# 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

**Abstract.** Most existing one-shot skeleton-based action recognition focuses on raw low-level information (*e.g.*, joint location), and may suffer from local information loss and low generalization ability. To alleviate these, we propose to leverage text description generated from large language models (LLM) that contain high-level human knowledge, to guide feature learning, in a global-local-global way. Particularly, during training, we design 2 prompts to gain global and local text descriptions of each action from an LLM. We first utilize the global text description to guide the skeleton encoder focus on informative joints (*i.e.*,global-to-local). Then we build non-local interaction between local text and joint features, to form the final global representation (*i.e.*, local-to-global). To mitigate the asymmetry issue between the training and inference phases, we further design a dual-branch architecture that allows the model to perform novel class inference without any text input, also making the additional inference cost neglectable compared with the base skeleton encoder. Extensive experiments on three different benchmarks show that CrossGLG consistently outperforms the existing SOTA methods with large margins, and the inference cost (model size) is only 2.8% than the previous SOTA. Code is available at CrossGLG.

**Keywords:** 3D skeleton-based action recognition · One-shot · LLM

## 1 Introduction

Driven by the accessibility of low-cost 3D cameras like the Microsoft Kinect [43], 3D skeleton-based action recognition has emerged as an active area of research.

† Yang Xiao and Wenzheng Zeng are corresponding authors.

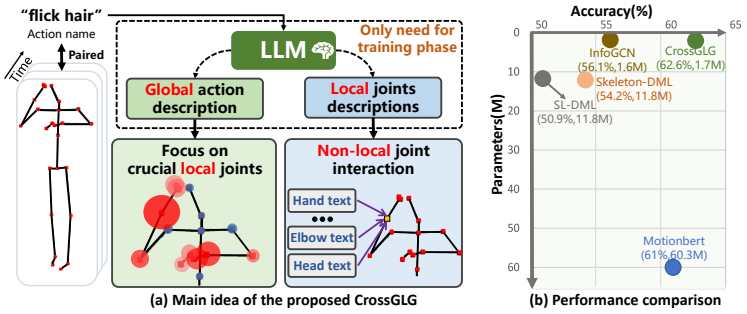


**Fig. 1:** (a) Main idea of the CrossGLG: We propose to leverage text description generated from large language models (LLM) that contain high-level human knowledge to guide feature learning, in a global-local-global way. In global-to-local (block in green), the larger the radius of the circle around a joint, the more important that joint is. In local-to-global (block in blue), non-local interaction establishes connections between all textual features and all skeleton features at the joint level to summarize the high-level global action representation. (b) Performance comparison on NTU120 [19] dataset (100 base classes): CrossGLG outperforms the current SOTA in effectiveness and efficiency.

Despite the remarkable progress, most works [17, 18, 27, 40] focus on fully supervised settings that heavily rely on large-scale annotated data for feature learning, introducing high annotation costs. One feasible way to alleviate this is to conduct one-shot 3D action feature learning.

To this end, some pioneer [19, 22, 23, 37, 45, 46] attempts have been made. They are all based on the raw 3D skeleton data, focusing on low-level information (*e.g.*, joint localization). While they tend to suffer from the following defeats: (1) Existing methods generally can not focus on the crucial local areas, which leads to the loss of important details. (2) The lack of high-level semantic information to guide the models makes it difficult to grasp the deep overall motion characteristics, resulting in a weak generalization for unseen actions.

Some studies [3,10] in the fields of psychology and neuroscience show that humans can easily identify critical motion clues and extrapolate local observations into global conclusions to recognize actions, even with a limited amount of observations. Inspired by this, our motivation is to utilize high-level human knowledge to guide feature learning, for effective one-shot action recognition. Thanks to the recent success of large language models trained on massive data [4, 20, 25, 38], we can manage to obtain high-level human knowledge from the text they generate. Thus, we propose CrossGLG, a novel architecture that utilizes knowledgeable text descriptions to guide skeleton feature learning in a global-local-global way.

Particularly, as shown in Fig. 1(a), the paradigm can be divided into 2 main phases: global-to-local (block in green), and local-to-global (block in blue). We first a utilize global action description to guide the skeleton encoder to focus on local informative joints (*i.e.*, global-to-local). The global action description is obtained from LLMs by a designed prompt, which illustrates the important joints when a specific action occurs. We regard those joints as informative joints and

design a Joint Importance Determination Module to let the skeleton encoder focus on those informative joints. Such global-to-local guidance can let the skeleton encoder focus more on the important local clues for effective representation.

Based on the enhanced local joint features, we further build a non-local interaction between local joint skeleton features and local joint-level motion descriptions generated from the LLM, to achieve the so-called local-to-global. Within it, effective information exchange between local joint-level text-text, skeleton-skeleton, and skeleton-text has been established to enhance the global-aware ability of local features, forming a global summary of the action. In conclusion, the proposed cross-modal global-local-global guiding strategy lets the skeleton encoder first identify critical motion clues and then extrapolate local observations into global conclusions to effectively recognize actions.

Another issue is that for practical applications, the network is not capable of using text data during inference. To mitigate the asymmetry issue between the training and inference phases, we design a dual-branch architecture. Specifically, we divide the whole framework into a skeleton encoding branch and a cross-modal guidance branch, where the skeleton encoding branch is only responsible for skeleton feature encoding, and the cross-modal guidance branch is responsible for guiding the feature learning of the skeleton encoding branch, in the aforementioned global-local-global way. During training, we introduce a shared classifier to those two branches to reduce the feature gap between them. Such design not only lets features in the skeleton branch learn from the cross-modal guidance branch during training, but also keeps those two-branch features consistent so that only the skeleton encoder is needed during inference, without the need for any additional modal data.

