations for anchor, positive, and negative examples. For an anchor shape encoding $\mathbf{S}_{m_1 s_1 v_1}$, where body s_1 performs motion m_1 from viewpoint v_1, the positive example is $\mathbf{S}_{m_2 s_1 v_2}$ (same body, different motion and viewpoint), and the negative example is $\mathbf{S}_{m_1 s_2 v_1}$ (same motion and viewpoint, different body). The shape disentanglement loss is defined as:

$$\mathcal{L}_{\text{shape}} = \mathcal{L}_{\text{triplet}}(\mathbf{S}_{m_1 s_1 v_1}, \mathbf{S}_{m_2 s_1 v_2}, \mathbf{S}_{m_1 s_2 v_1}). \tag{4}$$

Viewpoint Disentanglement Loss: Similarly, we apply the triplet loss to disentangle the viewpoint representation $\mathbf{V}$ from the motion and shape. In this case, for an anchor $\mathbf{V}_{m_1 s_1 v_1}$ representing the viewpoint encoding of motion m_1 performed by body s_1 from viewpoint v_1, the positive example is $\mathbf{V}_{m_2 s_2 v_1}$ (different motion and body, same viewpoint), and the negative example is $\mathbf{V}_{m_1 s_1 v_2}$ (same motion and body, different viewpoint). The viewpoint disentanglement loss is:

$$\mathcal{L}_{\text{viewpoint}} = \mathcal{L}_{\text{triplet}}(\mathbf{V}_{m_1 s_1 v_1}, \mathbf{V}_{m_2 s_2 v_1}, \mathbf{V}_{m_1 s_1 v_2}). \tag{5}$$

Motion Disentanglement Loss: We apply a set of motion-specific loss functions to disentangle the motion representation $\mathbf{M}$ from the shape representation $\mathbf{S}$ and the viewpoint representation $\mathbf{V}$, while capturing both intra-class variations and subtle differences in motor imitation. We employ two motion disentanglement losses: one for training on the synthetic dataset and another for refining the model's performance on real-world data where the differences in motor imitation are more nuanced. For the synthetic dataset, we extend the triplet loss into a quadruplet loss as in [26], to ensure that the model is sensitive to small variations within the same motion class. This quadruplet loss introduces a semi-positive example, which represents a variation within the same motion class. Given an anchor motion encoding $\mathbf{M}_{m_1 s_1 v_1}$ for motion m_1 performed by body s_1 from viewpoint v_1, a positive example $\mathbf{M}_{m_1 s_2 v_2}$ with the same motion but different body and viewpoint, a semi-positive example $\mathbf{M}_{m_2 s_2 v_2}$ with a variation within the same motion class but different body and viewpoint, and a negative example $\mathbf{M}_{m_3 s_1 v_1}$ with a different motion but same body and viewpoint, the quadruplet loss can formally be defined as:

$$\begin{aligned}\mathcal{L}_{\text{motion}} = \mathcal{L}_{\text{triplet}}(\mathbf{M}_{m_1 s_1 v_1}, \mathbf{M}_{m_1 s_2 v_2}, \mathbf{M}_{m_3 s_1 v_1}) \\ + \beta\{\|\mathbf{M}_{m_1 s_1 v_1} - \mathbf{M}_{m_2 s_2 v_2}\|_2 - \gamma \cdot \text{var}(m_1, m_2)\}.\end{aligned} \tag{6}$$

The first term is a triplet loss that ensures that the anchor motion remains closer to the positive example than to the negative example. The second term controls the sensitivity of the motion encoding to intra-class variations by penalizing the Euclidean distance between anchor and semi-positive example. The third term penalizes a variation score between the characteristics vectors v_{m_1} and v_{m_2} of the anchor and semi-positive example and is defined as:

$$\text{var}(m_1, m_2) = \frac{\|\mathbf{v}_{m_1} - \mathbf{v}_{m_2}\|_1}{2 \times |\mathbf{v}_{m_1}|}. \tag{7}$$

These characteristics vectors contain variables such as energy, distance, and height, which influence the shape movement and are provided as metadata in the dataset. Finally, β and γ are scaling factors that adjust the impact of each term in the loss.

Nuanced Motion Loss: For the participant data, where motor imitation assessments require distinguishing even more subtle differences between a target and an imitated motion, we introduce a nuanced motion loss to capture these fine distinctions. This loss penalizes the distance between the motion encodings, using the DTW distance between the corresponding pose sequences as a dynamic margin that guides how "close" or "far apart" these motion encodings should be. Given a pair of pose sequences – $\mathbf{J}_{\text{actor}}$, representing the actor's movements and, $\mathbf{J}_{\text{imit}}$, representing a person's imitated movements – and their corresponding motion encodings, $\mathbf{M}_{\text{actor}}$ and $\mathbf{M}_{\text{imit}}$, respectively, the loss is defined as:

$$\mathcal{L}_{\text{nuanced}} = \|\mathbf{M}_{\text{actor}} - \mathbf{M}_{\text{imit}}\|_2^2 + \delta \cdot \text{dist}(\text{DTW}(\mathbf{J}_{\text{actor}}, \mathbf{J}_{\text{imit}})), \tag{8}$$

where the function $\text{dist}(\text{DTW}(\mathbf{J}_{\text{actor}}, \mathbf{J}_{\text{imit}}))$ calculates the distance between the pose sequences after alignment with Dynamic Time Warping (DTW). The Euclidean distance between the motion encodings $\|\mathbf{M}_{\text{actor}} - \mathbf{M}_{\text{imit}}\|_2^2$ is influenced by this DTW distance, which acts as a margin, and the parameter δ is a scaling factor that modulates its impact. If the DTW distance between the pose sequences is small (indicating strong alignment in the movements), the encodings are encouraged to be closer together than the ones with a larger DTW distance (indicating less alignment).

