Loose Social-Interaction Recognition in Real-world Therapy Scenarios

Abid Ali[1 2] Rui Dai[1] Ashish Marisetty[1] Guillaume Astruc[1]
Monique Thonnat[1] Jean-Marc Odobez[3] Susanne Thümmler[1 2] Francois Bremond[1 2]

[1]INRIA [2]University Cote d'Azur [3]Idiap

Abstract

The computer vision community has explored dyadic interactions for atomic actions such as pushing, carrying-object, etc. However, with the advancement in deep learning models, there is a need to explore more complex dyadic situations such as loose interactions. These are interactions where two people perform certain atomic activities to complete a global action irrespective of temporal synchronisation and physical engagement, like cooking-together for example. Analysing these types of dyadic-interactions has several useful applications in the medical domain for social-skills development and mental health diagnosis.

To achieve this, we propose a novel dual-path architecture to capture the loose interaction between two individuals. Our model learns global abstract features from each stream via a CNNs backbone and fuses them using a new Global-Layer-Attention module based on a cross-attention strategy. We evaluate our model on real-world autism diagnoses such as our Loose-Interaction dataset, and the publicly available Autism dataset for loose interactions. Our network achieves baseline results on the Loose-Interaction and SOTA results on the Autism datasets. Moreover, we study different social interactions by experimenting on a publicly available dataset i.e. NTU-RGB+D (interactive classes from both NTU-60 and NTU-120). We have found that different interactions require different network designs. We also compare a slightly different version of our method (details in Section 3.6) by incorporating time information to address tight interactions achieving SOTA results.

1. Introduction

Human activity recognition has been an active research area in the computer vision community for a wide range of applications, including health care, video surveillance, personality development, sports analytics, robotics, and so on. In this domain, the analysis of human-human interaction and, more specifically, of dyadic interaction plays a central role. Dyadic interaction recognition has useful applications in the medical domain for social skills development, parent-child interaction therapy (PCIT), autism spectrum disorder (ASD) diagnosis, mental health diagnosis, education, etc.

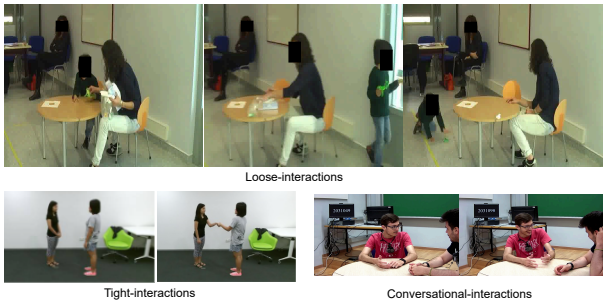

Figure 1. Different dyadic interaction types.

Dyadic interactions can be categorised into three types i) *tight interactions*, ii) *conversational interactions*, and iii) *loose interactions* as illustrated in Figure 1. Tight interactions are synchronised atomic actions with physical contact involved, such as shaking hands, hugging, etc. Moreover, tight interactive activities are composed of a few seconds with limited intra-class variance and high temporal synchronisation. They have been thoroughly studied in computer vision research. In the past decade, several high-performing Deep Convolutional Neural Networks (CNNs) models [17, 30, 38] have been designed to classify tight interactive actions with more than 90% accuracy on lab-simulated public datasets such as SBU [37], ShakeFive [32], NTU-RGB+D [20]. Similarly, conversational interactions, such as engaging with one or more people during a meeting, a debate, or a talk, have been investigated in the literature for several purposes, including personality modelling [1, 4, 6, 25]. A characteristic of such conversation-based interactions is that people have little mobility, and are primarily shot from front-facing cameras, and the main analysis goal is to detect and model the streams of conversational activities they exhibit like talking, eye-gaze aversion or contacts, facial expressions, and minimal hand gestures.

In contrast, **loose interactions** are complex dyadic interactions, where two people individually perform a combination of asynchronous and asymmetric atomic actions that complete the global task without direct physical involvement. Loose dyadic interactions are long actions (more than one minute long) representing complex real-world scenarios such as **compound social interactions** (individuals performing different atomic actions within an activity), **spontaneous acting** (subjects acting freely without any specific guidance), **asynchronous and asymmetrical events** (both people performing independent and different atomic actions of their own, but together they complete an interactive activity such as celebrating a birthday), and **without physical engagement**. This type of activity usually consists of a leader and an assistant or helper. For example, in the activity of cooking, there is a leader (chef) who performs the main activity (cooking) by interacting in a loose manner with an assistant or helper (for instance, the assistant helps in chopping vegetables, providing required ingredients, etc.). The combination of such asynchronous and asymmetrical atomic actions generally has a global interaction for cooking activity. This type of weak interaction has not been explored much in the computer vision community to date. Therefore, there is a lack of research on recognising these complex activities that have loose asynchronous human-human interactions.