The proposed CrossGLG makes the feature encoder learn fine-grained high-level semantic information from the text description both globally and locally. As shown in Fig. 1(a), in the action 'flick hair', the joints of the arm and hand are of greater importance than joints such as the spine. Meanwhile, our visualization results show that for the novel class action that was not seen during training, the model without fine-tuning can still pay more attention to the informative clues. This implies that the model's spatial information grasping ability has been greatly improved and can really focus on important local regions with strong generalization ability. Extensive experiments on three benchmark (*i.e.*, NTU RGB+D 60 [31], NTU RGB+D 120 [19], and Kinetics [15]) also validate the effectiveness of the proposed method. Our method achieves state-of-the-art performance in terms of accuracy and efficiency. Notably, CrossGLG outperforms the SOTA methods by 6.6% and 5.9% in the experimental settings of 20 and 80 base classes on NTU RGB+D 120 [19], while maintaining a model size that is only 2.8% of the previous SOTA [46] during inference.

In summary, our contributions are summarized below:

• We are the first to consider text descriptions from large language models to facilitate one-shot skeleton-based 3D human action recognition.

• We design CrossGLG, a novel architecture that utilizes knowledgeable text descriptions to guide skeleton feature learning in a global-local-global way, for effective one-shot 3D action recognition.

• We propose a dual-branch architecture to address the asymmetry issue between training and testing.

## 2 Related Work

**One-shot 3D skeleton-based action recognition.** To alleviate the labor-consuming issues in 3D action recognition, one-shot skeleton-based action recognition has emerged as a research focal point and some efforts [9, 47] have been paid. These works focus on enhancing the adaptation ability of feature extractor and classifier from base class to novel class via imitating the one-shot test cases using base class data during training. There is also a portion of work [19, 22, 23, 37, 42, 46] that opts for another approach, matrix learning techniques. They enhance the discriminative properties of generated features by exploring the construction of latent feature space, widening inter-class feature distances while narrowing intra-class gaps.

However, these works only focus on the low-level information of the skeleton sequence (*i.e.*, joint locations). Due to the lack of guidance from high-level semantic information, these methods face problems such as low generalization ability. Although APSR [19] has tried to introduce semantic information, this introduction is too little. And it cannot realize the detection of important joints during inference, which limits the improvement of the discriminative properties of the test features. Unlike these previous approaches, we utilize human-knowledge-rich action description texts to introduce high-level semantic information to guide the learning of skeleton features from both global and local perspectives. Our research focus is orthogonal and complementary to existing works, such as the matrix learning ideas.

**Skeleton action recognition with cross-modal information.** Some existing works also utilize auxiliary information to facilitate skeleton feature learning, which is mainly based on two types of additional inputs: visual data [5, 13, 14, 30, 32, 35, 44] and non-visual data [39]. The first class mainly utilizes RGB or depth data. They design a two-stream network to process visual data and skeleton data separately and perform score fusion, feature fusion, or co-learning. However, these visual data may suffer from the problems of expensive acquisition, noise, and inefficient semantic knowledge. In the second category, GAP [39] performs contrastive learning between skeleton features and text features to facilitate feature learning. Despite the effectiveness, it treats all local features equally, which reduces the discriminative power. Besides, it simply conducts contrastive learning within each local-level and global-level skeleton-text feature pair, lacking the cross and non-local interactions between different parts of the body, which limits the global-aware ability of local features. Moreover, it is designed for fully supervised setting, and the experiment (Tab. 1) demonstrates its inferior generalization performance in the more challenging one-shot setting. Suitable design under

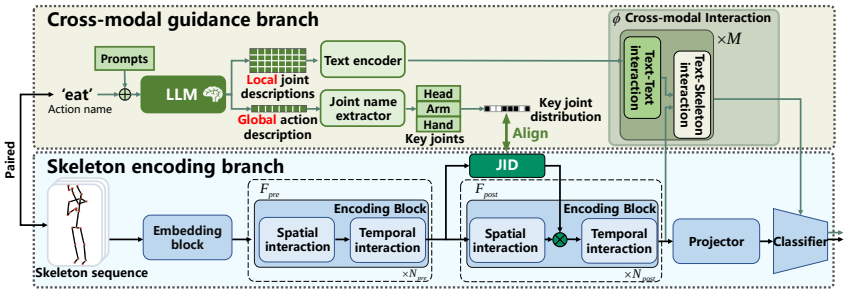


**Fig. 2:** Overview of model architecture: JID (Joint Importance Discrimination) outputs joint significance from skeleton features. Green highlights cross-modal guidance for skeleton learning. Blue indicates the skeleton encoding branch, used only for novel class inference without textual input.

the challenging one-shot learning paradigm has not been well concerned yet. To our knowledge, we are the first to consider knowledgeable text descriptions to guide skeleton feature learning for one-shot 3D action recognition. We propose a novel cross-modal guidance framework that runs in a global-local-global way. Experiments demonstrate the significant superiority of our novel design.

**Large Language Model (LLM).** Benefiting from unique training mechanisms [8, 16, 20, 25, 26, 38], massive parameters [4, 28], and sufficient training data [29], large language models (*e.g.*, ChatGPT [24]) typically can understand a wider and more complex language context. This enables them to better understand the semantics and contextual relevance in the text and maintain strong text generation capabilities. That is, given the corresponding prompt, LLMs can generate logical and detailed results. Inspired by this, we propose to utilize the generative power of large language models to assist us in extracting global and local strongly semantic heuristic information about actions.