2) Reconstruction Loss: The reconstruction loss ensures that the latent representations contain sufficient information to accurately reconstruct the original pose sequence. This loss helps maintain the integrity of the learned representation while simultaneously validating the completeness of the disentangled components: motion ($\mathbf{M}$), shape ($\mathbf{S}$), and viewpoint ($\mathbf{V}$). The reconstruction loss is computed as:

$$\mathcal{L}_{rec}(\mathbf{J}, \hat{\mathbf{J}}) = \frac{1}{T}\frac{1}{J}\sum_{t=1}^{T}\sum_{j=1}^{J} \|\mathbf{J}_t^j - \hat{\mathbf{J}}_t^j\|_2^2, \tag{9}$$

where $\mathbf{J}_t^j$ and $\hat{\mathbf{J}}_t^j$ represent the 2D coordinates of joint j at time t in the original and reconstructed sequences, respectively.

3) Total Loss: The total loss integrates the disentanglement, reconstruction, and nuanced motion losses. For the synthetic data, the focus is on disentangling motion, shape, and viewpoint while ensuring that the model can reconstruct the original pose sequences. The total loss for synthetic data is given by:

$$\mathcal{L}_{\text{total-syn}} = \lambda_{\text{dis}}(\mathcal{L}_{\text{shape}} + \mathcal{L}_{\text{viewpoint}} + \mathcal{L}_{\text{motion}}) + \lambda_{\text{rec}}\mathcal{L}_{rec}, \tag{10}$$

where λ_{dis} and λ_{rec} are weighting factors that control the contribution of each component.

For the participant data, we apply a different combination of losses to ensure that the model can reconstruct real sequences

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2025.3637089

and capture the nuanced differences in motor imitation. The total loss for the participant data is given by:

$$\mathcal{L}_{\text{total-real}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{nuanced}}\mathcal{L}_{\text{nuanced}}, \tag{11}$$

where the weights λ_{rec} and λ_{nuanced} balance the two losses.

During training, the model learns from both synthetic and participant data in a mixed process. The overall total loss is a weighted sum of the synthetic and real data losses:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{syn}}\mathcal{L}_{\text{total-syn}} + \lambda_{\text{real}}\mathcal{L}_{\text{total-real}}, \tag{12}$$

where λ_{syn} and λ_{real} are weighting factors that balance the contributions of the synthetic and participant data during training. By combining these losses, the model benefits from the strengths of both datasets, leveraging synthetic data for disentanglement and participant data for adapting to the nuanced complexities of real-world motor imitation.

IV. Computing Motion Imitation Score

Having learned robust and disentangled representations of motion, shape, and viewpoint during training, the goal of CAMI-2DNet is to compute an imitation score that quantifies the similarity between the actor's motion and the person's imitated motion. As discussed before, this score is critical for evaluating motor imitation performance for diagnosing ASCs.

To compute this score, CAMI-2DNet focuses solely on the motion encodings, eliminating the influence of nuisance factors such as differences in body shape or camera viewpoint. This allows for a more accurate and fair comparison of the movements between the actor and a person. In this section, we detail the process CAMI-2DNet uses to compute the imitation score, beginning with the encoding of the pose sequences, followed by refining and aligning the motion encodings, and concluding with the calculation of cosine similarity between the actor's and the person's motion encodings.

1) Encoding: Given two pose sequences, $\mathbf{J}_{\text{actor}}$ representing the actor's movements and $\mathbf{J}_{\text{imit}}$ representing the person's imitated movements, we first encode these sequences using the trained encoder f_{enc}, which generates three disentangled components for both the actor and the person: motion encoding $\mathbf{M}$, shape encoding $\mathbf{S}$, and viewpoint encoding $\mathbf{V}$:

$$(\mathbf{M}_{\text{actor}}, \mathbf{S}_{\text{actor}}, \mathbf{V}_{\text{actor}}) = f_{\text{enc}}(\mathbf{J}_{\text{actor}}; \theta_{\text{enc}}), \tag{13}$$

$$(\mathbf{M}_{\text{imit}}, \mathbf{S}_{\text{imit}}, \mathbf{V}_{\text{imit}}) = f_{\text{enc}}(\mathbf{J}_{\text{imit}}; \theta_{\text{enc}}). \tag{14}$$

