Supplementary Material for Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition

S1. Supplementary on Component Studies

S1.1. Details of Individual Pathway Models

In Section 4.3, we validated our proposed disentangled multi-scale aggregation scheme by contrasting its effectiveness on the individual pathways of the STGC blocks. Table S1 shows the individual pathway model architectures used for producing the results in Table 1 of the main paper. (For MS-GCN, MS-TCN, and MS-G3D block definitions, refer to Fig. 3 of the main paper.)

The factorized pathway models were referred to as "GCN" and "G3D" in Table 1, respectively. Keeping consistent notations, we use T, N, C to denote the number of frames, the number of nodes in the skeleton graph, and the number of feature channels, respectively. N is dataset-dependent. The blue highlighted blocks in Table S1 indicate the modules where the number of scales, K, to aggregate from skeleton graphs are adjusted as in Table 1 of the main paper. This applies to both the spatial graph $\mathcal{G}$ and the spatial-temporal graph $\mathcal{G}_{(\tau)}$.

Factorized Pathway	**G3D Pathway**	**Output Size**
Input		$300 \times N \times 3$
MS-GCN MS-TCN MS-TCN	MS-G3D	$300 \times N \times 96$
MS-TCN		$300 \times N \times 96$
MS-GCN MS-TCN (stride = 2) MS-TCN	MS-G3D (stride = 2)	$150 \times N \times 192$
MS-TCN		$150 \times N \times 192$
MS-GCN MS-TCN (stride = 2) MS-TCN	MS-G3D (stride = 2)	$75 \times N \times 384$
MS-TCN		$75 \times N \times 384$
Global Average Pooling		$1 \times 1 \times 384$
FC + Softmax		# Action Classes

Table S1. **Model architectures for individual pathways** used in Section 4.3 and Table 1. The number of scales K for MS-GCN/MS-G3D are adjusted as in Table 1 at blue highlighted blocks. The output size tuples denote number of frames (T), number of joints (N), and feature channels (C), respectively. $N = 25$ for NTU RGB+D 60 [4] and NTU RGB+D 120 [3] and $N = 18$ for Kinetics Skeleton 400.

S1.2. Visualizing Table 1 & Further Discussions

Fig. S1 visualizes Table 1 of the main paper using line charts. We can observe that when the underlying graph is denser (*i.e.* comparing spatial graphs $\mathcal{G}$ used by MS-GCN and spatial-temporal graphs $\mathcal{G}_{(\tau)}$ used by MS-G3D), the performance gaps between the two multi-scale aggregation schemes are wider. Note that when the number of scales $K = 1$, the two schemes are equivalent since multi-scale aggregation schemes do not apply.

The values of $K \in \{12, 5\}$ for the factorized/G3D pathways in the final multi-pathway model are empirically determined and involve performance/complexity trade-off. In Table 1 and Fig. S1, the values $K \in \{1, 4, 8, 12\}$ serve as evenly spread anchors for interpolation. $K = 0$ is omitted as it implies that no neighborhood aggregation is performed and graph convolutions are reduced to fully connected layers over the node features.

For spatial skeleton graphs (Fig. S1(a)), we observe that when residual masks $\mathbf{A}^{\text{res}}$ are not added (red plots of Fig. S1(a)), the accuracy gaps between the two aggregation schemes narrow as K reaches maximum. We argue that this is due to the overly sparse spatial graphs with large K under disentangled aggregation (*e.g.* excluding self-loops, $\tilde{\mathbf{A}}_{(12)}$ only has 5 edges with graph definition from NTU RGB+D [4]), where the extra branches with sparse graphs counteract the benefits of disentangled aggregation as well as multi-scale aggregation in general.

On spatial-temporal graphs (Fig. S1(b)), the effectiveness of disentangled aggregation is more evident due to the increased graph density. However, since G3D is intended to capture on *local* spatial-temporal features, a moderate scale of K gives the best performance.

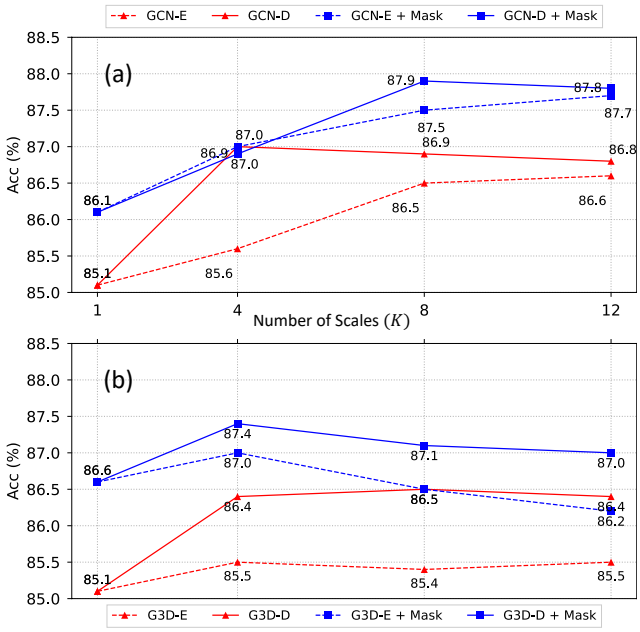

Figure S1. (Zoom in for best view.) Line charts for Table 1 of the main paper. Accuracy (%) comparison between our proposed disentangled aggregation scheme (-D) and the adjacency powering-based aggregation scheme (-E) on the Cross Subject setting of the NTU RGB+D 60 dataset using joint data only. (a) Comparing schemes using spatial graphs with GCN. (b) Comparing schemes using spatial-temporal graphs with G3D.