## 3 Method

Here we will introduce the proposed CrossGLG, a novel cross-modal guided one-shot skeleton-based human action recognition architecture. We will briefly recap the task formulation of one-shot action recognition in Sec. 3.1. The architecture overview will be illustrated in Sec. 3.2, followed by detailed descriptions of each component (from Sec. 3.3 to Sec. 3.6).

### 3.1 Preliminary

In one-shot 3D action recognition, we are given a labeled base class dataset $\mathcal{D}_{base}$ and a novel class dataset $\mathcal{D}_{novel}$, where the two datasets do not overlap. $\mathcal{D}_{novel}$ consists of N action classes, where one sample of each class along with its label is drawn to form the support set $\mathcal{S}_{novel}$, and the others to form the query set $\mathcal{Q}_{novel}$. The goal is to train the model using the $\mathcal{D}_{base}$ to complete the training of the model and make adaptations on $\mathcal{S}_{novel}$ to make predictions for $\mathcal{Q}_{novel}$.

### 3.2 Architecture Overview

The architecture of the proposed CrossGLG is shown in Fig. 2, which contains two branches: the skeleton encoding branch and the cross-modal guidance branch. The skeleton encoding branch takes a skeleton sequence as input, encodes it, and ultimately performs action classification. We utilize the proposed cross-modal guidance branch to guide the skeleton feature learning. Specifically, based on the paired action name of the input skeleton sequence, we design two kinds of prompts that allow an LLM (*e.g.*, ChatGPT) to generate both global action description, and local joint-level motion description of the action (Sec. 3.3). We use the obtained global action description to guide the skeleton encoder focusing on local informative joints, by aligning the output distribution between the designed Joint Importance Determination Module (JID) and the key joints distribution derived from the global action description. We further introduce a cross-modal interaction module $\phi$ to build strong non-local interactions between local joint features and joint-level text features, and summarize the high-level global information for the entire action. The two-branch outputs will share the same classifier for the final action classification. During inference, only the skeleton encoding branch will be used to conduct one-shot novel class classification, without using any auxiliary text information.

### 3.3 Derive Knowledgeable Action Descriptions from Large Language Model (LLM)

Our motivation is to utilize high-level human knowledge to guide the skeleton feature learning, for a more effective one-shot action recognition. Text is actually a good carrier to involve such knowledge, as it is low-noise and rich in high-level semantic information. Thanks to the recent success of Large Language Models (LLM) [4,20,25,38] trained on massive web data, we can manage to obtain high-level human knowledge from the text they generate. To this end, we designed two prompts (*i.e.*, global action prompt and joint motion prompt) to obtain both global and local joint-level descriptions of the action, as shown in Fig. 3. The global description of the specified action and its joint motion description can be obtained from the LLM by simply changing the 'action name' (*e.g.*, '[hand waving]' in the given example) and '[joint-list]' (*e.g.*, the defined 25 joints in the NTU RGB+D 60 dataset [31]) in the global action prompt and joint motion prompt. 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 two types of text will be subsequently used to guide the skeleton feature learning.

### 3.4 Cross-Modal Global-Lobal-Global Guidance

We propose dual guidance: (1) using global action description to guide the skeleton encoder focus on local informative joints (*i.e.*, global-to-local guidance), and (2) establishing non-local interactions between local joint-level text and skeleton features, to achieve a kind of local-to-global cross-modal guidance.

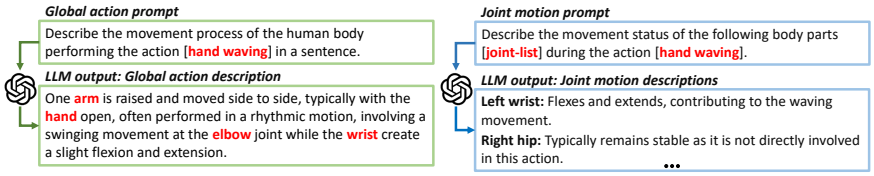


**Fig. 3:** The paradigm to acquire knowledgeable action description texts through a large language model (ChatGPT [4] is used in our implementation).

**Global-to-Local Textual Guidance Obtaining local informative joints from the global action description text.** As shown in Fig. 3, the acquired global action description indicates which areas are required to accomplish a particular movement from a global perspective (*i.e.*, the word in red color). Those informative regions are actually more essential to recognize the action. Therefore, our motivation is to extract those key areas from the global action description and guide the feature encoder to pay more attention to them. Particularly, we utilize the Noun Phrase Extraction module of Stanford CoreNLP [21] to extract the nouns in the global action description. Among these nouns, only joints and body parts such as 'arm' are left. Moreover, the body parts are converted to the corresponding joints in the human skeleton defined by the dataset used. For example, in the NTU RGB+D 120 dataset [19], arms correspond to shoulders, elbows, and wrists, and eyes correspond to the head. Thus, we can get the distribution of key joints from the global description of each action. Let $K \in \mathbb{R}^V$ be the distribution matrix of key joints of an action, where $K_j = 1$ if the $j$-th is a key joint of the action. Here, $V$ represents the number of joints.

