CrossGLG: LLM Guides One-shot Skeleton-based 3D Action Recognition in a Cross-level Manner

Tingbing Yan[1], Wenzheng Zeng[1†], Yang Xiao[1†], Xingyu Tong[1], Bo Tan[1], Zhiwen Fang[2], Zhiguo Cao[1], and Joey Tianyi Zhou[3,4,5]

[1] Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2] School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
[3] CFAR, Agency for Science, Technology and Research, Singapore
[4] IHPC, Agency for Science, Technology and Research, Singapore
[5] Centre for Advanced Technologies in Online Safety (CATOS), Singapore
{yantingbing,wenzhengzeng,Yang_Xiao,xy_tong,bo_tan,zgcao}@hust.edu.cn, fzw310@smu.edu.cn, zhouty@cfar.a-star.edu.sg

Å1 CrossGLG from a plug-and-play module perspective

Actually, CrossGLG can be regarded as a plug-and-play module that enhances the feature learning of different existing skeleton encoders (*e.g.*, MotionBERT [8], HDGCN [3], and InfoGCN [1]). All we need to do is to change the base skeleton encoder in our skeleton encoding branch. To verify its effectiveness, as shown in Tab. Å1, we conducted experiments on NTU120 [4] and applied CrossGLG to MotionBERT [8], HDGCN [3], and InfoGCN [1], respectively. It can be seen that CrossGLG delivers improvements in all settings. Moreover, CrossGLG only brings a neglectable cost compared with the original skeleton encoders, thanks to our dual-branch design. This proves that CrossGLG can be used as a plug-and-play module with good generalization capability.

#Base Classes	20	40	60	80	100	Para(M)
MotionBERT [8]	35.5	54.3	56.5	52.8	61.0	60.3
MotionBERT [8]+CrossGLG	**51.0**(↑15.8)	56.9(↑2.6)	58.4(↑1.9)	54.2(↑1.4)	62.5(↑1.5)	60.8(+0.5)
HDGCN [3]	39.0	50.4	55.8	49.8	51.9	1.7
HDGCN [3]+CrossGLG	43.0(↑4.0)	56.7(↑6.3)	57.4(↑1.6)	55.9(↑6.1)	56.8(↑4.9)	1.8(+0.1)
InfoGCN [1]	37.0	53.9	<u>58.8</u>	<u>55.7</u>	56.1	1.6
InfoGCN [1]+CrossGLG	45.3(↑8.3)	**56.8**(↑2.9)	**62.1**(↑3.3)	**61.6**(↑5.9)	**62.6**(↑6.5)	1.7(+0.1)

Table Å1: The improvements CrossGLG can bring to different backbones for one-shot action recognition on NTU 120 [4].

† Yang Xiao and Wenzheng Zeng are corresponding authors.

Å2 Implementation details

The skeleton encoding branch is based on InfoGCN [1]. During training, the learning rate is set to 0.05, batch_size is set to 128, and the number of training epochs is 110. α_1 and α_2 are set to 0.5 and 0.2, respectively. The random seed is 0. During testing, the batch_size is also set to 128. When applying GAP [6] to the 1-shot task, as described in its paper, the learning rate is set to 0.1, and the trade-off parameter λ is set to 0.8. Additionally, in the experiments of Sec. Å1, the learning rate for HDGCN [3] is set to 0.05. The learning rate for Motion-BERT's [8] backbone is set to 0.00008. α_1 and α_2 are also set to 0.5 and 0.2, respectively, while other settings remain unchanged from their initial configurations. We use Pytorch to implement our work. All models in our experiments were done on two NVIDIA RTX 3090 GPUs.