S1.3. Details of Graph Connectivity Ablations

In Section 4.3 and Table 3 of the main paper, we compared different graph connectivity patterns for the spatial-temporal subgraph $\mathcal{G}_{(\tau)}$ on which G3D modules perform unified spatial-temporal graph convolutions. Here, we provide matrix expressions and visualizations for the listed connectivity patterns in Table 3.

Grid-like graph connectivity refers to setting the spatial-temporal graph block adjacency matrix $\tilde{\mathbf{A}}_{(\tau)}$ to $\tilde{\mathbf{A}}^{\text{grid}}_{(\tau)}$ as defined in Eq. S1, where the submatrices along the main diagonal are set to $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, the superdiagonal/subdiagonal submatrices are set to $\mathbf{I}$, and the rest is set to $\mathbf{0}$. Fig. S2(a) illustrates the connectivity pattern when the window size $\tau = 3$. Intuitively, this setting connects skeleton graphs at different frames into larger graphs by simply introducing temporal *self-edges* between adjacent frames in the window.

$$\tilde{\mathbf{A}}^{\text{grid}}_{(\tau)} = \begin{bmatrix} \tilde{\mathbf{A}} & \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \tilde{\mathbf{A}} & \mathbf{I} & \ddots & \vdots \\ \mathbf{0} & \mathbf{I} & \tilde{\mathbf{A}} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \mathbf{I} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I} & \tilde{\mathbf{A}} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N} \quad \text{(S1)}$$

Grid-like + dense self-edges connectivity refers to setting $\tilde{\mathbf{A}}_{(\tau)}$ to $\tilde{\mathbf{A}}^{\text{densegrid}}_{(\tau)}$ as defined in Eq. S2, which adds node self-edges between *every pair* of frames on top of the "**grid-like**" setting, thus introducing *dense* self-connections along the temporal dimension. Fig. S2(b) illustrates the connectivity pattern when $\tau = 3$. Unlike the "**grid-like**" pattern where the *effective* temporal window size is 3 regardless of τ because nodes are only connected to their adjacent temporal neighbors, this pattern gives a more direct comparison to our cross-spacetime connectivity pattern (Eq. 5 of the main paper) since the number of temporal self-edges are now correlated to τ.

$$\tilde{\mathbf{A}}^{\text{densegrid}}_{(\tau)} = \begin{bmatrix} \tilde{\mathbf{A}} & \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{I} & \tilde{\mathbf{A}} & \mathbf{I} & \ddots & \vdots \\ \mathbf{I} & \mathbf{I} & \tilde{\mathbf{A}} & \ddots & \mathbf{I} \\ \vdots & \ddots & \ddots & \ddots & \mathbf{I} \\ \mathbf{I} & \cdots & \mathbf{I} & \mathbf{I} & \tilde{\mathbf{A}} \end{bmatrix} \in \mathbb{R}^{\tau N \times \tau N} \quad \text{(S2)}$$

Cross-spacetime edges as defined by Eq. 5 of the main paper are illustrated in Fig. S2(c) with $\tau = 3$.

S2. Class-wise Accuracy Comparison

Table S2 shows the class-wise performance comparison between our model and our baseline (Js-AGCN) [5] on the Cross Subject setting of NTU RGB+D 60 dataset, using the joint data only. Here, the "Factorized" setting and the "Factorized + G3D" setting refer to the third row ("MS-GCN (Factorized Pathway) Only") and the last row ("with 2 MS-G3D Pathways") of Table 2 of the main paper, respectively.

It can first be observed that our final model consistently outperforms the baseline across different action classes. We can also see that the full model (last column) outperforms the factorized pathway model (second last column) in most action classes, due to the ability of G3D to pick up complex spatial-temporal dependencies that may aid classification.

However, in the classes where long-range dependencies are more important (*e.g.* actions like "make a phone call/answer phone", "nausea or vomiting condition", and "eat meal/snack" generally have less motion over time compared to other actions), the full model slightly underperforms the factorized pathway model. We argue that this is because the full model puts significant weighting on regional spatial-temporal reasoning (using G3D pathway) in cases where the features from long-range modeling (using factorized pathway) may be more useful.