**Joint Importance Determination Module (JID).** We design JID to let the skeleton encoder focus on the extracted informative joints. As shown in Fig. 2, the base skeleton encoder is based on InfoGCN [7], which contains an embedding block, $N$ encoding blocks, a projector, and a classifier. Each encoding block performs a spatial interaction and a temporal interaction on the skeleton features. Spatial interaction refers to the interaction of different joint features at the same time, while temporal interaction refers to the interaction of the same joint feature at different times. We divide the $N$ encoding blocks into two parts: The first part $F_{pre}$ has $N_{pre}$ blocks and the second part $F_{post}$ has $N_{post}$ blocks. After processing the input 3D skeleton sequence $X \in \mathbb{R}^{T_{in} \times V \times C_{in}}$ by the embedding block, we first feed it into $F_{pre}$ to obtain a skeleton feature $f_{pre} \in \mathbb{R}^{T_{pre} \times V \times C_{pre}}$ that gets processed with certain spatio-temporal information. Then we obtain a preliminary joint overall motion feature $\overline{f}_{pre} \in \mathbb{R}^{V \times C_{pre}}$ by pooling $f_{pre}$ along the time dimension. Here, $T_{in}$, $T_{pre}$ denote the sequence length. $C_{in}$, $C_{pre}$ denote the channel lengths of the input and $f_{pre}$. We feed the $\overline{f}_{pre}$ into the Joint Importance Discriminator (JID) module to obtain the importance of each joint motion feature, where JID is composed of two linear layers and a softmax layer. JID receives the $\overline{f}_{pre}$ and outputs the joint importance $k_{out} \in \mathbb{R}^V$ of $X$, where $k^i_{out}$ represents the importance of the $i$-th joint and it holds that $0 \leq k^i_{out} \leq 1$ for all $1 \leq i \leq V$. The generated $k_{out} \in \mathcal{R}^V$ is used to reweight the feature

$f_{post} \in \mathcal{R}^{T \times V \times C_{post}}$ as Eq. (1),

$$[f_{post}]_{ij}^{t} = k_{out}^{i} \cdot [f_{post}]_{ij}^{t}, t \in [1, T], i \in [1, V], j \in [1, C_{post}], \tag{1}$$

where i and j iterate over the joint and channel dimensions respectively. This can help the encoder focus more on local informative joints for effective action representation. To make JID produce more reasonable results, we design a calibration loss to align its output to the key joints distribution $k_{gt} \in \mathbb{R}^{V}$ that is derived from the global action text description:

$$\mathcal{L}_{calibrate} = MSE(k_{out}, k_{gt}), \tag{2}$$

where MSE is the standard MSE-loss. By aligning $k_{out}$ to $k_{gt}$, JID can benefit from the high-level knowledge from the global action description to output a more reasonable joint importance result.

**Local-to-Global Cross-Modal Interaction.** Although the skeleton modality provides simple and efficient information about the body, it may also suffer from sparse representation and potential noise [33]. If the model only focuses on the motion of local skeleton joints, it will be difficult to capture the high-level overall motion features, to summarize as an action-level representation. This may reduce the effectiveness of feature representation and hurt the generalization capability, especially under the one-shot setting. Therefore, we further design a local-to-global cross-modal interaction module $\phi$. Within it, non-local interaction between joint-level motion description and joint-level skeleton features is built to summarize the high-level global representation of the entire action.

Specifically, we use a pre-trained text encoder (*e.g.*, DeBerta [11] in our implementation) to process the joint-level motion descriptions (illustrated in Sec. 3.3) to get their feature embeddings $t \in \mathbb{R}^{V \times C_{txt}}$. We pass the skeleton sequence $X$ through the skeleton encoding blocks $F_{pre}$ and $F_{post}$ to get the feature $f_{post} \in \mathbb{R}^{T \times V \times C_{post}}$, and pool $f_{post}$ along the time dimension to get the overall joint features $\overline{f}_{post} \in \mathbb{R}^{V \times C_{post}}$. We project $t$ and $\overline{f}_{post}$ to a public feature space $P$ by 2 learnable MLPs, resulting in $p_{txt} \in \mathbb{R}^{V \times C_p}$ and $p_{ske} \in \mathbb{R}^{V \times C_p}$, respectively. Here, $C_{txt}$ and $C_{post}$ are respectively the channel lengths of the text feature space and skeleton space, and $C_p$ is the channel length of the public feature space. After that, we feed $p_{txt}$ and $p_{ske}$ into $\phi$ for interaction. $\phi$ has a total of M layers of interaction blocks, where $\phi^i$ is the $i$-th block in $\phi$ ($1 \leq i \leq M$). Each of these blocks has the same architecture. In each block $\phi^i$, we build non-local interactions in text-text, skeleton-skeleton, and skeleton-text perspectives.

First, We apply self-attention to $p_{txt}$, enabling each joint text feature to get information from others. As shown in Eq. (3), $p_{txt}$ is fused with the non-local semantic context information, resulting in $p_{txt}^{i} \in \mathbb{R}^{V \times C_p}$:

$$p_{txt}^{i} = MHA(Q, K, V = p_{txt}), \tag{3}$$

where MHA [36] is the multi-head attention.We then utilize text feature $p_{txt}^{i}$ to guide the integration of joint-level skeleton features $p_{st}^{i-1}$ to obtain $p_{st}^{i} \in \mathbb{R}^{V \times C_p}$

through a cross-attention module, where $p_{st}^0$ is $p_{ske}$, as shown in Eq. (4),

$$p_{st}^i = MHA(Q = p_{txt}^i, K = p_{st}^{i-1}, V = p_{st}^{i-1}), \tag{4}$$

After that, we sum up the text features $p_{txt}^i$ and skeleton features $p_{st}^i$ to fuse the knowledge of the two modalities via MLP processing, as shown in Eq. (5),

$$p_{st}^i = MLP(p_{txt}^i + p_{st}^i). \tag{5}$$

After going through the fusion operation of $M$-layer blocks, the output $p_{st}^M \in \mathbb{R}^{V \times C_p}$ of $\phi$ is with fine-grained high-level semantic information.