Å3 Training and evaluation protocol on Kinetics

We followed the skeleton extraction as well as the processing method in [7] and sampled each skeleton sequence evenly over 60 frames. Since kinetics [2] did not have an official 1-shot setting before, we designed a fair base classes set and novel classes set split following NTU120 [4]. As described in Sec. 3.1, the model will be trained using a $\mathcal{D}_{base}$ consisting of all the samples from the training set. One sample from each of the Novel classes will be randomly selected to form a support set $\mathcal{S}_{novel}$ to help the model categorize the unseen samples from the Novel classes. The support set $\mathcal{S}_{novel}$ is kept constant in the same setting. The following are the splits for the 20 base classes and the 40 base classes, respectively:

- **20-class:**
  - **Train:** [2, 22, 42, 62, 82, 102, 122, 142, 162, 182, 202, 222, 242, 262, 282, 302, 322, 342, 362, 382]
  - **Novel:** [3, 23, 43, 63, 83, 103, 123, 143, 163, 183, 203, 223, 243, 263, 283, 303, 323, 343, 363, 383]
- **40-class:**
  - **Train:** [2, 4, 22, 24, 42, 44, 62, 64, 82, 84, 102, 104, 122, 124, 142, 144, 162, 164, 182, 184, 202, 204, 222, 224, 242, 244, 262, 264, 282, 284, 302, 304, 322, 324, 342, 344, 362, 364, 382, 384]
  - **Novel:** [3, 23, 43, 63, 83, 103, 123, 143, 163, 183, 203, 223, 243, 263, 283, 303, 323, 343, 363, 383]

Å4 The impact of hyperparameters α_1 and α_2

α_1 and α_2 are used to balance the overall effect of the proposed losses $\mathcal{L}_{calibrate}$ and $\mathcal{L}_c$. Tab. Å2 shows that the performance is relatively robust to their values.

α_1				α_2			
0	0.2	0.5	0.6	0	0.2	0.3	0.5
42.5	43,6	45.3	44.9	43.3	45.3	44.8	45.1

Table Å2: Choice for α_1 and α_2 at 20 base classes settings on NTU120 [4].

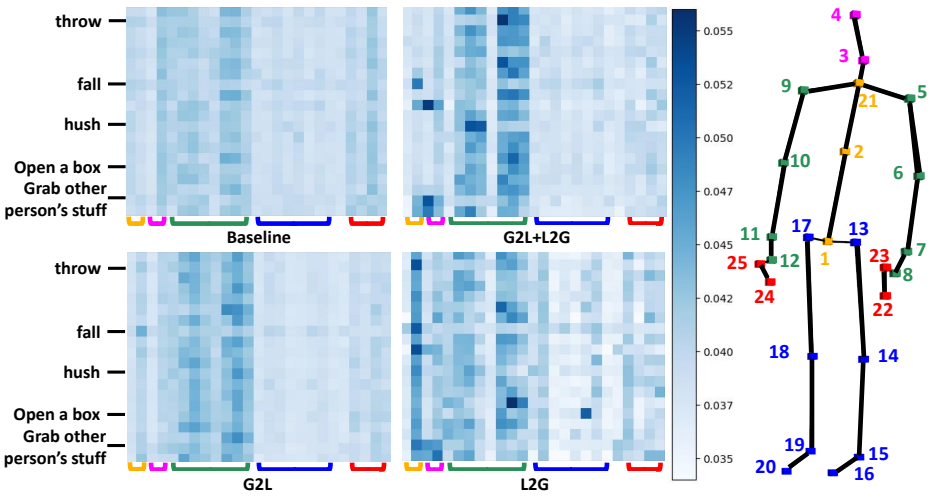

Fig. Å1: Visual analysis of spatial attention within the skeleton encoder. In each matrix, each row represents the importance of all encoding blocks for each joint feature in a novel action. The deeper the color, the higher the importance.

Å5 Limitations

CrossGLG realizes the introduction of human knowledge in text from two perspectives, global-to-local and local-to-global, respectively. Although results from different experimental settings have demonstrated the effectiveness of CrossGLG, it still has some room for improvement in some extremely difficult-to-distinguish situations. For example, there are still a certain number of misjudgments in the actions 'yawn' and 'hush'.