2) Optimizing Motion Encodings: Once the original pose sequences $(\mathbf{J}_{\text{actor}}, \mathbf{J}_{\text{imit}})$ have been encoded, we refine their motion encodings $(\mathbf{M}_{\text{actor}}, \mathbf{M}_{\text{imit}})$ to improve the reconstruction of the original pose sequences. The refinement is carried out by minimizing the reconstruction loss $\mathcal{L}_{\text{rec}}$, which measures the difference between the original pose sequences and their reconstructed versions $\hat{\mathbf{J}}_{\text{actor}} = f_{\text{dec}}(\mathbf{M}_{\text{actor}}, \mathbf{S}_{\text{actor}}, \mathbf{V}_{\text{actor}}; \theta_{\text{dec}})$ and $\hat{\mathbf{J}}_{\text{imit}} = f_{\text{dec}}(\mathbf{M}_{\text{imit}}, \mathbf{S}_{\text{imit}}, \mathbf{V}_{\text{imit}}; \theta_{\text{dec}})$, while keeping shape $(\mathbf{S}_{\text{actor}}, \mathbf{S}_{\text{imit}})$ and viewpoint $(\mathbf{V}_{\text{actor}}, \mathbf{V}_{\text{imit}})$ encodings and the decoder f_{rec} freezed. The minimization objective is given by:

$$\min_{\mathbf{M}_{\text{actor}}} \mathcal{L}_{\text{rec}}(\mathbf{J}_{\text{actor}}, f_{\text{dec}}(\mathbf{M}_{\text{actor}}, \mathbf{S}_{\text{actor}}, \mathbf{V}_{\text{actor}}; \theta_{\text{dec}})), \tag{15}$$

$$\min_{\mathbf{M}_{\text{imit}}} \mathcal{L}_{\text{rec}}(\mathbf{J}_{\text{imit}}, f_{\text{dec}}(\mathbf{M}_{\text{imit}}, \mathbf{S}_{\text{imit}}, \mathbf{V}_{\text{imit}}; \theta_{\text{dec}})). \tag{16}$$

3) Computing the Imitation Score: After optimizing the motion encodings, we ignore the shape and viewpoint encodings and focus solely on comparing the motion representations $\mathbf{M}_{\text{actor}}$ and $\mathbf{M}_{\text{imit}}$. Before computing the similarity between the two motion encodings, we first temporally align the encodings using DTW to account for any differences in timing or duration between the actor's and the person's motions. Following the alignment, we compute the similarity between the motion encodings using cosine similarity, which provides a quantitative measure of how closely the encoded representations of the two motions align. The cosine similarity is given by:

$$\text{score}(\mathbf{M}_{\text{actor}}, \mathbf{M}_{\text{imit}}) = \frac{\mathbf{M}_{\text{actor}} \cdot \mathbf{M}_{\text{imit}}}{\|\mathbf{M}_{\text{actor}}\|_2 \|\mathbf{M}_{\text{imit}}\|_2}. \tag{17}$$

To enhance the interpretability of the imitation assessment, the final score is computed as a weighted average of the cosine similarity of motion encodings of the actor and an individual for the $\mathcal{S}$ body segments as discussed in Section II-B. The final imitation score is computed as:

$$\text{CAMI}(\mathbf{M}_{\text{actor}}, \mathbf{M}_{\text{imit}}) = \sum_{\mathcal{S} \subset \mathcal{J}} w_{\mathcal{S}} \cdot \left[\text{score}(\mathbf{M}^{\mathcal{S}}_{\text{actor}}, \mathbf{M}^{\mathcal{S}}_{\text{imit}})\right]_+, \tag{18}$$

where $w_{\mathcal{S}}$ represents the weight assigned to the body segment $\mathcal{S}$, which is a subset of $\mathcal{J}$, the full set of body joint indices. Each body segment $\mathcal{S}$ corresponds to a specific set of joint indices (*e.g.*, joints for the left arm or right leg). The weights are determined via cross-validation grid-search hyperparameter tuning. By isolating body segments, CAMI-2DNet localizes the assessment to specific areas of the body, providing insight into which body part contributes to the imitation differences. For visualization examples of these localized assessments, please refer to Appendix IV-A. The final CAMI score ranges between 0 and 1, and quantifies how well the person imitates the actor's movement, with higher values indicating greater similarity and better imitation performance. This score, after normalization using the minimum and maximum values, is then used to discriminate people with ASCs from NT.

V. Experiments

In this section, we provide an overview of the synthetic and participant datasets, including details about the participants and experimental procedures. Moreover, we present experiments comparing the effectiveness of our deep-learning-based method, CAMI-2DNet, relative to the non-deep-learning methods, CAMI-3D, CAMI-2D, and Human Observation Coding (HOC). The evaluation focuses on construct validity, reliability, and diagnostic classification performance.

A. Dataset details

1) Synthetic Actors and Real Actions Dataset: We employed the synthetic motion dataset SARA [26], created using Adobe Mixamo [27], to gather sequences of poses from a variety of 3D characters, each with a unique body shape, performing the same motions under kinematic constraints. This dataset comprises motion sequences from 18 different 3D characters across four action categories: Combat, Adventure, Sport, and

Dance. Each action sequence comprises a minimum of 32 frames, with a total of 4,428 base motions (*e.g.*, dancing, jumping), having noticeable intra-class variations, resulting in a total of 103,143 variations. Each frame in these sequences contains the 3D coordinates of 17 joints from various body parts, and samples were generated through 2D projection.