### 3.5 Dual-Branch Architecture

During inference, we actually can not acquire the text information (*e.g.*, action name) when the network is conducting classification on novel unseen action classes, which means the cross-modal guidance branch can not be used for testing. To solve the asymmetry issue between the training and inference phases, we design a dual-branch architecture that lets the features from both the skeleton encoding branch and the cross-modal guidance branch go through a shared classifier, for action classification during training. Such design not only lets features in the skeleton branch learn from the guidance branch during training, but also keeps those two-branch features consistent so that the guidance branch can be removed during inference. This echoes the concept of distilling complex network knowledge into a simpler architecture, similar to that seen in [1, 6, 12]. However, our design delves further, investigating the alignment of knowledge across modalities, thus enhancing cross-modal feature learning.

Specifically, after obtaining the fusion feature $p_{st}^M$ with both skeleton and textual knowledge, we project it back into the skeleton feature space and pool it along the joint dimension to obtain $f_{out}^c \in \mathbb{R}^{C_{post}}$ as the final output of the cross-modal guidance branch. Similarly, $f_{post} \in \mathbb{R}^{T \times V \times C_{post}}$ in the skeleton encoding branch is also projected and pooled to the same feature space, resulting in $f_{out}^s \in \mathbb{R}^{C_{post}}$. We take $f_{out}^s$ as the final output of the skeleton encoding branch. We pass the outputs $f_{out}^c$ and $f_{out}^s$ of the two branches separately through a shared MLP classifier with a softmax layer to obtain the normalized probabilities $\hat{y}_c$ and $\hat{y}_s$. After that, we use the standard cross-entropy loss to compute the classification losses $\mathcal{L}_c$ and $\mathcal{L}_s$ of the two branches, respectively,

$$\mathcal{L}_c = \mathcal{L}_{CE}(\hat{y}_c, y), \tag{6}$$

$$\mathcal{L}_s = \mathcal{L}_{CE}(\hat{y}_s, y), \tag{7}$$

where y is the action label. In view of optimization, the gradient of the cross-modal guidance branch could be back-propagated into the encoding blocks. Thus, the fine-grained high-level semantic information in the joint motion text is passed into the encoding blocks. The enhancement of the encoding blocks can be seen in Fig. 5. In the inference phase, we only need the skeleton encoding branch that already incorporates human knowledge for skeleton action recognition, without using any auxiliary text information. Such design also makes our model more efficient during inference, because only the skeleton encoder will be used.

### 3.6 Overall Learning Scheme

The overall training loss of the network is shown in Eq. (8), where $\alpha_1$ and $\alpha_2$ are the hyperparameters of $\mathcal{L}_{calibrate}$ and $\mathcal{L}_c$, respectively:

$$\mathcal{L}_{overall} = \mathcal{L}_s + \alpha_1 \mathcal{L}_{calibrate} + \alpha_2 \mathcal{L}_c. \tag{8}$$

At the end of the training, we freeze the skeleton encoding branch and use the distribution calibration method [41] for one-shot action classification.

## 4 Experiment

To demonstrate the advantages of CrossGLG, we perform one-shot skeleton action recognition on three large-scale datasets, NTU RGB+D 60 [31], NTU RGB+D 120 [19] and Kinetics [15]. We compare our model with current SOTA methods and conduct ablation studies to verify the effect of each component.

### 4.1 Datasets and Settings

**NTU RGB+D Dataset 60 (NTU 60)** [31] is a large-scale dataset, which contains 56, 578 videos with 60 action labels and 25 joints for each body, including interactions with individual and pairs activities. It has a total of 10 categories of novel actions in the one-shot setting.
**NTU RGB+D Dataset 120 (NTU 120)** [19] represents the most extensive dataset for action recognition, comprising 114, 480 videos annotated with 120 action labels. Recorded across diverse settings, involving 106 subjects and utilizing 32 distinct setups, the dataset includes a total of 20 categories of novel actions within the one-shot setting.
**Kinetics [15].** Besides the aforementioned two widely used benchmarks [19,31], we further conduct experiment on the challenging Kinetics dataset, with a fair 1-shot setting. It contains around 650, 000 video clips retrieved from YouTube. The videos cover as many as 400 human action classes, ranging from daily activities, sports scenes, to complex actions with interactions. Please see the appendix for its evaluation protocol under the 1-shot setting.
**Implementation details.** The skeleton encoding branch contains 9 encoding blocks with the same training setting as InfoGCN [7]. For the cross-modality guidance branch, 3 interaction blocks are used for fusing skeleton and texture features. The hyperparameters $\alpha_1$ and $\alpha_2$ illustrated in Sec. 3.6 are 0.5 and 0.2, respectively. DeBERTa-V2-Xlarge [11] is chosen as the text encoder. A single NVIDIA RTX 3090 GPU is used for training and testing.

### 4.2 Comparison with State-of-the-Art Methods

**NTU 120.** From Tab. 1, it can be seen that CrossGLG consistently outperforms the current SOTA methods in large margins, within all experimental settings with different numbers of base classes on NTU 120 [19]. CrossGLG also only

| #Base Classes | 20 | 40 | 60 | 80 | 100 | Para(M) |
|---|---|---|---|---|---|---|
| APSR [19] | 29.1 | 34.8 | 39.2 | 42.8 | 45.3 | - |
| uDTW [37] | 32.2 | 39.0 | 41.2 | 45.3 | 49.0 | - |
| SL-DML [23] | 36.7 | 42.4 | 49.0 | 46.4 | 50.9 | 11.8 |
| Skeleton-DML [22] | 28.6 | 37.5 | 48.6 | 48.0 | 54.2 | 11.8 |
| ALCA-GCN [45] | 38.7 | 46.6 | 51.0 | 53.7 | 57.6 | - |
| MotionBERT [46] | 35.5 | 54.3 | 56.5 | 52.8 | 61.0 | 60.3 |
| InfoGCN [7] | 37.0 | 53.9 | 58.8 | 55.7 | 56.1 | 1.6 |
| InfoGCN [7]+GAP [39] | 35.06 | 54.8 | 50.82 | 53.23 | 59.9 | 1.6 |
| InfoGCN [7]+CrossGLG | **45.3** | **56.8** | **62.1** | **61.6** | **62.6** | 1.7 |