Nevertheless, for actions that have strong spatial-temporal joint correlations for making predictions (*e.g.* "drop", "take off a shoe", and "touch head (headache)"),

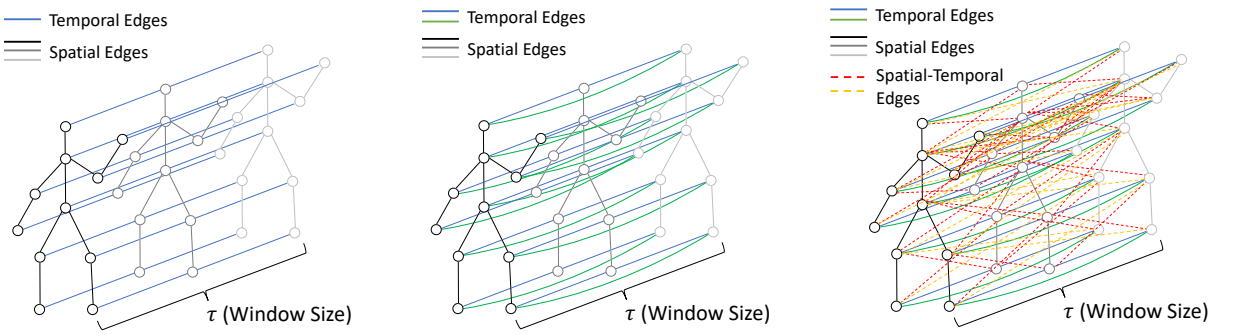

Figure S2. (Best viewed with color and zoomed in.) Visualization of various **graph connectivity settings** presented in Table 3 of the main paper. For all figures, the window size $\tau = 3$. Spatial edges with different brightness (black, gray, light gray) correspond to skeletons at different time frames. (a) Visualizing "**grid-like**" connectivity. Temporal edges connect nodes with themselves between adjacent frames. (b) Visualizing "**grid-like + dense self-edges**" connectivity. The green edges denote the extra temporal self-edges other than the adjacent temporal edges. (c) Visualizing "**cross-spacetime edges**" connectivity. The extra red and yellow edges denote the cross-spacetime edges between all pairs of frames. To help interpret the cross-spacetime connectivity, one can observe that the head joint of the first skeleton (in black) is connected to the body center joints of the second and the third skeleton, and vice versa.

the full model outperforms the baseline and the factorized model to greater extents. See Section S3.1 for examples.

S3. Visualizations

S3.1. Qualitative Analysis

Fig. S3 illustrates snapshots of an action sample in the class "drop" that is correctly classified by our model but incorrectly classified by the baseline model [5] as "tear up paper". In this case, the action closely resembles the action "tear up paper" due to the hand movements away from each other (frames (3) to (5)). However, the distinguishing clue lies in the upper body movements and the head joint movements: from frames (5) to (8), it can be observed that the skeleton leans slightly forward as an item has been dropped after frame (4), and the head joint correspondingly moves downward to reflect that the person looks at the dropped item. Being able to classify this ambiguous sample suggests that our model is capable at picking up useful spatial-temporal dependencies like these for making predictions.

Fig. S4 illustrates snapshots of an action sample in the class "take off a shoe" that is correctly classified by our model but incorrectly classified by our model baseline [5] as "wear a shoe". In this case, the skeletons became noisy in the middle of the action (frames (4) to (6)). However, strong visual cues exist at the beginning and the end of the action. The upper body movements (bending upper body) in frames (2) and (3) and the hand movement towards the feet in frame (3) have spatial-temporal correlations with the subsequent lower body movements of the foot when the person is actually taking off the shoe near frame (5). This is then correlated to the following upper body movement (standing straight) in frames (7) and (8). In particular, because a person is less likely to "wear a shoe" by starting with a shoe in his/her hand at the beginning of the action, it is more reasonable to interpret frame (3) as approaching to the foot to "take off a shoe". Being able to make a correct prediction in this case also suggests that our model was able to pick up these spatial-temporal cues even when the most representative body movements are obscured.

Fig. S5 illustrates snapshots of an action sample in the class "touch head (headache)" that is correctly classified by our model but incorrectly classified by the baseline [5] as "wear on glasses". Similar to the "drop"example shown in Fig. S3, this sample is fairly ambiguous at a first glance. However, the distinguishing feature lies in the co-occurring movements across spacetime between the head joint and the hand joints (frames (3) to (7)), as well as the magnitude of the head joint movements (compare frame (1) and (8); intuitively, "wear on glasses" would usually involve less head movements). The fact that our model makes a correct prediction on this sample suggests that it was able to correlate useful joint movement patterns across spacetime.

S3.2. Failure Cases

Fig. S6 illustrates the confusion matrix on the Cross Subject setting of the NTU RGB+D 60 dataset. It can be observed that while our model correctly classifies most of the classes, it confuses similar actions such as "writing" and "typing on a keyboard".