Additionally, in autism diagnosis, each ADOS [22] module corresponds to different tasks (actions of Loose-Interaction dataset) for severity evaluation. For example, the activity of *imitation* responds to an analysis of child attention, gaze, and social skills. Classifying these loose interactions helps us to use each module for its appropriate task.

Existing deep learning models such as I3D [5], X3D [10], and SlowFast [11] etc. perform flawlessly on such tight interactive activities as they are atomic and temporally synchronised (kissing or hugging each other). On the contrary, as discussed above, loose interactions are complex, having no such temporal synchronisation, neither are they symmetrical, and therefore are challenging for existing methods. Such interactions require a model that can exchange abstract-level information between the two individuals at different levels. This limits the capabilities of current models to be applied to such activities.

Furthermore, existing two-stream models, with early-fusion struggle to handle asynchronous and asymmetrical interactions (needs well-defined temporal synchronisation like the ones we see in tight interactions to perform well), while late-fusion models do not exchange sufficient information between the two streams. Therefore, mid-level global feature modelling is necessary to recognise loose interactions.

Taking into account the above challenges, we propose a new architecture for loose-interaction recognition in social activities. The main contributions of this paper are as follows.

- To our knowledge, we are the first to propose a new task of collaborative loose interactions, to focus on the recognition of asynchronous and asymmetrical loose social interactions in dyadic situations.

- We propose a novel dual-path network for joint action recognition (composite social-interactive activities). The dual paths learn high-low-level multiscale visual features individually from two distinct inputs (leader and assistant) using **3D-CNNs**. The global abstract features are obtained through **Abstract Projection**. The action is recognised by performing a fusion via a novel **Global Layers Attention** (GLA) mechanism.

- We validate our method on a real-world dataset, depicting the loose social interactions of a clinician with a child during an autism diagnosis. Autism diagnosis data are recorded during the assessment of young children with ASD following the ADOS-2 protocols [22]. Besides this real-world dataset, we perform experiments on other datasets e.g., Autism, and NTU-RGB+D.

2. Related Work

Video Classification: CNNs have been very successful in learning 3D spatio-temporal representations for human activity recognition [5]. Two-stream methods commonly used in combination of RGB and optical flow [12], with a special emphasis on video classification. SlowFast network [11] has demonstrated the possibility of combining representations of different temporal resolutions (i.e. frame rates) to improve action recognition. Recently, with the advent of Transformers, several methods improved action recognition by incorporating attention. MViT [9] proposed pooling attention to learn spatio-temporal features at different scale. Video-Swin [21] improved MViT using 3D shifted window modules for self-attention with patch merging after each spatial downsampling.

Furthermore, Foundation models such as CLIP [29], DINOV2 [24], and VideoMAE [31] has been very useful for down-stream tasks such as action recognition, and action localisation [27]. However, these methods are focused on general action recognition with little or no attention towards human-human interactions. Therefore, we adapt 3DCNN backbone to model pyramid of spatio-temporal features at higher spatial resolution (early layers) to low-level visual information (deeper layers) from two individual RGB inputs. On top of that we utilise cross-attention mechanism

of transformers to build our interaction recognition architecture.

Human Interaction Recognition is a subdomain of recognition of actions. Lately, researchers combine CNN, RNN and GCN to recognise interactions from skeleton data [28, 36]. For instance, DR-GCN [38] learns geometric and relative attention features from the two skeletons using their dyadic relational graph module. [14] uses mid-fusion of 3-stream GCNs using inter and intra-body graphs to recognise interactions between two skeletons. Dyadformer [6] proposed cross-subject layers using audio video inputs from two individuals to predict the personality of both individuals in long videos. However, most of these methods are focused on dyadic short interactions with proper temporal synchronisation. A good skeleton input can effectively recognise short actions such as *shaking hands* but could not model more complex loose interaction.

ASD Recognition: The Autism Diagnostic Observation Schedule (ADOS) [22], a standard semi-structured test, was created by psychologists to identify ASD. The aim of ADOS, which can last up to two hours (four 30-minute sessions) and requires expert skills to carry out, is to assess the degree of social insufficiency in children. They designed individual modules to evaluate gaze, face gestures, body gestures and social-interactions of child during each session. In recent years, researchers have developed computer vision algorithms to address ASD behaviour recognition. Action recognition systems can use articulated posture structures, appearance, and motion information to study behavioural cues to address ASD diagnosis [2, 8, 23, 26]. [2] uses two-stream I3D in a late-fusion manner to recognise autistic actions. Recently, [26] proposed a guided weakly supervised method. They augment target autistic action classes with a general video dataset using posterior maximum likelihood for better behavioural posture learning. Unfortunately, these techniques focus primarily on repetitive or stemming behaviours of the child, ignoring the social complexity that occurs in interaction circumstances, which is a crucial component of ASD diagnosis.