2) Participant Dataset: In addition, we incorporated participant data from neurotypical (NT) individuals and individuals with Autism Spectrum Conditions (ASCs) to train and evaluate our method. These participant data were collected as part of a wider-scale study examining imitation skills in autism.

a) Participants: The participant dataset included 185 people aged 6 to 12 years, comprising 82 children with ASCs and 103 neurotypical (NT) children. We refer to this dataset as CAMI-185. Among these participants, 47 participants (27 with ASCs, 20 NT) have HOC score annotations, forming a subset we refer to as CAMI-47. See Appendix III-A for details on the participant demographies and socioeconomic status. As detailed in [15], and consistent with other prior methods for assessing motor imitation [28], the HOC procedure was designed to derive semi-quantitative scores of motor imitation performance, with summed ratings of individual movements, each scored as inaccurate (0) or accurate (1). Two expert raters, trained by a pediatric neurologist (SHM), scored each participant, with inter-rater and intra-rater reliability established.

The CAMI-47 subset was used to evaluate the performance of the methods against HOC and to train the CAMI-3D method.[1] The autism diagnoses were based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [29] criteria as applied by a board-certified Child Neurologist (SHM) with over 30 years of clinical and research experience with children with ASC. Research-reliable assessors confirmed the diagnosis on-site using the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [30]. The parent-report version of the Social Responsiveness Scale, Second Edition (SRS-2) [31] was also administered. To participate, children needed a full-scale IQ score of at least 80 or at least one index score of 80 (verbal comprehension, visual-spatial, or fluid reasoning index) on the Wechsler Intelligence Scale for Children-Fifth Edition [32]. Additionally, to account for autism-associated differences in general motor abilities, we used the Movement Assessment Battery for Children (mABC), Second Edition [33]. Ethics approval was obtained from the Johns Hopkins University School of Medicine Institutional Review Board before the study began. Written informed consent was obtained from all participants' legal guardians, and verbal assent was obtained from all children. Recruitment was conducted through local schools and community events. Participants were invited to the Center for Neurodevelopmental and Imaging Research at the Kennedy Krieger Institute for two-day visits and received $100 compensation for their time.

b) Procedure: Children participated in an imitation task involving two movement sequences, (Sequence 1, Sequence 2), each repeated across two trials (Trial A and Trial B). The two sequences included different types of movements (Sequence 1: 14 movements, Sequence 2: 18 movements) that were relatively unfamiliar to the participants (e.g., moving arms up and down like a puppeteer), lacked an end goal, and required the simultaneous movement of multiple limbs. See Appendix III-B for details on the movement types. The movements were designed to assess imitation of a continuous series of fluid, dynamic movements involving simultaneous articulation of multiple joints. This design provided for assessment of autism-associated difficulties with imitation, in particular challenges with movements requiring dynamic visual-motor integration [5], [34]–[38]. The stimulus video was displayed on a large TV screen, showing an actor performing dance-like whole-body movements without any background music or sound. The children's movements were recorded using two Kinect Xbox cameras at 30 frames per second, one positioned in front of the child and the other at the back. 3D data was used for CAMI-3D and video data from the front camera was used for CAMI-2D and CAMI-2DNet analysis. In the CAMI-185 dataset, 182 participants completed Trial A of both sequences, while fewer participants completed Trial B (Sequence 1: 54, Sequence 2: 61). For the CAMI-47 subset, participants in Trial A were 43 for Sequence 1 and 46 for Sequence 2, while in Trial B they were 40 for Sequence 1 and 36 for Sequence 2.

B. Results

1) Construct Validity and Test Re-test Reliability: To verify the construct validity of our method relative to CAMI-3D and CAMI-2D, we analyzed their correlation with the scores obtained from HOC across all sequences and trials of the CAMI-47 dataset. The results, as illustrated in Figure 2a, show strong positive correlations between the three methods and HOC. Specifically, the correlation coefficients were 0.69 for CAMI-3D, 0.66 for CAMI-2D, and 0.70 for CAMI-2DNet. Notably, CAMI-2DNet, which operates entirely without supervision from HOC, demonstrated the highest correlation with HOC scores. This strong correlation not only highlights the accuracy and reliability of CAMI-2DNet but also underscores its potential as a highly effective tool for analyzing the participant data independently of HOC supervision. Despite strong correlation, indicating similar coarse-level assessment, CAMI-2DNet and HOC differ in diagnostic discrimination: HOC's subjective binary scoring can overlook subtle variations in imitation, whereas CAMI-2DNet's fine-grained features capture nuanced patterns, yielding superior diagnostic discrimination (see next).

2) Diagnostic Classification Ability: We evaluated the performance of our method, CAMI-2DNet, relative to CAMI-3D, CAMI-2D, and HOC in classifying children into diagnostic groups by computing the receiver-operating characteristic (ROC) curve across all sequences and trials of the CAMI-47 dataset. Larger areas under the curve (AUC) indicate better discriminative ability, as shown in Figure 2b. CAMI-3D demonstrated the highest performance with an AUC of 0.859, indicating its superior capability in distinguishing between diagnostic groups. The 3D nature of this method likely contributes to its better performance. The CAMI-2D method, which operates on 2D video data, showed an AUC of 0.789. While its performance is slightly lower than that of CAMI-3D, it provides valuable discriminative ability, comparable

[1]CAMI-2D does not require training the linear regression as it uses the weights learned by CAMI-3D for regressing the CAMI scores.