Fig. S7 shows two action samples from these two classes. The distinguishing features for these samples ("typing on a keyboard" and "writing") are the minor hand and forearm movements, but they are ineffectively captured by our model. Other challenges that are evident in these samples

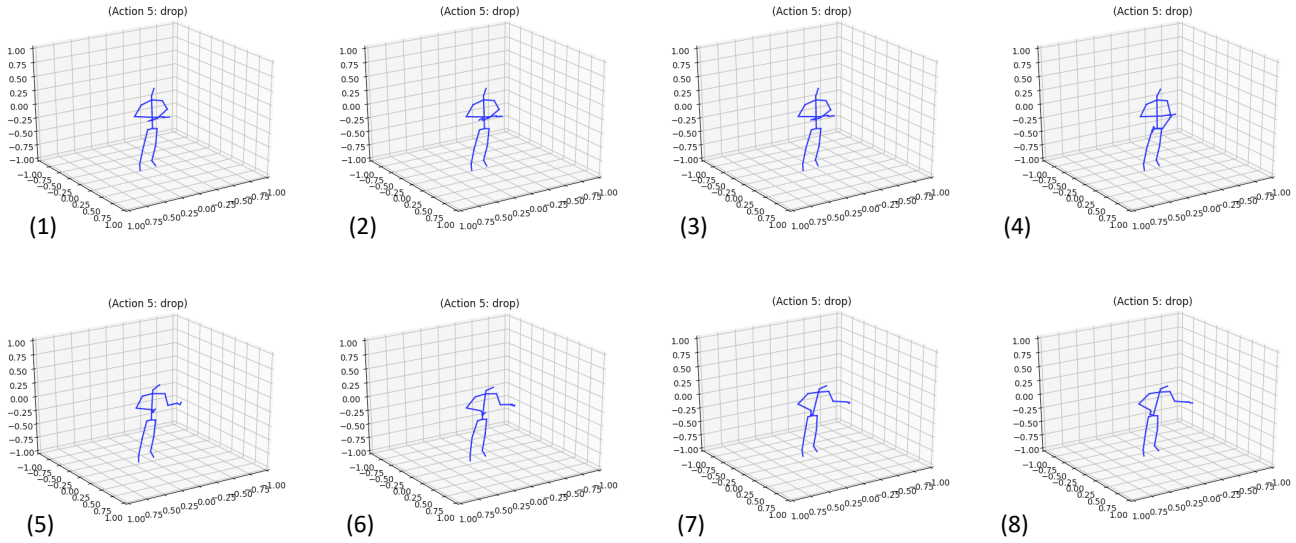

Figure S3. (Zoom in for best view.) Snapshots of an action sample from the class "**drop**" from the test set of the Cross Subject setting of NTU RGB+D 60 dataset. Frames are labeled in chronological order. Our model correctly classifies this sample while the baseline model [5] classifies this as "**tear up paper**".

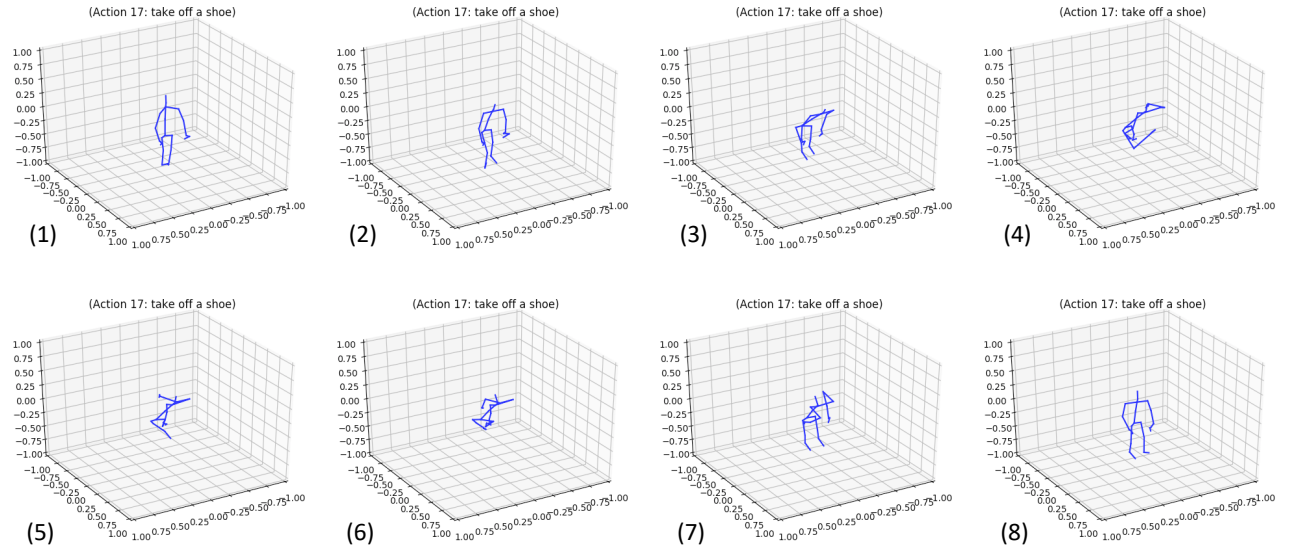

Figure S4. (Zoom in for best view.) Snapshots of an action sample from the class "**take off a shoe**" from the test set of the Cross Subject setting of NTU RGB+D 60 dataset. Frames are labeled in chronological order. Our model correctly classifies this sample despite the noisy skeletons illustrated in frames (4), (5), (6), while the baseline model [5] classifies this as "**wear a shoe**".

include noises in the skeleton data (*e.g.* changing distances between the two shoulders), and that different actions can have very similar poses (*e.g.* both skeletons are sitting).