3. Proposed Method

In this section, we discuss our proposed architecture for recognising complex social activities that have asynchronous loose interactions. Our network addresses the key challenge: How to effectively model interactive activities that are asymmetrical and have no temporal synchronisation at the frame-level? To handle this situation, we came up with a dual path architecture. The network consists of four parts: i) **Convolution Backbone**, ii) **Abstract Projection Module**, iii) **Global Layers-Attention Module**, and iv) **Classification Head**, as shown in Figure 2. This is an end-to-end learning architecture. Each module is explained in the following.

3.1. Terminology description

Prior to modelling loose asynchronous interactions, we need to establish the roles of each individual involved in the interaction. Asynchronous loose interactions usually have a leader (focused on completing the whole action/task) and an assistant or helper (helping with minimal atomic actions). These roles could be reversed if the child is autistic (less socially active). Therefore, our input terms are defined as "Leader" and "Assistant" in the architecture.

3.2. Convolution Backbone

Our backbone learns the spatio-temporal multi-scale features of two distinct individual inputs in a dual-stream fashion. Both paths shares style design and parameters of convolution backbone reducing computation costs. We adapt a multi-scale 3D-CNNs having a pyramid of features with early layers operating at high spatio-temopral resolution modelling low-level visual features, and deeper layers at spatio-temporally coarse, but complex high-dimensional information. Let X be the video snippet; then $X_{leader}^{b\times c\times t\times h\times w}$ and $X_{assistant}^{b\times c\times t\times h\times w}$ are the cropped images of the leader and assistant to extract $g^{b\times c\times t\times h\times w}$ coarse-fine features from the convolution backbone at different levels ($block_3$, $block_4$, and $block_5$).

For each input X_{leader}, $X_{assistant}$ we extract the g_{leader}, and $g_{assistant}$, respectively. As both individuals perform these actions interactively (asynchronously and asymmetrically), the idea is to extract different symbolic spatio-temopral features in different blocks for fusion. As, the two individuals interact in a loose manner, fusing them locally (spatio-temporal level) does not benefit. Therefore, we need to compute global abstract information of each block for fusion, as shown in Figure 2(b). Experimentally, we find that utilising features from the last three blocks can efficiently recognise a complex activity (details in Section 5).

3.3. Abstract Projection

Before concatenating, coarse-fine block-level features needs to be encoded to a common embeddings. First, we average-pool the spatio-temporal dimensions of each block to obtain a global context. As these loose actions are executed jointly, both individuals perform separate tasks to complete the entire activity regardless of time synchronisation at the frame-level: for instance, in the **Joint-game**[1] activity, the clinician (leader) assembles different toys such as a dollhouse and plays with the child (assistant), where the child sets up a toy-lounge-furniture that involves *pick and place small chair, and desk, etc.* and the clinician *pick and move around a small doll*. Therefore, extracting a global context of the spatio-temporal features can better recognise

[1]The activity is a part of ADOS assessment

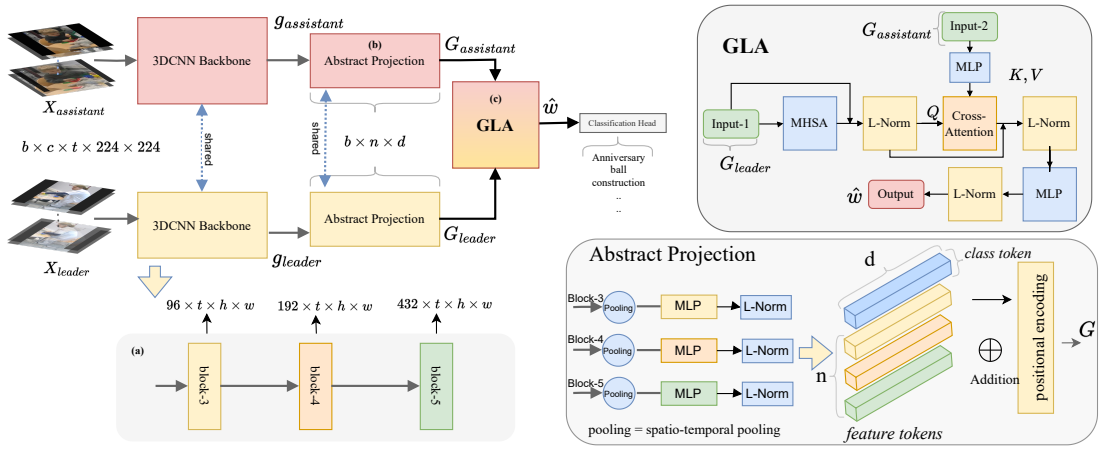

Figure 2. Our proposed architecture consists of (a) the Convolution backbone, (b) the Abstract Projection module, and (c) the GLA module. The model takes the input (leader and assistant) and outputs the action prediction score through the classification head.

these types of action. Second, we project these block-level features through MLP layer to learn abstract features of higher order at coarse-fine (blocks) level for each input as in Eq. 1. We use GELU as activation function.