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2025.3637089

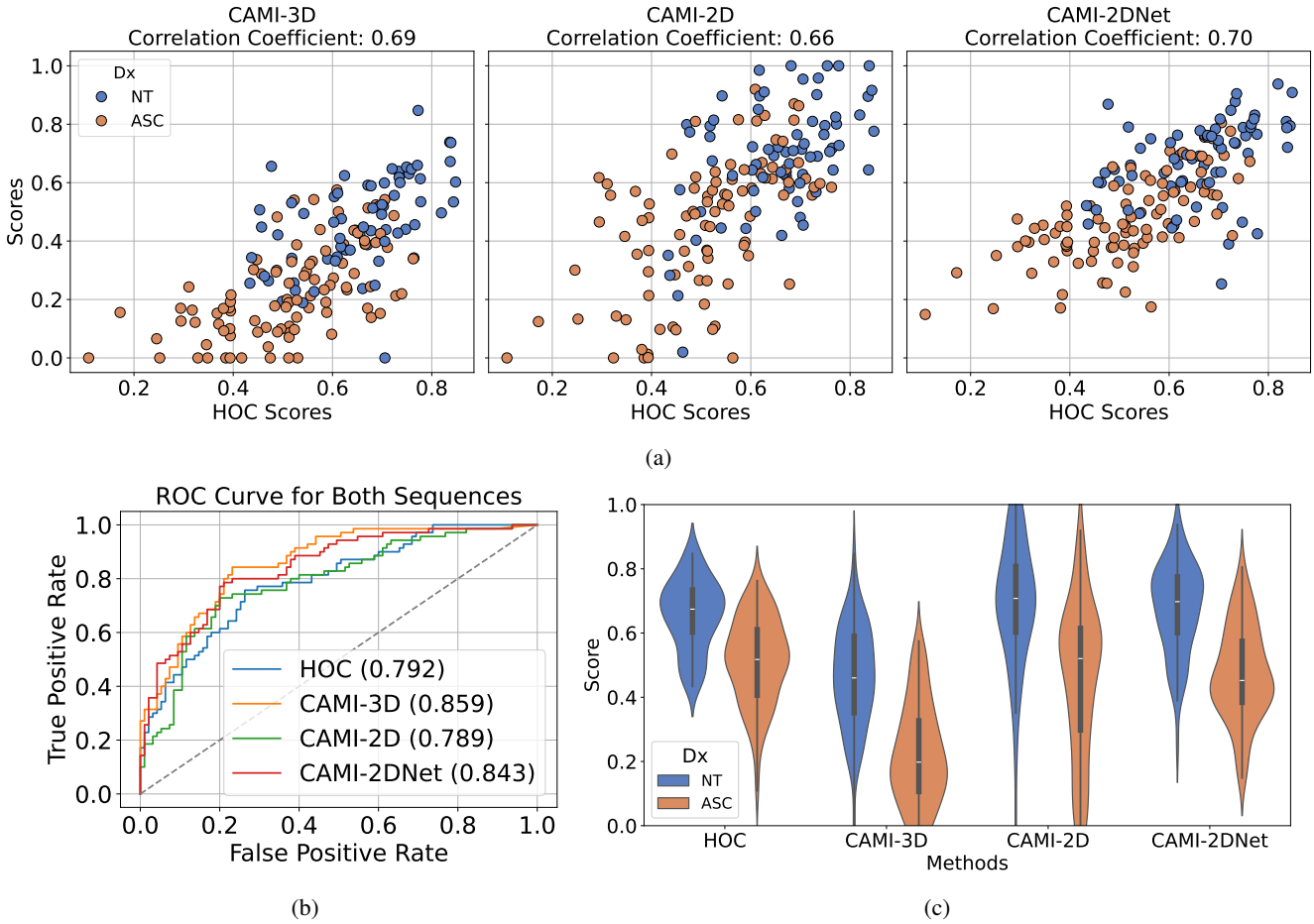


Fig. 2: Comparing CAMI-2DNet, CAMI-2D, CAMI-3D, and human observation coding (HOC) on the CAMI-47 dataset (27 ASCs, 20 NT). **(a) Correlation with HOC Scores**: Scatter plots showing the correlation between HOC scores and the scores from CAMI-3D, CAMI-2D, and CAMI-2DNet. CAMI-2DNet has the highest correlation with HOC scores. **(b) ROC Curve for Both Sequences**: Receiver operating characteristic (ROC) curve: true positive rate vs. false positive rate as classification threshold is varied. The Area Under the Curve (AUC) indicates the diagnostic ability of the different methods. CAMI-2DNet (AUC = 0.843) demonstrates comparable performance to CAMI-3D (AUC = 0.859) and superior performance over both HOC (AUC = 0.792) and CAMI-2D (AUC = 0.789). **(c) Violin Plot of Scores**: The violin plots illustrate the distribution of scores for ASC and NT groups across the four methods. CAMI-2DNet not only shows a clear separation between the ASC groups but also displays less variability within each group, highlighting its robustness and reliability.

to the HOC method (AUC = 0.792). The reliance on 2D data without leveraging the depth information of 3D data might account for this difference. CAMI-2DNet achieved an AUC of 0.843, demonstrating comparable performance to CAMI-3D and superior performance over both CAMI-2D and HOC. CAMI-2DNet's ability to operate directly on video data without requiring HOC annotations during training offers a significant practical advantage. This makes CAMI-2DNet a highly effective and efficient tool for diagnostic classification, balancing high performance with operational simplicity. Furthermore, the violin plots, in Figure 2c, showing the distribution of scores for the NT and ASC groups across the four methods, demonstrate that CAMI-2DNet not only exhibits a distinct separation between ASCs and NT groups but also shows relatively reduced variability within each group, underscoring its robustness and reliability.