**Table 1:** Accuracy (%) comparison on NTU120 [19].

| #Base Classes | 10 | 20 | 30 | 40 | 50 |
|---|---|---|---|---|---|
| uDTW [37] | 56.9 | 61.2 | 64.8 | 68.3 | 72.4 |
| MotionBERT [46] | **58.3** | 61.0 | 70.0 | 70.3 | 74.5 |
| InfoGCN [7] | 51.1 | 62.1 | 65.7 | 72.1 | 72.3 |
| InfoGCN+CrossGLG | 57.9 | **67.1** | **70.9** | **73.4** | **75.6** |

**Table 2:** Accuracy (%) comparison on NTU60 [31].

| #Base Classes | 20 | 40 |
|---|---|---|
| MotionBERT [46] | 13.2 | 16.6 |
| InfoGCN [7] | 13.3 | 18.2 |
| InfoGCN [7]+CrossGLG | **17.4** | **19.2** |

**Table 3:** Accuracy (%) comparison on Kinetics [15].

brings a neglectable cost (0.1M) compared with the based skeleton encoder [7], making it much more efficient than the previous SOTA MotionBERT [46] (only 2.8% of its model size) during inference. Besides, we also apply GAP [39], the current SOTA text-guided method designed for fully supervised setting, under one-shot learning setting. It can be seen that when generalizing to the one-shot learning setting, its performance is even lower than the base encoder InfoGCN [7] in some cases. Meanwhile, CrossGLG can consistently enhance the performance of the based skeleton encoder with neglectable cost, showing the strong superiority of the proposed method.

**NTU 60.** We also compare SOTA methods with CrossGLG on NTU 60 dataset [31]. As in Tab. 2, CrossGLG still outperforms the most SOTA method [46] in the vast majority of settings. Although our method exhibits a 0.4% lower performance compared to MotionBERT [46] in the setting of 10 base classes, we attribute this to the relatively weaker base encoder [7] we employed, and as shown in the bottom two lines of Tab. 2, CrossGLG actually brings a significant performance improvement (6.8%) to the base encoder, showing its effectiveness.

**Kinetics.** We also compare SOTA methods with CrossGLG on the Kinetics dataset [15]. From Tab. 3, it can be observe that even in more complex and challenging environments, our method still outperforms the current SOTA method [46]. This further illustrates the superiority and generalization ability of our method in skeleton-based one-shot action recognition.

### 4.3 Ablation Studies

We conduct ablation studies on NTU120 [19] to evaluate the components.

**Analysis of global-to-local textual guidance (G2L).** The result is shown in Tab. 4. It can be observed that G2L can consistently improve one-shot action

| G2L | L2G | #Base classes | | | | |
|---|---|---|---|---|---|---|
| | | 20 | 40 | 60 | 80 | 100 |
| × | × | 37 | 53.9 | 58.8 | 55.7 | 56.1 |
| ✓ | × | 43.3 | 54.7 | 60.9 | 58.7 | 61.7 |
| × | ✓ | 42.5 | 54.9 | 61.7 | 60.1 | 58.6 |
| ✓ | ✓ | **45.3** | **56.8** | **62.1** | **61.6** | **62.6** |

**Table 4:** Comparison of performance with vs. without proposed components.

| $N_{pre}$ | #Base Classes | | | | |
|---|---|---|---|---|---|
| | 20 | 40 | 60 | 80 | 100 |
| 3 | 41.1 | 50.8 | **61.9** | 54.5 | 57.7 |
| 5 | **43.3** | **54.7** | 60.9 | **58.7** | **61.7** |
| 7 | 43.2 | 54.1 | 58.7 | 58.5 | 58.0 |
| 9 | 38.7 | 51.8 | 54.4 | 57.3 | 55.3 |

**Table 5:** Performance Analysis of Insertion Position $N_{pre}$ for JID Module .

| Text-Text interaction | #Base classes | | | | |
|---|---|---|---|---|---|
| | 20 | 40 | 60 | 80 | 100 |
| × | 41.7 | 53.8 | 59.5 | 49.6 | 55.9 |
| ✓ | **42.5** | **54.9** | **61.7** | **60.1** | **58.6** |

**Table 6:** Analysis of text self-attention module's role in joint skeleton-textual features interaction (*phi*) on NTU 120 [19].

| LLM | Acc(%) |
|---|---|
| ChatGPT [4] | 45.3 |
| Gemini [34] | 45.2 |
| Qwen [2] | 45.7 |

**Table 7:** Robustness of CrossGLG under varying LLM on NTU 120 [19], 20 base classes.

recognition across all settings, which essentially demonstrates the effectiveness of the proposed global-to-local textual guidance. We also experimented with the insertion location of the proposed JID module. The encoder has a total of 9 encoding blocks. JID takes the output of the $N_{pre}$-th block as input. For example, when $N_{pre}$ is 9, it means that the output of JID is directly applied to the output of the 9-th layer. The results of our experiments for the setting of $N_{pre}$ are shown in Tab. 5. It can be seen that the overall classification accuracy of the model is best when the output of the 5-th block is fed to JID. Our analysis is that when the insertion position of the JID is too forward, the skeleton features fed into the JID are not sufficiently fused with spatio-temporal information, and it is difficult for the JID to effectively determine the importance of each joint through such shallow features. On the other hand, when the insertion position is too far back, the subsequent blocks that can be affected by the JID are too few so it is difficult to influence the feature representations.