S4. Additional Training Details

In this section we provide other training details in addition to the configurations mentioned in Section 4.2.

Parameter counts with different window sizes. As shown in Table 2 of the main paper, different configurations of G3D with different window sizes τ have different number of parameters. In this case, the extra parameters for larger τ are introduced by the fully connected layers at the end of each G3D module (see Fig. 3(d) of the main paper) for collapsing the temporal windows, rather than the graph convolution itself (Eq. 7 of the main paper).

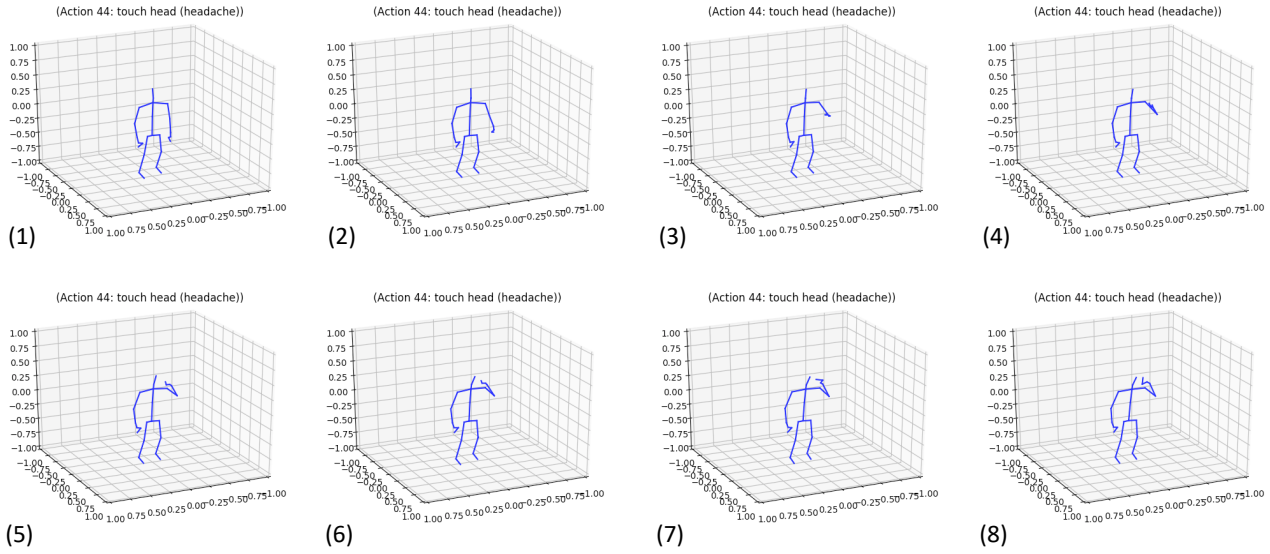

Figure S5. (Zoom in for best view.) Snapshots of an action sample from the class "**touch head (headache)**" from the test set of the Cross Subject setting of NTU RGB+D 60 dataset. Frames are labeled in chronological order. Our model correctly classifies this sample while the baseline model [5] classifies this as "**wear on glasses**".

Mixed Precision Training. Sliding temporal windows with size τ over an input graph sequence imply duplicating the input tensor by a factor of τ and operating on a larger graph adjacency of size $\tau N \times \tau N$. Thus, large values of τ many incur large memory costs. When appropriate, we used mixed precision training[1] to reduce GPU memory usage.

S5. Future Work

Attend to movements with different magnitudes. An interesting direction for future work is to design architectures to pick up minor body movements (*e.g.* of hands and feet) which may be critical for action classification. More generally, it would be useful to design mechanisms for distinguishing the importance of joint movements of different *magnitudes*. For example, the leg movements of the skeleton in Fig. S7 (left) travel greater distances than the hand movements, but the leg movements are less important for classifying the action ("typing on a keyboard").

Other GNN layers. In this work we primarily adopt GCNs [2] as the message passing module for skeleton graphs (spatial or spatial-temporal). One drawback of GCNs is their homogeneous neighborhood averaging (i.e. all neighbors have nearly the same importance under aggregation) which limits the model capacity. While in this work we deployed graph residual masks $\mathbf{A}^{\text{res}}$ to mitigate the problem, these masks are optimized for all possible actions and completely without constraints. Looking further, it would be interesting to experiment other GNN layers like graph attention layers [6] that compute individual weights of neighbors and graph isomorphism layers [7] that have higher expressive powers.

Neighborhood Sampling. As the window size τ becomes larger for G3D blocks, the resulting spatial-temporal graph $\mathcal{G}_{(\tau)}$ will have higher average node degrees due to the extra cross-spacetime edges. In these cases, it would be interesting to explore sampling based neighborhood aggregation methods like GraphSAGE [1] that may enable a trade-off between model complexity (as influenced by the size of $\mathcal{G}_{(\tau)}$) and performance.

S6. Further Acknowledgements

The authors would also like to thank Henry W. F. Yeung and Meng Zhou for useful discussions and Yuk Ying Chung for computing resources.