$$L = MLP(activation(AvgPool3D(X))) \quad (1)$$

The three projected layers are fused, obtaining $G^{b\times n\times d}$, where n represents the number of layers, which is three in our case, and d is the embedded feature vector of size 768. At this point, we concatenate a learnable classification token clfToken$^{b\times 1\times 768}$ in dimension n along with the addition of a 1-D positional encoding PE$^{1\times 4\times d}$ as shown in Eq. 2. The same process is performed for g_{leader} and $g_{assistant}$ individually.

$$G^{b\times(n+1)\times d} = LNorm(fuse(L^{b\times d} * n, clf) + PE) \quad (2)$$

where G is the projected embedding of the blocks features for each stream, LNorm, L, n, clf, and PE represent layer normalisation [3], embedded encoding, the number of layers, class-token, Positional Encoding, respectively. The abstract projection G is executed separately for both paths g_{leader}, and $gassistant$, obtaining an embedding G_{leader}, and $G_{assistant}$ (as illustrated in Figure 2(b)).

3.4. Global-Layer-Attention

So far, the dual-paths independently learn abstract context and encode them through the abstract projection module as embeddings. Now, we need to integrate these embeddings from G_{leader}, and $G_{assistant}$ in such a way that we capture the asynchronous loose interaction between the two participants. To this end, we rely on an attention mechanism.

$$Q = LNorm(G_{leader} + MHSA(G_{leader})) \quad (3)$$

$$K = V = MLP(G_{assistant})) \quad (4)$$

In particular, let Q be the query from G_{leader} (in Eq. 3) and the key and value $\{K, V\}$ (defined in Eq. 4) from $G_{assistant}$ to perform a cross-attention in between the two inputs as in Eq. 5. MHSA [33] stands for multi-head self-attention, MLP indicates multilayer perceptron.

Our intuition is that the global high-level information (key-value pairs) of one person can be used to attend to the global context of the other person for better loose interaction recognition (as illustrated in Figure 2(c)) in an asynchronous and asymmetrical manner. We implement this intuition with the help of the proposed *Global-Layer-Attention*. This module takes an input G_{leader} and the corresponding vector from the interacted person $G_{assistant}$ and uses the multi-head cross attention mechanism to obtain the refined vector $\hat{w}^{b\times n+2\times d}$ (as in Eq. 5), where b, n, 2, d defines the batch-size, the number of layers, classification tokens, and the embedded dimensions (as illustrated in Figure 2)(c). T stands for transpose in Eq. 5.

$$\hat{w} = softmax(\frac{Q * K^T}{\sqrt{d_k}}) * V \quad (5)$$

In our implementations, we use eight tokens, including two classification tokens and three abstract feature layers of each stream (obtained in Section 3.3) to perform cross attention between G_{leader}, and $G_{assistant}$. First, we apply a single MHSA layer with eight attention heads in G_{leader} followed by layer-normalisation to get Q. Before performing a cross-attention between G_{leader} and $G_{assistant}$ we obtained $\{K, V\}$ by applying a single MLP layer on $G_{assistant}$. Finally, a cross-attention is performed between Q and $\{K, V\}$. We use layer normalisation and add skip-connections at certain positions, as shown in Figure 2(c).

3.5. Classification Head

The output is taken by applying an MLP layer to the classification token of $\hat{Y}^{b\times 2\times D}$ taken from the final output of

the Global-Layer-Attention module, where b is the batch size, 2 denotes the two classification tokens (one for each stream) and D is the 768 embeddings.

3.6. Temporal Synchronisation Modelling

To best accommodate temporally synchronised tight interactive actions, we made slight changes to our existing architecture. This slightly different design of our model captures frame-level temporal information in a synchronised and symmetrical manner. The main difference is, how we project the CNN layers, with specific changes in the Abstract Projection module. This variant of our model uses only the last CNN-block $block_5$ by pooling only spatial features, preserving the temporal information for both streams. The feature token is generated from $G^{b\times t\times d}$, where b is the batch-size, t represents the temporal frames (32 for NTU) and d are the embedded features as in Eq. 6. The rest of the model is the same.

$$G = MLP(Relu(AvgPool2D(block\_5\_layer))) \quad (6)$$

4. Experiments

We have studied two types of interaction, i) loose interactions, and ii) tight interactions, and explored which modelling strategy (global abstract context, or temporal modelling) is effective for them.

In Section 4.4 we experiment with the Loose-Interaction and the Autism [26] datasets to study interactions in social therapy situations. We have found that global abstract features from the CNNs backbone can effectively address asymmetrical and temporally asynchronous interactions.

On top of that, we study the NTU-RGB+D [20] dataset for tight interactions. Our study concludes that tight interactive actions are temporally synchronised and that a global abstract feature approach is not helpful. Tight interactions can be addressed with a slight change in the proposed architecture to model temporal information, as explained in Section 3.6.