**Analysis of local-to-global cross-modal interaction (L2G).** As can be observed from Tab. 4, when L2G is added, the classification accuracy under different experimental settings is also improved in all cases. This proves that the proposed L2G successfully conveys rich high-level semantic context to guide the skeleton feature learning. We also conducted experiments on the text-text interaction within the interaction block $\phi$. As shown in Tab. 6, it brings consistent performance gain across all settings. We argue this is because it can summarize non-local semantics to facilitate the subsequent skeleton-text interaction.

**Robustness to LLMs.** Here we evaluate the robustness of CrossGLG when using different LLM for text generation. As shown in Tab. 7, CrossGLG is robust with different LLM (performance difference less that 0.5%).

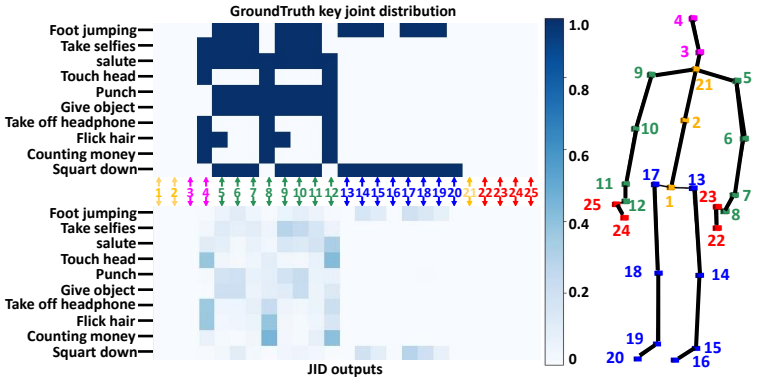


**Fig. 4:** Visualization of the output joint importance from JID. The upper half of the distribution of Ground Truth key joints is extracted from the global action description. The bottom half is the output of the Joint Importance Discrimination (JID) module.

## 5 Visual Analysis

We conducted experiments on NTU120 [19] (20-base) to visualize the intermediate results of the network to further analyze the effectiveness of CrossGLG.

**Visualization of the output joint importance from JID.** We visualize the output of JID (the bottom part of Fig. 4) and compare it with the ground truth key joint distribution (the upper part of Fig. 4). It can be seen that JID successfully learned to predict joint importance which is similar to the GT key joint distribution. At the same time, JID does not simply fit the GT distribution, but also adjusts the weights appropriately according to the observed skeleton features. For example, the GT key joint of 'Squat down' derived from the global action description contains joints around both leg and arm. However, the movement of arm is actually not strongly correlated with the action of squatting down. This is also learned by JID, as it produces low weights on joints around arms, which further demonstrates its capacity for high-level semantic understanding.

**Attention visualization of the skeleton encoder.** We visualize the attention of the skeleton encoder, which is shown in Fig. 5. In each visualization matrix, each row represents the spatial attention of each joint in an action. The color assigned to each joint on the right side corresponds to the color representing its respective position at the bottom of the matrix. It can be seen that G2L, L2G, and their combination significantly improve the spatial clue-grasping ability of the model. For example, compared to the baseline, they improved the attention for the arm and finger parts in both 'touch head' and 'arm swings'. In the actions 'sit down' and 'squat down', there is more spatial attention to the leg joints.

We also perform a visual analysis of the spatial attention of the novel class actions. 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, and the results are shown in Fig. 6. It can be seen that compared to the baseline, our method pays more attention to the hand features in the actions 'Take off glasses'

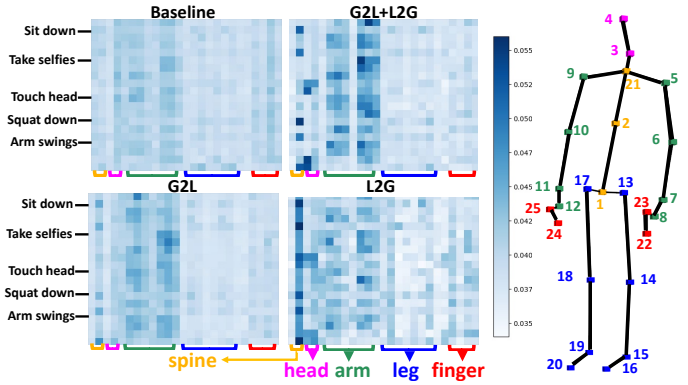


**Fig. 5:** Visual analysis: Spatial attention in skeleton encoder. Matrix rows depict importance of encoding blocks for joint features; darker colors indicate higher significance.

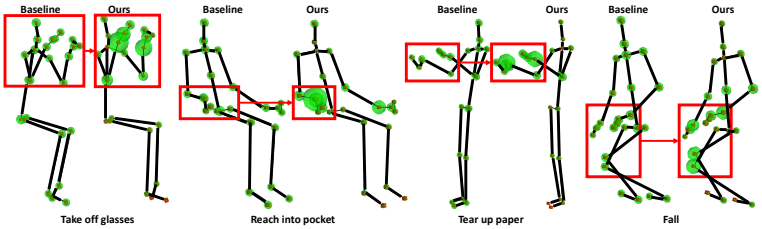


**Fig. 6:** Spatial attention visualization analysis on novel class actions. The spheres around each joint indicate how much attention the model pays to it. We draw the 'Take off glasses', 'Peach into pocket', 'Tear up paper' and 'Fall' actions for illustration.

and 'Tear up paper'. In the action 'reach into pocket', our method pays more attention to the right hand that reaches into the pocket. In the 'Fall' action, our model not only focuses on key joints such as the knees and hands but also pays attention to areas like the spine, as the fall is an action relevant to the whole body. The ability of key spatial information capturing in unseen action classes demonstrates the strong generalization of our method.