Å6 Visual analysis

We visualize the attention of the skeleton encoder to analyze its behaviors, which is shown in Fig. Å1. In each visualization matrix, each row represents the spatial attention of each joint in a novel action. The color assigned to each joint on the right side in Fig. Å1 corresponds to the color representing its respective position at the bottom of the matrix. At the end of the training of the model, we freeze the skeleton encoding branch and directly examine the spatial attention of the novel actions. G2L, L2G, and a combination of the two can focus more on the

really important joints in the actions than the baseline. For example, in the actions 'throw' and 'open a box', the arms and hands are more important. In the action 'fall', which involves the joints of the whole body, the network gives similar importance to each joint according to the characteristics of the action, meaning that the network pays attention to all joints.

Å7 Text descriptions

We use large language models (such as ChatGPT [5]) to obtain the global action description and joint motion descriptions for each action. The global action description indicates which joints and body parts are important from a global perspective. The joint-level motion description provides fine-grained high-level semantic information from a local perspective. The global motion descriptions and joint motion descriptions for several actions are shown below:

Wave hand

- **Global action description:** one **arm** is raised and moved side to side, typically with the **hand** open, often performed in a rhythmic motion, involving a swinging movement at the **elbow** joint while the **wrist** creates a slight flexion and extension
- **joint motion descriptions:**
  - **Base of spine:** Generally remains stable, providing a balanced foundation for the upper body.
  - **Mid of spine:** Typically straight and upright, supporting the upper body's movements.
  - **Neck:** Allows for flexibility and rotation, facilitating head movement.
  - **Head:** May tilt slightly to the side or stay upright, often facing the direction of the hand waving.
  - **Left shoulder:** Lifts and rotates slightly as the arm moves to wave.
  - **Left elbow:** Bends and straightens rhythmically to create the waving motion.
  - **Left wrist:** Flexes and extends, contributing to the waving movement.
  - **Left hand:** Performs a waving motion, lifting and lowering in a friendly gesture.
  - **Right shoulder:** Remains engaged in supporting the arm's movement.
  - **Right elbow:** Bends and straightens in coordination with the waving action.
  - **Right wrist:** Flexes and extends, mirroring the movements of the left wrist.
  - **Right hand:** Performs the waving action, lifting and lowering in a friendly gesture.
  - **Left hip:** Typically stable during hand waving, but may sway slightly with body movement.
  - **Left knee:** Generally remains straight but can bend slightly with body weight shifts.

- **Left ankle:** Often stationary or may have subtle movements with body weight shifts.
- **Left foot:** Remains grounded, providing stability.
- **Right hip:** Similar to the left hip, usually stable during hand waving.
- **Right knee:** Generally straight but may bend slightly with body weight shifts.
- **Right ankle:** Often stationary or may have subtle movements with body weight shifts.
- **Right foot:** Remains grounded, providing stability.
- **Spine:** Supports an upright posture, allowing for coordination of upper body movements.
- **Tip of Left hand:** Moves in a waving motion, extending and retracting.
- **Left thumb:** Provides stability during the waving action.
- **Tip of Right hand:** Mimics the waving motion of the left hand, extending and retracting.
- **Right thumb:** Offers stability and support during the waving action.

Squat down

– **Global action description:** the **knees** and **hips** bend while the torso lowers, maintaining an upright posture, and the **arms** may extend forward or rest on the thighs, with the body weight shifting towards the heels, engaging the **leg** muscles for support and balance.
– **joint motion descriptions:**
  - **Base of spine:** The base of the spine is actively engaged in the action, supporting the weight of the upper body as it moves downward.
  - **Mid of spine:** The mid of the spine also participates by flexing forward slightly as the upper body leans forward during the squat.
  - **Neck:** The neck remains in line with the spine, maintaining a neutral position.
  - **Head:** The head aligns with the spine and does not actively move during the squat.
  - **Left shoulder:** The left shoulder remains stationary, with no significant movement.
  - **Left elbow:** The left elbow may slightly flex as the left arm moves with the body during the squat.
  - **Left wrist:** The left wrist maintains its natural position, without significant movement.
  - **Left hand:** The left hand is relaxed and usually hangs down as the body squats.
  - **Right shoulder:** Similar to the left shoulder, the right shoulder remains stationary.
  - **Right elbow:** The right elbow may slightly flex as the right arm moves with the body during the squat.
  - **Right wrist:** The right wrist maintains its natural position, without significant movement.