The ROC curve for both sequences, consisting of two trials in CAMI-47, is shown in Figure 3 (a-d). For instance, in both trials of Sequence 1, CAMI-2DNet achieved the highest AUC (Trial A: 0.885, Trial B: 0.887), outperforming CAMI-3D (Trial A: 0.794, Trial B: 0.84), CAMI-2D (Trial A: 0.785, Trial B: 0.839, and HOC (Trial A: 0.785, Trial B: 0.854). For Sequence 2 - Trial A, CAMI-2DNet achieved an AUC of 0.825, which is higher than CAMI-2D (0.802)[2] and HOC (0.762), but lower than CAMI-3D (0.906). In Sequence 2 - Trial B, CAMI-2DNet and CAMI-2D both achieved an AUC of 0.863, while CAMI-3D scored the highest at 0.906 and HOC obtained 0.834. These trials demonstrate the consistent and high performance of CAMI-2DNet across different se-

[2]The results for CAMI-2D reported in this paper differ from those reported in the original CAMI-2D paper [20] due to two main reasons: i) the current evaluation included 46 participants, compared to the 40 participants in the original study, and ii) the analysis in this paper employs cross-validation, whereas the original paper did not use cross-validation in its evaluation. These factors contribute to variations in the performance outcomes observed.

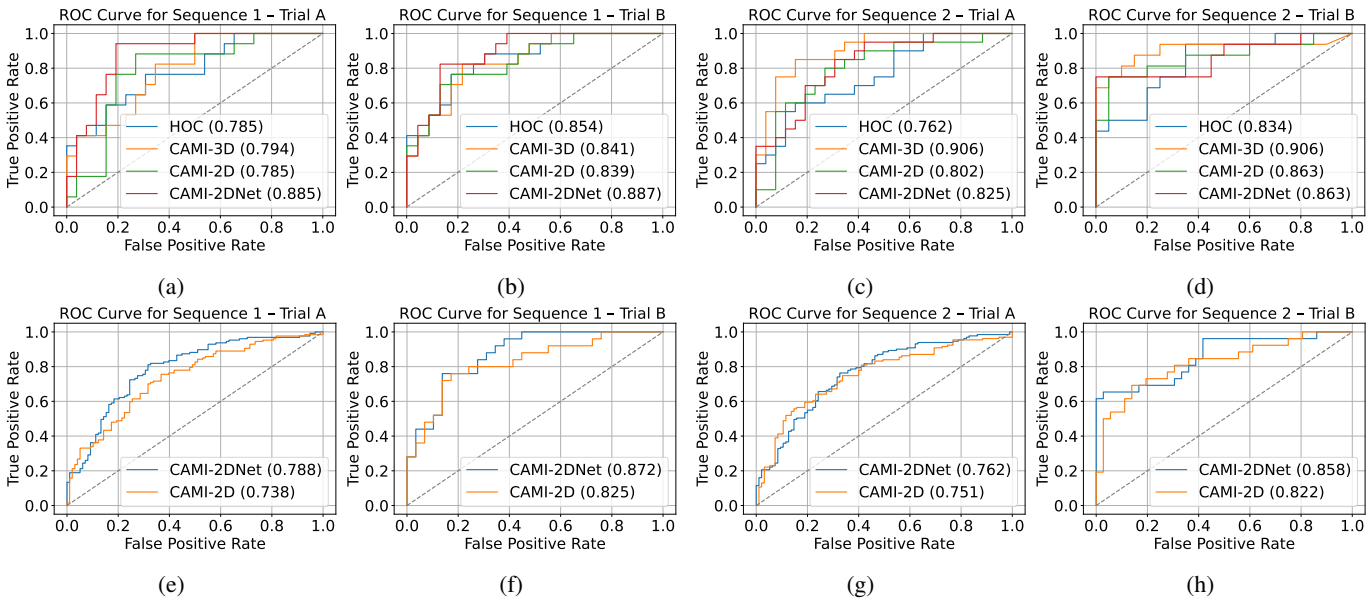


Fig. 3: **Receiver Operating Characteristic (ROC) curves comparing the diagnostic performance of HOC, CAMI-3D, CAMI-2D, and CAMI-2DNet across two datasets: CAMI-47 and CAMI-185.** The top row (a-d) presents results on the CAMI-47 dataset for two sequences, each consisting of two trials. CAMI-2DNet consistently outperforms HOC and CAMI-2D and demonstrates comparable or superior performance to CAMI-3D. The bottom row (e-h) shows results on the CAMI-185 dataset, comparing CAMI-2DNet with CAMI-2D across two sequences and two trials. CAMI-2DNet achieves higher diagnostic accuracy in all trials, demonstrating a higher AUC than CAMI-2D.

quences and trials, reinforcing its capability as a practical and effective tool for diagnostic classification.