4.1. Loose-Interaction Dataset

The Loose-Interaction dataset is actual children's assessment sessions recorded with clinicians at the hospital. 132-hour sessions were recorded following the ADOS-2 protocol to study the visual behaviour of children with the severity of autism. Each child was diagnosed with a possible autism disorder during different interactive ADOS-2 activities. Long videos were classified into nine (9) interaction classes i.e., *anniversary, playing with bubbles, playing with ball, construction, demonstration, describing-image, imitation, joint-game, and puzzle*. Each action video is 2 - 4 minutes long, depending on the activity. Blurred, distorted, and out-of-frame videos were discarded, yielding a total of 845 trimmed videos, with 9 classes, captured with an HD camera at 30 fps. Some subjects did not perform the same activity; therefore, the dataset is not subject-oriented and highly imbalanced. In this paper, a total of 87 unique children's hour-long videos were used out of 132 videos. The statistics of the dataset are given in Table 1

Action	**# of clips**	**# of unique children**
Anniversary	118	63
playing with bubbles	110	63
playing with ball	67	45
construction	253	38
demonstration	45	27
describing image	43	36
imitation	121	57
joint game	41	26
puzzle	47	30
Total	845	87(overlap exists)

Table 1. Loose-Interaction dataset statistics.

We intend to release the dataset in multiple modalities after ethical approval.

4.2. Public Datasets

Currently, there is no publicly available dataset for asynchronous loose interactions that fits our needs. Therefore, we evaluated our model on other closely related public datasets such as **Autism** [26] and **NTU-RGB+D** [20].

The **Autism** dataset was designed for the behavioural study of children with ASD under stress. It is more focused on the child and their autistic (repetitive) actions. It has some sort of loose interactions (*in most videos, the clinician performs the activity with the child in a weak interactive manner*), but not asynchronous. Furthermore, the actions are fine-grained, short. There are 1333 clips with a total of 8 action classes. More description provided in the supplementary materials.

In the **NTU-RGB+D** [20] dataset we consider only the interactive actions from both NTU-RGB+D 60 and 120 datasets achieving a total of 26 action classes in 13k video clips. Although actions fall into the category of tight-interactive activities, they could be useful for the validation of our model and for studying different methods that work for tight-interaction recognition.

4.3. Experimental Details

Complex loose interactions videos are temporally long and require higher temporal modelling (more than 60 frames) to capture the action completely. We have found that global abstract features from the CNNs backbone can effectively address such long complex (asymmetrical and temporally asynchronous) interactions. Utilising 3D-CNNs as our backbone has 2 main benefits. First, 3D-CNNs such

Methods	Input	Acc.% Mean
2S-DRAGCN [38]	2P skeleton	33.84
GWSDR$_{rgb+flow}$ [26]	scene	52.33
Mvit [9]		56.37
X3D [10]		63.50
DinoV2+TCN [24]		63.98
VideoMAE$_{finetuned}$ [31]		64.50
SlowFast [11]	2P tracklets	20.06
X3D$_{earlyfusion}$ [10]		30.03
GWSDR$_{rgb}$ [26]		41.43
CoarseFine [16]		46.06
Mvit$_{latefusion}$ [9]		58.01
X3D$_{latefusion}$ [10]		64.01
DinoV2+TCN$_{latefusion}$ [24]		65.86
VideoMAE$_{fine-tuned}$ [31]		66.71
Proposed †	2P tracklets	37.03
Proposed		**72.04**

Table 2. Baseline results on the Loose-Interaction dataset. † defines our temporal model, 2P: means cropped tracklets of both persons, separately.

Methods	Input	Modality	Acc. % (CS)
ST-LSTM [28]	2P	Skeleton	63.00
IRN$_{inter+intra}$ [28]			77.70
GCA-LSTM [28]			73.00
ST-GCN [38]			80.20
AS-GCN [38]			73.13
2S-AGCN$_{uni_joint}$ [38]			83.19
2S-AGCN$_{uni-bone}$ [38]			85.25
2S-DRAGCN [38]			90.56
PoseC3D$_{limb}$ [7]	scene		94.91
PoseC3D$_{joint}$ [7]			95.85
I3D [5]	scene	RGB	82.00
Swin-Transformer [21]			92.52
SlowFast [11]	2P		93.70
Proposed			95.02
Proposed †			**96.25**

Table 4. Results in the NTU-RGB+D dataset for interactive actions and comparison with the SOTA methods. $P1$ and $P2$ represent the results of using a single-person tracklet without interactions. †: means our temporal variant model. 1P (single), 2P (both) tracklets.

as X3D [10] have longer temporal modelling capabilities due to multiscale temporal pooling compared to Transformers (Mvit, VideoSwin) [9, 33]. X3D can operate at 64 - 120 frames input with a low computational cost compared to VideoSwin. Second, training 3D-CNNs with a smaller dataset is comparatively better compared to Transformers without requiring additional training strategies.