[1] https://github.com/NVIDIA/apex

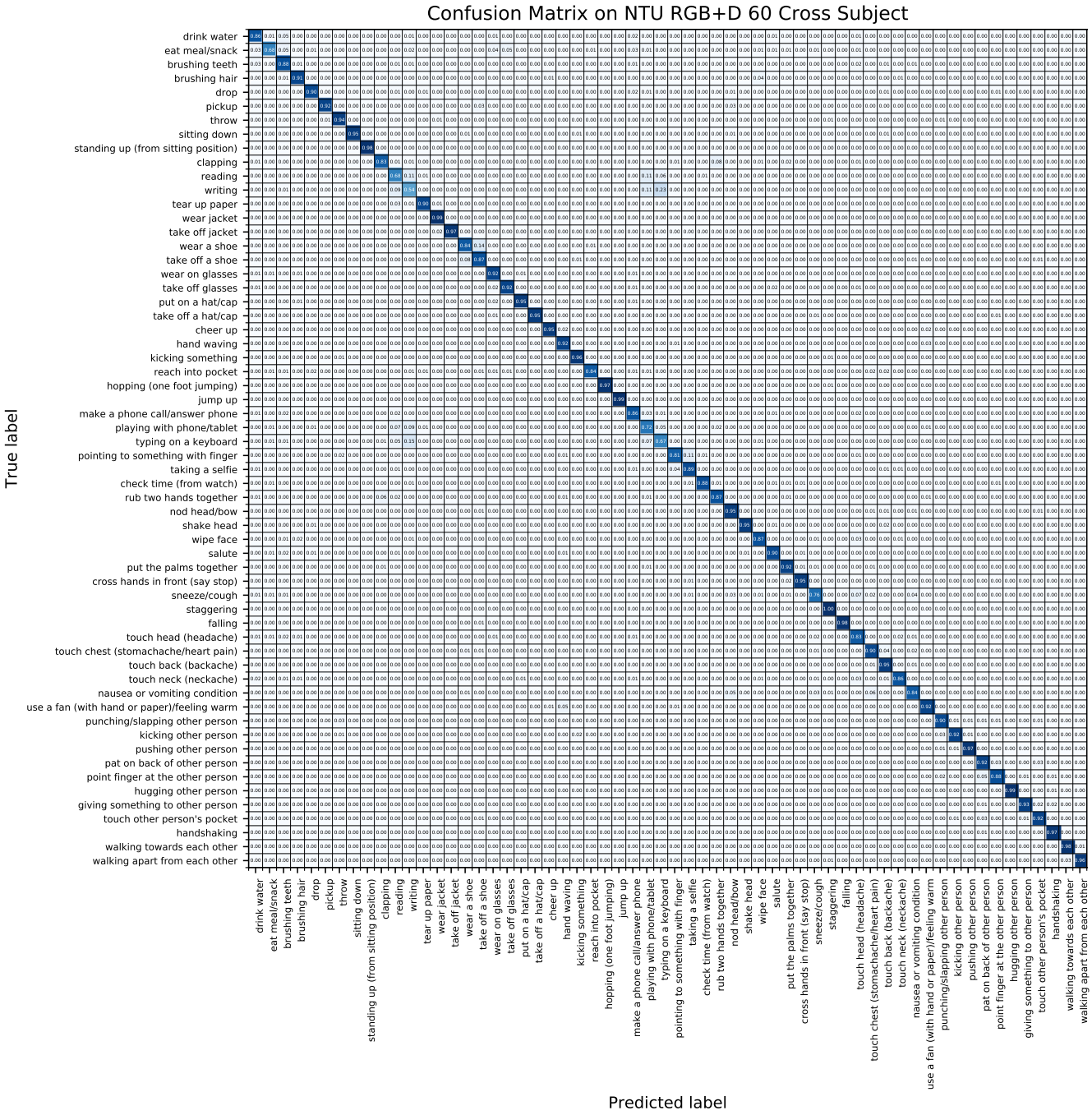

Figure S6. (Best viewed with color and zoomed in.) **Confusion matrix** on the Cross Subject setting of the NTU RGB+D 60 dataset, using the joint data only. Similar action pairs such as "writing" vs. "typing on a keyboard", and "reading" vs. "playing with phone/tablet" is more frequently misclassified by our model.