The ROC curves across both sequences and trials in the CAMI-185 dataset further highlight CAMI-2DNet's consistent superiority compared to CAMI-2D. As shown in Figure 3 (e-h), CAMI-2DNet achieves higher AUC scores in both trials of Sequence 1 (Trial A: 0.787 vs. 0.737, Trial B: 0.853 vs. 0.824) and Sequence 2 (Trial A: 0.767 vs. 0.751, Trial B: 0.856 vs. 0.824). Overall, CAMI-2DNet outperforms CAMI-2D and HOC in terms of diagnostic ability and maintains a strong correlation with HOC scores. Moreover, CAMI-2DNet performs comparably to CAMI-3D while offering greater practicality by operating directly on video data and without needing labor-intensive HOC annotations and ad hoc normalization steps.

VI. Limitations and Future Directions

While CAMI-2DNet shows strong promise as a practical and scalable tool for motor imitation assessment, it has some limitations that open avenues for future research. First, learning disentangled motion representations is based on synthetic data, which may not fully capture the variability and complexity of real-world human movements. To bridge this gap, we employed integrated training with real participant data, but future work should explore other ways, such as domain adaptation. Second, deploying CAMI-2DNet in real-world clinical or mobile settings introduces challenges related to computational efficiency and hardware limitations, and comprehensive clinical assessment. Optimizing the model for lightweight inference is a critical next step toward broader accessibility and scalability. It should also be noted that motor imitation difficulties captured by CAMI-2DNet represent only one aspect of autism. Thus, CAMI-2DNet can be a useful tool that complements, but does not necessarily replace, other diagnostic and treatment assessments that tap into autism diagnostic criteria, such as social-communicative difficulties and restricted interests and repetitive behaviors. Finally, motor imitation behaviors could be influenced by cultural and environmental factors. Expanding training and evaluation datasets to include diverse populations would help improve the model's robustness and fairness across different demographic and cultural contexts.

VII. Conclusion

We introduced CAMI-2DNet, a scalable and interpretable deep learning-based approach to motor imitation assessment in video data. CAMI-2DNet uses 2D pose estimation techniques to extract 2D joint trajectories from the video. These trajectories are then mapped to a motion representation that is disentangled from nuisance factors such as body shape and camera viewpoint. A motor imitation score is then computed by comparing the motion representation of an individual to that of the actor. Our experiments demonstrate that CAMI-2DNet performs on par with CAMI-3D in discriminating ASC vs neurotypical children, and outperforms both HOC and CAMI-2D, while offering greater practicality by operating directly on video data and without the need for ad hoc data normalization and HOC annotations. These results highlight CAMI-2DNet as an effective and accessible tool for assessing motor imitation in children with ASCs and related developmental conditions.

REFERENCES

[1] A. Masi, M. M. DeMayo, N. Glozier, and A. J. Guastella, "An overview of autism spectrum disorder, heterogeneity and treatment options," *Neurosci. Bull.*, vol. 33, pp. 183–193, 2017.

[2] A. N. Bhat, "Motor impairment increases in children with autism spectrum disorder as a function of social communication, cognitive and functional impairment, repetitive behavior severity, and comorbid diagnoses: A SPARK study report," *Autism Res.*, vol. 14, pp. 202–219, 2021.

[3] H. Over and M. Carpenter, "The social side of imitation," *Child Dev. Perspect.*, vol. 7, no. 1, pp. 6–11, 2013.

[4] R. Santra, C. Pacheco, D. Crocetti, R. Vidal, S. H. Mostofsky, and B. Tunçgenç, "Evaluating Computerised Assessment of Motor Imitation (CAMI) for identifying autism-specific difficulties not observed for attention-deficit hyperactivity disorder or neurotypical development," *Br. J. Psychiatry*, pp. 1–8, 2025.

[5] D. E. Lidstone and S. H. Mostofsky, "Moving toward understanding autism: Visual-motor integration, imitation, and social skill development," *Pediatr. Neurol.*, vol. 122, pp. 98–105, 2021.

[6] M. B. Nebel, A. Eloyan, C. A. Nettles, K. L. Sweeney, K. Ament, R. E. Ward, A. S. Choe, A. D. Barber, J. J. Pekar, and S. H. Mostofsky, "Intrinsic visual-motor synchrony correlates with social deficits in autism," *Biol. Psychiatry*, vol. 79, no. 8, pp. 633–641, 2016.

[7] L. A. Edwards, "A meta-analysis of imitation abilities in individuals with autism spectrum disorders," *Autism Res.*, vol. 7, pp. 363–380, 2014.

[8] F. Chiarotti and A. Venerosi, "Epidemiology of autism spectrum disorders: A review of worldwide prevalence estimates since 2014," *Brain Sci.*, vol. 10, no. 5, p. 274, 2020.

[9] K. Lyall, L. A. Croen, J. L. Daniels, M. Fallin, C. Ladd-Acosta, B. K. Lee, B. Y. Park, N. W. Snyder, D. E. Schendel, H. E. Volk, G. C. Windham, and C. Newschaffer, "The changing epidemiology of autism spectrum disorders," *Annu. Rev. Public Health*, vol. 38, pp. 81–102, 2017.

[10] S. Michelet, K. Karp, E. Delaherche, C. Achard, and M. Chetouani, "Automatic imitation assessment in interaction," in *Int. Workshop on Human Behavior Understanding*, 2012.