All networks were pre-trained on the kinetics-400 [5] dataset. We kept the same training protocols for all experiments with a batch size of 8 and trained them for 100 epochs. For the proposed architecture, we have used the SGD optimiser with an initial learning rate of 0.003 and a momentum of 0.9 at training time.

We use pre-processing explained in the supplementary materials to extract tracklets (individual bounding boxes) for the Loose-Interaction dataset. We used skeleton information for the extraction of tracklets in NTU-RGB+D, and Yolov5 [15] with DeepSORT [35] for the Autism dataset. Additional details are provided in the Supplementary Materials.

Method	Acc.%
ECO [39]	61.4
TSN [34]	68.0
R(2+1)D [13]	69.8
I3D [5]	69.3
TSM [19]	69.8
TSN+DR$_{rgb+flow}$ [26]	70.1
TSN+GWS+DR$_{rgb+flow}$ [26]	72.5
GWSDR$_{rgb+flow}$ [26]	75.1
Proposed	**76.3**
Proposed$_{rgb+flow}$	**78.6**

Table 3. Results and comparison with SOTA on the Autism dataset.

4.4. Experiments on ASD datasets

Experiments on Loose-Interactions: we first evaluate our proposed dual-path architecture on the Loose-Interaction dataset, the results are reported in Table 2. Handling the long temporal duration of the Loose-Interaction dataset, we perform several experiments with different temporal sizes, including 32, 64, 80, and 120 frames. The best results were achieved with the 80-frame snippet. All baseline models follow the same protocols. For a fair comparison, an additional TCN layer is used for longer temporal modelling in transformer-based baselines such as Mvit [9] and VideoMAE [31].

We compare our proposed method with existing 3D-CNNs, GCN, and Transformer-based architectures.

GWSDR [26] uses an additional optical-flow modality to capture motion. Interestingly, this method did not perform well on Loose-Interaction dataset. We notice, their model greatly benefits from a weak co-learning strategy by training on other similar atomic actions found in large datasets (Kinetics). However, the loose-Interaction dataset has unique and composite actions, different from the Kinetics action classes. Thus, pretraining in such a manner does not fully help the model to converge. Secondly, the actions are longer, asymmetrical, and temporally asynchronous for this method to capture well. 2S-DRAGCN [38] uses skeletons for dyadic-interactions. However, the model is designed to better capture synchronised atomic interactions

Input	Acc. %
MHSA($G_{assistant}$), MLP(G_{leader})	71.3
MLP($G_{assistant}$), MHSA(G_{leader})	72.0

Table 5. Impact of swapping the inputs of the GLA module.

(as in the case of NTU-RGB+D) compared to complex actions in Loose-Interaction dataset.

We further investigate SOTA transformers-based action classification architectures for loose-interaction tasks. We experiment with MViT [18] (multiscale design), VideoMAE [31] (general mask learning model) and DINOV2 [24] (foundation model). We use MViT-small with an additional TCN layer for longer temporal pooling. However, the model did not converge due to the size of the dataset. Next, we use fine-tuning strategies by freezing a few layers at a time to let the model learn and converge for each input separately. Later, we use a latefusion strategy as shown in Table 2. We experiment with the same strategy for VideoMAE [31]. A common problem with these methods is the small size of the dataset and higher computational costs. Furthermore, they works well on atomic actions compared to long complex activities. Also, the architectures perform poorly if trained separately, reducing computational costs, but could not converge fully if both inputs are jointly trained.

Lastly, we use a multilayer TCN model for temporal modelling of spatial features extracted from DinoV2 [24] and MLP layers for classification. DINOV2+TCN performs well with only temporal modelling but is unable to capture more complex actions due to no spatial interaction learning.

Experiments on Autism: in this section, we perform experiments on the Autism dataset where we use 32 frames as input snippets. Our proposed network achieves new SOTA results (Table 3) in the Autism dataset with an increase in accuracy of 1.2%. The existing methods in Table 3 use full-frame in their experiments. GWSDR [26] utilises an additional optical flow stream in conjunction with the rgb scene to capture atomic actions in this dataset. Furthermore, they greatly benefited from other large-scale datasets using their guided weak supervision technique. However, our proposed method achieves higher accuracy by just using rgb tracklets of the two individuals. To make a fair comparison with them, we further use an additional optical flow stream for this experiment in latefusion manner, achieving an additional 2.3% increase in accuracy.

4.5. Experiments on tight interactions

On the NTU-RGB+D dataset, we only used interactive action classes. For this reason, we compare our method with the SOTA methods in interactive action classes, as shown in Table 4. Our main network achieves results comparable to those of the SOTA. We further extend our network to model temporal synchronisation. Our slightly changed network achieves SOTA results in the NTU-RGB+D dataset as reported in Table 4. Our original proposed method is also comparable to that of SOTA. The validation of these two designs on the loose and tight interactive datasets shows the usability of different network design strategies for different synchronisations and symmetry.