NTU RGB+D 60 Action Class	ID	Baseline [5]	Factorized (Ours)	Factorized + G3D (Ours)
drink water	1	78.8	83.9 (+5.1)	85.8 (+7.0, +1.9)
eat meal/snack	2	70.2	71.3 (+1.1)	67.6 (-2.6, -3.7)
brushing teeth	3	81.0	84.6 (+3.6)	87.5 (+6.5, +2.9)
brushing hair	4	85.0	89.0 (+4.0)	90.8 (+5.8, +1.8)
drop	5	81.5	85.1 (+3.6)	89.8 (+8.3, +4.7)
pickup	6	90.5	89.8 (-0.7)	91.6 (+1.1, +1.8)
throw	7	90.2	92.4 (+2.2)	94.2 (+4.0, +1.8)
sitting down	8	90.5	95.2 (+4.7)	94.5 (+4.0, -0.7)
standing up (from sitting position)	9	93.8	96.0 (+2.2)	98.2 (+4.4, +2.2)
clapping	10	75.5	78.0 (+2.5)	82.8 (+7.3, +4.8)
reading	11	65.2	62.3 (-2.9)	68.5 (+3.3, +6.2)
writing	12	52.6	50.7 (-1.9)	54.4 (+1.8, +3.7)
tear up paper	13	87.5	92.3 (+4.8)	90.4 (+2.9, -1.9)
wear jacket	14	98.2	97.8 (-0.4)	98.9 (+0.7, +1.1)
take off jacket	15	94.9	97.5 (+2.6)	96.7 (+1.8, -0.8)
wear a shoe	16	78.0	82.1 (+4.1)	83.9 (+5.9, +1.8)
take off a shoe	17	77.4	83.6 (+6.2)	87.2 (+9.8, +3.6)
wear on glasses	18	90.1	88.6 (-1.5)	91.6 (+1.5, +3.0)
take off glasses	19	89.8	93.1 (+3.3)	91.6 (+1.8, -1.5)
put on a hat/cap	20	93.0	95.2 (+2.2)	95.2 (+2.2, +0.0)
take off a hat/cap	21	94.5	93.8 (-0.7)	94.9 (+0.4, +1.1)
cheer up	22	88.3	94.2 (+5.9)	95.3 (+7.0, +1.1)
hand waving	23	90.1	92.0 (+1.9)	92.3 (+2.2, +0.3)
kicking something	24	95.3	96.7 (+1.4)	96.4 (+1.1, -0.3)
reach into pocket	25	81.8	83.6 (+1.8)	83.9 (+2.1, +0.3)
hopping (one foot jumping)	26	95.6	96.7 (+1.1)	97.5 (+1.9, +0.8)
jump up	27	97.5	99.3 (+1.8)	99.3 (+1.8, +0.0)
make a phone call/answer phone	28	85.5	90.2 (+4.7)	86.2 (+0.7, -4.0)
playing with phone/tablet	29	64.4	69.1 (+4.7)	72.4 (+8.0, +3.3)
typing on a keyboard	30	65.8	68.4 (+2.6)	67.3 (+1.5, -1.1)
pointing to something with finger	31	73.2	79.7 (+6.5)	81.2 (+8.0, +1.5)
taking a selfie	32	85.1	88.4 (+3.3)	89.1 (+4.0, +0.7)
check time (from watch)	33	83.7	90.2 (+6.5)	88.4 (+4.7, -1.8)
rub two hands together	34	84.8	85.5 (+0.7)	86.6 (+1.8, +1.1)
nod head/bow	35	87.7	94.2 (+6.5)	94.6 (+6.9, +0.4)
shake head	36	88.7	91.6 (+2.9)	94.5 (+5.8, +2.9)
wipe face	37	81.9	87.3 (+5.4)	87.3 (+5.4, +0.0)
salute	38	89.9	89.1 (-0.8)	89.9 (+0.0, +0.8)
put the palms together	39	89.1	90.6 (+1.5)	92.4 (+3.3, +1.8)
cross hands in front (say stop)	40	93.1	92.8 (-0.3)	94.9 (+1.8, +2.1)
sneeze/cough	41	69.2	73.9 (+4.7)	75.7 (+6.5, +1.8)
staggering	42	97.1	97.8 (+0.7)	99.6 (+2.5, +1.8)
falling	43	95.6	97.5 (+1.9)	97.8 (+2.2, +0.3)
touch head (headache)	44	71.4	79.0 (+7.6)	83.3 (+11.9, +4.3)
touch chest (stomachache/heart pain)	45	87.7	90.6 (+2.9)	90.2 (+2.5, -0.4)
touch back (backache)	46	93.1	91.7 (-1.4)	94.9 (+1.8, +3.2)
touch neck (neckache)	47	80.4	88.4 (+8.0)	86.2 (+5.8, -2.2)
nausea or vomiting condition	48	83.3	87.3 (+4.0)	84.0 (+0.7, -3.3)
use a fan (with hand or paper)/feeling warm	49	89.1	89.5 (+0.4)	91.6 (+2.5, +2.1)
punching/slapping other person	50	88.7	92.0 (+3.3)	89.8 (+1.1, -2.2)
kicking other person	51	90.9	90.9 (+0.0)	92.0 (+1.1, +1.1)
pushing other person	52	91.3	95.7 (+4.4)	97.1 (+5.8, +1.4)
pat on back of other person	53	87.3	89.5 (+2.2)	92.4 (+5.1, +2.9)
point finger at the other person	54	83.0	88.8 (+5.8)	87.7 (+4.7, -1.1)
hugging other person	55	97.1	98.2 (+1.1)	98.5 (+1.4, +0.3)
giving something to other person	56	92.8	91.3 (-1.5)	93.1 (+0.3, +1.8)
touch other person's pocket	57	88.0	92.4 (+4.4)	92.0 (+4.0, -0.4)
handshaking	58	96.4	95.3 (-1.1)	97.1 (+0.7, +1.8)
walking towards each other	59	98.5	97.8 (-0.7)	97.8 (-0.7, +0.0)
walking apart from each other	60	95.7	95.7 (+0.0)	96.0 (+0.3, +0.3)

Table S2. **Class-wise classification accuracy** (%) for each of the 60 classes on the Cross Subject setting of the NTU RGB+D 60 Skeleton dataset, using joint data only (*i.e.* no joint-bone two-stream fusion). We use the joint stream of 2s-AGCN [5] as the performance baseline. Values in the parentheses in the "**Factorized**" column are accuracy differences compared to the baseline. Values in the parentheses in the "**Factorized + G3D**" column are accuracy differences compared to the baseline and the "**Factorized**" pathway model, respectively.