## 6 Conclusion

We propose a novel cross-modal guidance framework CrossGLG that leverages human-knowledge-rich action descriptions from LLM to guide the skeleton feature learning, in a global-local-global way. Specifically, the global-to-local guidance enables the skeleton encoder to focus more on important local details. The local-to-global guidance establishes non-local interactions between joint-level motion descriptions and skeleton features to guide the generation of global action representations using fine-grained high-level semantic information. The dual-branch architecture could mitigate the asymmetry issue between the training and inference phases. Experiments verify the superiority of CrossGLG.

**Acknowledgments.** This work is jointly supported by the National Natural Science Foundation of China (Grant No. 62271221 and 62371219). Joey Tianyi Zhou is supported by the National Research Foundation, Prime Minister's Office, Singapore, and the Ministry of Communications and Information, under its Online Trust and Safety (OTS) Research Programme (MCI-OTS-001) and SERC Central Research Fund(Use-inspired Basic Research). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, or the Ministry of Communications and Information.

## References

1. Aguilar, G., Ling, Y., Zhang, Y., Yao, B., Guo, C.: Knowledge distillation from internal representations. pp. 7350–7357 (2020)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Blake, R., Shiffrar, M.: Perception of human motion. Annu. Rev. Psychol. **58**, 47–73 (2007)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Cai, J., Jiang, N., Han, X., Jia, K., Lu, J.: Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2735–2744 (2021)
6. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11923–11932 (2022). https://doi.org/10.1109/CVPR52688.2022.01163
7. 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)
8. Dai, D., Sun, Y., Dong, L., Hao, Y., Sui, Z., Wei, F.: Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. arXiv preprint arXiv:2212.10559 (2022)
9. Guo, M., Chou, E., Huang, D.A., Song, S., Yeung, S., Fei-Fei, L.: Neural graph matching networks for fewshot 3d action recognition. In: Proceedings of the European conference on computer vision. pp. 653–669 (2018)
10. Hadad, B., Schwartz, S., Maurer, D., Lewis, T.L.: Motion perception: a review of developmental changes and the role of early visual experience. Frontiers in integrative neuroscience **9**, 49 (2015)
11. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)
12. Hou, Z., Yu, B., Tao, D.: Batchformer: Learning to explore sample relationships for robust representation learning. In: CVPR (2022)
13. Jing, Y., Wang, F.: Tp-vit: A two-pathway vision transformer for video action recognition. In: 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2185–2189. IEEE (2022)

14. Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D.: Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Transactions on Systems, Man, and Cybernetics: Systems **49**(9), 1806–1819 (2018)
15. 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)
16. Knox, W.B., Stone, P.: Augmenting reinforcement learning with human feedback. In: ICML 2011 Workshop on New Developments in Imitation Learning (July 2011). vol. 855, p. 3 (2011)
17. 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)
18. Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., Tian, Q.: Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3595–3603 (2019)
19. 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)
20. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023)
21. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. pp. 55–60 (2014)
22. Memmesheimer, R., Häring, S., Theisen, N., Paulus, D.: Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3702–3710 (2022)
23. Memmesheimer, R., Theisen, N., Paulus, D.: Sl-dml: Signal level deep metric learning for multimodal one-shot action recognition. In: 2020 25th International conference on pattern recognition. pp. 4573–4580. IEEE (2021)
24. OpenAI: Introducing chatgpt. Website (2022), `https://openai.com/index/chatgpt`
25. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
26. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022)
27. Plizzari, C., Cannici, M., Matteucci, M.: Spatial temporal transformer network for skeleton-based action recognition. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III. pp. 694–701. Springer (2021)
28. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

29. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
30. Rani, S.S., Naidu, G.A., Shree, V.U.: Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. Materials Today: Proceedings **37**, 3164–3173 (2021)
31. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
32. Song, S., Liu, J., Li, Y., Guo, Z.: Modality compensation network: Cross-modal adaptation for action recognition. IEEE Transactions on Image Processing **29**, 3957–3969 (2020)
33. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence (2022)
34. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
35. Thoker, F.M., Gall, J.: Cross-modal knowledge distillation for action recognition. In: 2019 IEEE International Conference on Image Processing. pp. 6–10. IEEE (2019)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wang, L., Koniusz, P.: Uncertainty-dtw for time series and sequences. In: European Conference on Computer Vision. pp. 176–195. Springer (2022)
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
39. 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)
40. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
41. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. In: International Conference on Learning Representations (2021)
42. Yang, S., Liu, J., Lu, S., Hwa, E.M., Kot, A.C.: One-shot action recognition via multi-scale spatial-temporal skeleton matching. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
43. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia **19**(2), 4–10 (2012)
44. Zhao, R., Ali, H., Van der Smagt, P.: Two-stream rnn/cnn for action recognition in 3d videos. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4260–4267. IEEE (2017)
45. Zhu, A., Ke, Q., Gong, M., Bailey, J.: Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6038–6047 (2023)

46. 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)
47. Zou, Y., Shi, Y., Shi, D., Wang, Y., Liang, Y., Tian, Y.: Adaptation-oriented feature projection for one-shot action recognition. IEEE Transactions on Multimedia **22**(12), 3166–3179 (2020)