Network streams	**Acc.%**
leader	55.62
assistant	51.84
leader+assistant$_{LF}$	64.02
Proposed	**72.04**

Table 6. Importance of each stream in our design choice. LF: stands for late-fusion

Layer Selection	**Acc.%**
$conv_5$	60.0
All ($block_1 - block_5$)	48.0
$block_2, block_3, block_4, block_5$	66.0
$block_3, block_4, block_5$	**72.0**

Table 7. Experimenting with the number of layers to model abstract features.

Attention	Acc.%
Self-attention	38.5
Cross-Attention	72.0

GLA modules	Acc.%
2	44.2
1	72.0

Table 8. Ablation study on attention mechanism and analysing the number of GLA modules used.

Components	**Acc.%**
W/o AP, W/o GLA	64.0
W AP, W/o GLA	66.5
W AP, W GLA	**72.0**

Table 9. Experiments with and without the Abstract Projection (AP), and GLA module. **W** and **W/o** means **with** and **without**, respectively.

5. Ablation Study

Here, we discuss several ablation experiments that validate our design choices, using the Loose-Interaction dataset as it is the most relevant one. More specifically, we analyse i) the influence of the network inputs swapping; ii) the CNNs embedding layers; iii) the attention type; iv) the number of GLA modules used for attention; v) the importance of each component used.

To evaluate the impact of leader and assistant on model performance, we swapped the input of the GLA module. Table 5 validates that there is an asymmetric and asynchronous behaviour, where the leader is interested in com-

pleting the activity by having a loose interaction with the assistant. The small drop in accuracy we observe is due to the fact that when the child is autistic, they are not fully involved in the activity, as shown in Table 6. Thus, the information he/she carries is smaller compared to that of their partner. This validates our proposed way of processing the leader and assistant streams as in Figure 2(c).

Furthermore, to understand the usefulness of each input, we experiment with training the $X3D$ [10] model separately for both the leader and the assistant. The results in Table 6 show that the leader stream is more accurate in recognising activity than the assistant. Combining both streams using our proposed GLA module can greatly improve the prediction of such loose interactions. In addition, we provide more ablation study about the number of attention heads used in GLA module and depth of backbone in the Supplementary Materials Section 4.1.

Device	Flops (G)	Param (M)	Infer. Time Data (ms)	Infer. Time Model (ms)
CPU	-	-	1760	3830
GPU	76.99	10.54	1560	80.90

Table 10. (Computational complexity analysis. We use a 10 clips testing strategy. We show computational complexity in GFLOPs for a single clip input and inference time of the model and data in milliseconds. Provided inference time is on both CPU (Xeon Silver 4215) and Tesla v100 GPU for a single batch size. Infer. is short for Inference)

Next, we analyse the pyramid of high- and low-level features extracted from the 3D-CNN backbone for further projection. Specifically, features of $block_3$, $block_4$, and $block_5$ of the 3D-CNNs backbone. We compare this design choice with other experimental approaches from using only the $block_5$ block to utilising all blocks, $block_1 - block_5$, from both streams. The results of these experiments are given in Table 7. This analysis demonstrates that earlier low-level features are not as important compared to deep-layer features.

Moreover, the proposed method is based on cross-attention strategy to perform attention using the novel GLA module. A cross-attention between two different input streams can emphasise the correlation between them efficiently. We compared this design choice with self-attention to validate its importance. Table 8 defines this comparison between different attention approaches. Next, our network has only one GLA module for this fusion between the two paths using attention. However, we have noticed a drop in the efficiency of our network when increasing the complexity of the model (increasing the number of GLA modules). One possible reason for this is the small size of the dataset to fully use a more dense architecture. This analysis is reported in Table 8. Additionally, we have evaluated the usefulness of the Abstract Projection module and the GLA module by experimenting without each of them. Results are described in Table 9.

5.1. Computational Complexity Analysis

We evaluate our model's computational complexity using GLOPs and inference time for a single input to validate its use for devices with limited resources in real-time, as shown in Table 10. Our model is light-weight as the backbone and abstract projection modules are shared between the two paths, thus our model uses only 10.54M parameters having 76.99 GFLOPs. Furthermore, it takes only 3.8 seconds to run a single input video on the CPU (Xeon Silver 4215) and 1.76 seconds to process a single batch input. Similarly, with a single Tesla v100 GPU our model takes 0.80 seconds to run a single input video. This validates our model is efficient in terms of computational complexity and can work on resource-constrained devices in real-time.