[11] R. C. Schmidt, S. Morr, P. Fitzpatrick, and M. J. Richardson, "Measuring the dynamics of interactional synchrony," *J. Nonverbal Behav.*, vol. 36, pp. 263–279, 2012.

[12] A. Paxton and R. Dale, "Frame-differencing methods for measuring bodily synchrony in conversation," *Behav. Res. Methods*, vol. 45, pp. 329–343, 2013.

[13] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *CVPR*, 2009.

[14] A. Ravichandran, R. Chaudhry, and R. Vidal, "View-invariant dynamic texture recognition using a bag of dynamical systems," in *CVPR*, 2009.

[15] B. Tunçgenç, C. Pacheco, R. Rochowiak, R. Nicholas, S. Rengarajan, E. Zou, B. Messenger, R. Vidal, and S. H. Mostofsky, "Computerised Assessment of Motor Imitation (CAMI) as a scalable method for distinguishing children with autism," *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging*, vol. 6, no. 3, 2021.

[16] R. Bellman and R. Kalaba, "On adaptive control processes," *IRE Trans. Autom. Control*, vol. 4, no. 2, pp. 1–9, 1959.

[17] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Real-time multi-person 2d pose estimation using part affinity fields," *IEEE TPAMI*, vol. 43, no. 1, pp. 172–186, 2019.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *CVPR*, 2019.

[19] K. A. Kinfu and R. Vidal, "Efficient vision transformer for human pose estimation via patch selection," in *BMVC*, 2023.

[20] D. E. Lidstone, R. Rochowiak, C. Pacheco, B. Tunçgenç, R. Vidal, and S. H. Mostofsky, "Automated and scalable Computerized Assessment of Motor Imitation (CAMI) in children with autism spectrum disorder using a single 2d camera: A pilot study," *Res. Autism Spectr. Disord.*, vol. 87, p. 101840, 2021.

[21] D. Holden, T. Komura, and J. Saito, "Phase-functioned neural networks for character control," *ACM TOG*, vol. 36, no. 4, pp. 1–13, 2017.

[22] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM TOG*, vol. 35, pp. 1–11, 2016.

[23] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or, "Learning character-agnostic motion for motion retargeting in 2d," *arXiv preprint arXiv:1905.01680*, 2019.

[24] H. Coskun, D. J. Tan, S. Conjeti, N. Navab, and F. Tombari, "Human motion analysis with deep metric learning," in *ECCV*, 2018.

[25] J. H. G. Williams, A. Whiten, and T. Singh, "A systematic review of action imitation in autistic spectrum disorder," *J. Autism Dev. Disord.*, vol. 34, pp. 285–299, 2004.

[26] J. Park, S. Cho, D. Kim, O. Bailo, H. Park, S. Hong, and J. Park, "A body part embedding model with datasets for measuring 2d human motion similarity," *IEEE Access*, vol. 9, pp. 36 547–36 558, 2021.

[27] Adobe Inc, "Mixamo." [Online]. Available: https://www.mixamo.com

[28] S. H. Mostofsky, P. Dubey, V. K. Jerath, E. M. Jansiewicz, M. C. Goldberg, and M. B. Denckla, "Developmental dyspraxia is not limited to imitation in children with autism spectrum disorders," *J. Int. Neuropsychol. Soc.*, vol. 12, no. 3, pp. 314–326, 2006.

[29] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. American Psychiatric Association, 2022.

[30] A. McCrimmon and K. Rostad, "Test review: Autism diagnostic observation schedule (ADOS-2) manual: Toddler module," *J. Psychoeduc. Assess.*, vol. 32, no. 1, pp. 88–92, 2014.

[31] T. P. Bruni, "Test review: Social responsiveness scale, 2nd ed. (SRS-2)," *J. Psychoeduc. Assess.*, vol. 32, no. 4, pp. 365–369, 2014.

[32] L. G. Weiss, V. N. Locke, T. Pan, J. G. Harris, D. H. Saklofske, and A. Prifitera, "Wechsler intelligence scale for children," *WISC-V*, 2019.

[33] S. E. Henderson, D. A. Sugden, and A. L. Barnett, *Movement Assessment Battery for Children, 2nd ed. (MABC-2)*. Pearson Assessment, 2007.

[34] E. Gowen, "Imitation in autism: Why action kinematics matter," *Front. Integr. Neurosci.*, vol. 6, p. 117, 2012.

[35] L. K. MacNeil and S. H. Mostofsky, "Specificity of dyspraxia in children with autism," *Neuropsychology*, vol. 26, no. 2, p. 165, 2012.

[36] D. McAuliffe, A. S. Pillai, A. Tiedemann, S. H. Mostofsky, and J. B. Ewen, "Dyspraxia in asd: Impaired coordination of movement elements," *Autism Res.*, vol. 10, no. 4, pp. 648–652, 2017.

[37] R. P. Hobson and J. A. Hobson, "Dissociable aspects of imitation: A study in autism," *J. Exp. Child Psychol.*, vol. 101, pp. 170–185, 2008.

[38] L. Marsh, A. Pearson, D. Ropar, and A. Hamilton, "Children with autism do not overimitate," *Curr. Biol.*, vol. 23, no. 7, pp. R266–R268, 2013.