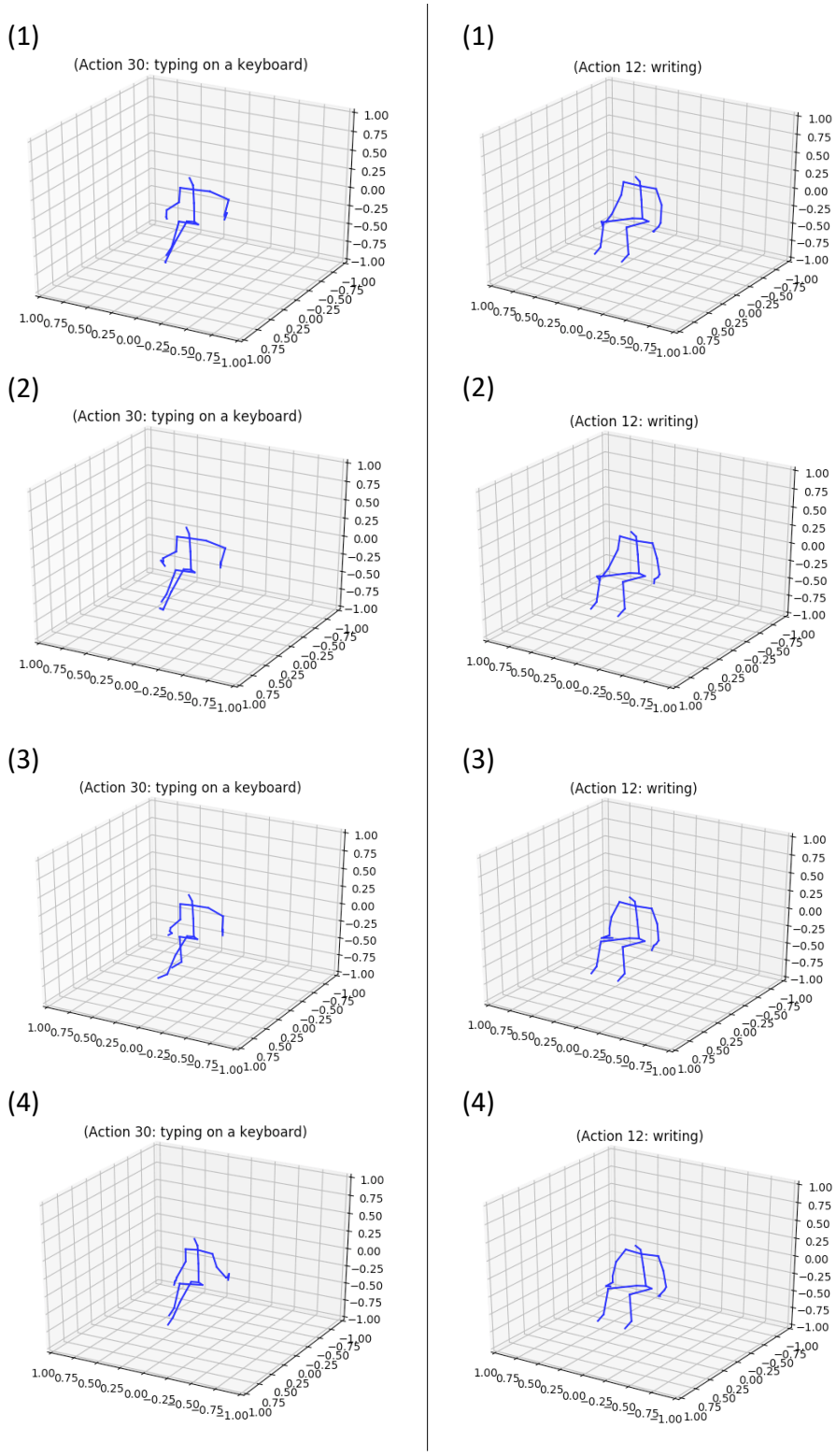

Figure S7. Two sample **failure cases** where our model fails to distinguish the action classes. Action snapshots are ordered from top to bottom in chronological order. **Left:** Visualizing a sample from action class "typing on a keyboard" misclassified as "writing" by our model. **Right:** Visualizing a sample from action class "writing" misclassified as "typing on a keyboard" by our model.

References

[1] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017. 5

[2] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 5

[3] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019. 1

[4] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 1

[5] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 2, 3, 4, 5, 7

[6] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[7] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019. 5