6. Conclusion and Future Work

Recognising complex social-interaction between two individuals performing an action is a challenging task. We propose a new direction for human-human action recognition having loose interactions. To address this challenging task, we design a new architecture that attends to temporally unsynchronised loose-interactive actions using our novel Global-Layer-Attention module. We validate our network in social therapy scenarios for Loose-Interactions and Autism datasets. Our model achieves SOTA results on these two datasets. To demonstrate our network generalisability in tight-interactive actions, we experiment with the NTU-RGB+D dataset. Our model achieves higher results by slightly changing the model design to capture synchronised interactions. Our proposed method has certain limitations to handle all types of social interactions. One possible solution would be to design an adaptive temporal synchronisation module that can model symmetrical and asymmetrical time synchronisation between two people.

In the future, our next goal is to detect the activities of autistic children in untrimmed videos. With the help of action detection and recognition methods, the final goal will be to generate severity reports for autistic children by automating the autism diagnosis process.

Acknowledgments

COFUND BoostUrCAreer program received funding from the European Union's Horizon 2020 under Marie Curie grant agreement No 847581. This work is also supported by the French government, through the ACTIVIS project managed by the National Research Agency (ANR) with the reference number ANR-19-CE19-0004.

References

[1] Tanay Agrawal, Dhruv Agarwal, Michal Balazia, Neelabh Sinha, and Francois Bremond. Multimodal personality recognition using cross-attention transformer and behaviour encoding. *arXiv preprint arXiv:2112.12180*, 2021. 1

[2] Abid Ali, Farhood F Negin, Francois F Bremond, and Susanne Thümmler. Video-based Behavior Understanding of Children for Objective Diagnosis of Autism. In *VISAPP 2022 - 17th International Conference on Computer Vision Theory and Applications*, Online, France, Feb. 2022. 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 70–79, 2022. 1

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 6

[6] David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B Moeslund, et al. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2177–2188, 2021. 1, 3

[7] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 6

[8] Rajagopalan et al. Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using rgbpose-slowfast network. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3356–3360, 2022. 3

[9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 2, 6

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 2, 6, 8

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 6

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016. 2

[13] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019. 6

[14] Yoshiki Ito, Quan Kong, Kenichi Morita, and Tomoaki Yoshinaga. Efficient and accurate skeleton-based two-person interaction recognition using inter-and intra-body graphs. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 231–235. IEEE, 2022. 3

[15] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, (Zeng Yifu), Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022. 6

[16] Kumara Kahatapitiya and Michael S. Ryoo. Coarse-fine networks for temporal activity detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8385–8394, June 2021. 6

[17] Pushpajit Khaire and Praveen Kumar. Deep learning and rgb-d based human action, human–human and human–object interaction recognition: A survey. *Journal of Visual Communication and Image Representation*, 86:103531, 2022. 1

[18] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, June 2022. 7

[19] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. corr abs/1811.08383 (2018), 1811. 6

[20] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 1, 5

[21] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, June 2022. 2, 6

[22] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3):205–223, 2000. 2, 3

[23] Farhood Negin, Baris Ozyer, Saeid Agahian, Sibel Kacdioglu, and Gulsah Tumuklu Ozyer. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders. *Neurocomputing*, 446:145–155, 2021. 3

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,

Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 6, 7

[25] Cristina Palmero, German Barquero, Julio CS Jacques Junior, Albert Clapés, Johnny Núnez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, et al. Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 4–52. PMLR, 2022. 1

[26] Prashant Pandey, Prathosh AP, Manu Kohli, and Josh Pritchard. Guided weak supervision for action recognition with scarce data to assess skills of children with autism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):463–470, Apr. 2020. 3, 5, 6, 7

[27] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2203–2213, 2023. 2

[28] Mauricio Perez, Jun Liu, and Alex C Kot. Interaction recognition through body parts relation reasoning. In *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part I 5*, pages 268–280. Springer, 2020. 3, 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[30] Islam Md Shafiqul, Mir Kanon Ara Jannat, Jin-Woo Kim, Soo-Wook Lee, and Sung-Hyun Yang. Hhi-attentionnet: An enhanced human-human interaction recognition method based on a lightweight deep learning model with attention network from csi. *Sensors*, 22(16):6018, 2022. 1

[31] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 6, 7

[32] Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. Spatio-temporal detection of fine-grained dyadic human interactions. In *Human Behavior Understanding: 7th International Workshop, HBU 2016, Amsterdam, The Netherlands, October 16, 2016, Proceedings 7*, pages 116–133. Springer, 2016. 1

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 6

[34] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 6

[35] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. 6

[36] Qing Ye, Haoxin Zhong, Chang Qu, and Yongmei Zhang. Human interaction recognition based on whole-individual detection. *Sensors*, 20(8):2346, 2020. 3

[37] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 1

[38] Liping Zhu, Bohua Wan, Chengyang Li, Gangyi Tian, Yi Hou, and Kun Yuan. Dyadic relational graph convolutional networks for skeleton-based human interaction recognition. *Pattern Recognition*, 115:107920, 2021. 1, 3, 6

[39] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 6