- **Right hand:** The right hand is relaxed and usually hangs down during the squat.
- **Left hip:** The left hip is actively involved in the squatting motion, allowing the left thigh to move backward.
- **Left knee:** The left knee flexes as the body lowers, allowing the squatting motion to occur.
- **Left ankle:** The left ankle allows for dorsiflexion, which aids in the squatting movement.
- **Left foot:** The left foot remains in contact with the ground, providing support.
- **Right hip:** Similar to the left hip, the right hip is actively involved in the squatting motion, allowing the right thigh to move backward.
- **Right knee:** The right knee flexes as the body lowers, mirroring the movement of the left knee.
- **Right ankle:** The right ankle allows for dorsiflexion, aiding in the squatting movement.
- **Right foot:** The right foot remains in contact with the ground, providing support.
- **Spine:** The entire spine is involved in the squatting motion, with the base and mid of the spine flexing slightly forward to maintain balance.
- **Tip of Left hand:** The tip of the left hand usually remains relaxed and hangs down during the squat.
- **Left thumb:** The left thumb remains in a relaxed position, with no significant movement.
- **Tip of Right hand:** Similar to the left hand, the tip of the right hand usually remains relaxed and hangs down.
- **Right thumb:** The right thumb remains in a relaxed position, with no significant movement.

Count money

– **Global action description:** the **hands** are involved in picking up, sorting, and organizing the bills or coins, while the **eyes** focus on visually inspecting and verifying the amount, and the **head** is engaged in mental calculations and cognitive processing to keep track of the counting progress
– **joint motion descriptions:**
  - **Base of spine:** The base of the spine generally remains stable while sitting or standing.
  - **Mid of spine:** The mid of the spine maintains its natural posture, with no significant movement.
  - **Neck:** The neck is often upright or slightly bent forward, depending on the position of the money.
  - **Head:** The head is typically upright, and the gaze is directed toward the money being counted.
  - **Left shoulder:** The left shoulder usually remains stable and doesn't actively move.

- **Left elbow:** The left elbow is bent at an angle to bring the currency closer for counting.
- **Left wrist:** The left wrist is typically flexed as it guides the movement of bills or coins.
- **Left hand:** The left hand holds and manipulates the money, separating and organizing it.
- **Right shoulder:** The right shoulder typically remains stable for balance.
- **Right elbow:** The right elbow may be bent slightly as it assists in the movement of money.
- **Right wrist:** The right wrist is often flexed to help manage the currency's position.
- **Right hand:** The right hand holds and manipulates the money, assisting in counting and organizing.
- **Left hip:** The left hip provides balance and support but doesn't actively move.
- **Left knee:** The left knee may be slightly bent or adjusted for comfort while sitting or standing.
- **Left ankle:** The left ankle generally maintains its neutral position.
- **Left foot:** The left foot is typically stationary during the counting process.
- **Right hip:** The right hip provides balance and support but doesn't actively move.
- **Right knee:** The right knee may be slightly bent or adjusted for comfort while sitting or standing.
- **Right ankle:** The right ankle generally maintains its neutral position.
- **Right foot:** The right foot is typically stationary during the counting.
- **Spine:** The spine retains its natural posture with no significant movement.
- **Tip of Left hand:** The tip of the left hand is actively engaged in handling and counting the money.
- **Left thumb:** The left thumb may assist in separating or flipping bills.
- **Tip of Right hand:** The tip of the right hand is actively engaged in handling and counting the money.
- **Right thumb:** The right thumb may assist in separating or flipping bills.

References

1. Chi, H.g., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20186–20196 (2022)
2. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
3. Lee, J., Lee, M., Lee, D., Lee, S.: Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10444–10453 (2023)
4. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: 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)
5. OpenAI: Introducing chatgpt. Website (2022), `https://openai.com/index/chatgpt`
6. Xiang, W., Li, C., Zhou, Y., Wang, B., Zhang, L.: Generative action description prompts for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10276–10285 (2023)
7. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI Conference on Artificial Intelligence (2018), `https://api.semanticscholar.org/CorpusID:19167105`
8. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15085–15099 (2023)