

SUGAR: Learning Skeleton Representation with Visual-Motion Knowledge for Action Recognition

Qilang Ye^{1 2}, Yu Zhou^{1 2*}, Lian He^{2 5}, Jie Zhang³, Xuanming Guo², Jiayu Zhang³, Mingkui Tan⁶, Weicheng Xie⁷, Yue Sun⁸, Tao Tan⁸, Xiaochen Yuan⁸, Ghada Khoriba⁹, Zitong Yu^{3 4*}

¹VCIP & TMCC & DISec, College of Computer Science & College of Cryptology and Cyber Science, Nankai University

²Zhongguancun Academy

³Great Bay University

⁴Dongguan Key Laboratory for Intelligence and Information Technology

⁵School of Computer Science & Technology, Beijing Institute of Technology

⁶South China University of Technology

⁷Shenzhen University

⁸Macao Polytechnic University

⁹Nile University

yql@mail.nankai.edu.cn

Abstract

Large Language Models (LLMs) hold rich implicit knowledge and powerful transferability. In this paper, we explore the combination of LLMs with the human skeleton to perform action classification and description. However, when treating LLM as a recognizer, two questions arise: 1) How can LLMs understand skeleton? 2) How can LLMs distinguish among actions? To address these problems, we introduce a novel paradigm named learning **S**keleton representation with **vis**Ual-**U**al-motion **know**ledge for **A**ction **R**ecognition (SUGAR). In our pipeline, we first utilize off-the-shelf large-scale video models as a knowledge base to generate visual, motion information related to actions. Then, we propose to supervise skeleton learning through this prior knowledge to yield discrete representations. Finally, we use the LLM with untouched pre-training weights to understand these representations and generate the desired action targets and descriptions. Notably, we present a Temporal Query Projection (TQP) module to continuously model the skeleton signals with long sequences. Experiments on several skeleton-based action classification benchmarks demonstrate the efficacy of our SUGAR. Moreover, experiments on zero-shot show that SUGAR is more versatile than linear-based methods.

Introduction

Skeleton-based action recognition (Shi et al. 2019a,b; Chen et al. 2021; Cai et al. 2023; Cheng et al. 2020) aims to model the spatio-temporal graph structure (Yan, Xiong, and Lin 2018) to classify actions. As representations of human behavior, skeletons satisfy practical applications such as human-computer interaction and intelligent monitoring

*Corresponding Authors (yzhou@nankai.edu.cn, yuzitong@gbu.edu.cn)

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

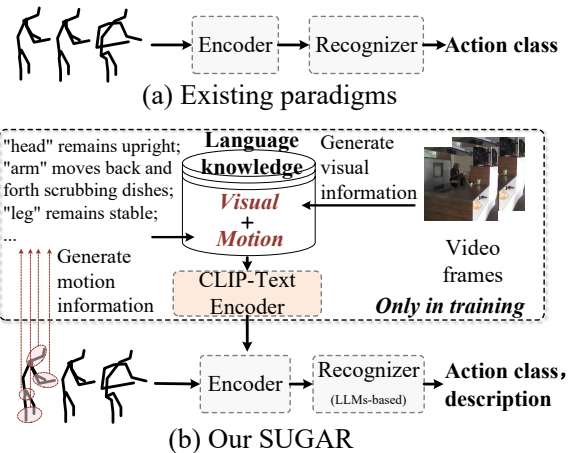


Figure 1: Comparison between existing action recognition paradigms and our SUGAR.

with lightweight data storage. Although training with skeleton data (e.g. human body joints) brings computational efficiency (Cheng et al. 2021; Noor and Park 2023), building an efficient recognizer for most activities of daily living is challenging and requires distinguishing similar fine-grained actions (Das et al. 2019). For example, the movement trajectories of drinking water and snacking are extremely similar, leading the recognizer to easily categorize them as one.

Recently, Large Language Models (LLMs) such as Vicuna (Chiang et al. 2023) and LLaMA (Touvron et al. 2023) have greatly influenced different domains. They can be effectively applied to many tasks that are not limited to plain text, but also to visual as well as audio tasks (Zhang, Li, and Bing 2023; Huang et al. 2023; Li et al. 2023b). A great deal of work has started to leverage the generated rich linguistic knowledge to improve their subtasks. LLMs can gener-

ate relevant information to help the corresponding network to learn linguistic knowledge and improve reasoning ability. Moreover, people have found that LLMs can be used not only as interpreters for text generation, but also as recognizers or classifiers to optimize downstream tasks (Ye et al. 2024). Specifically, fine-tuning on a large-scale multimodal corpus, LLMs can then adapt to different modal inputs.

But *what if the large language model served as an action recognizer?* Qu et al. (Qu, Cai, and Liu 2024) have given us answers. LLM contains rich human-centric knowledge after pre-training over a large corpus, it holds strong power to learn the input skeleton signal. Different from traditional linear-based methods (Yan, Xiong, and Lin 2018; Shi et al. 2019b), LLMs are not only capable of classifying but also describing actions. However, several problems remain to be explored: **1) How can LLMs understand skeleton?** LLMs can model language sequences and targets, but remain a challenge with input from other “non-human languages”. There is a gap between the skeleton and the text, and a lack of a universal encoder to align the two. Previous works (Qu, Cai, and Liu 2024; Zhou, Liu, and Wang 2023; Zhang et al. 2023; Wang, Boddeti, and Lim 2024) introduce an action-based vector quantized variational autoencoder (Van Den Oord, Vinyals et al. 2017) to learn discrete tokens, it provides a solution to transfer the skeleton into the LLM. However, LLMs are mostly pre-training over human language datasets, so it is difficult to ensure that the skeleton tokens are consistent with text tokens. **2) How can LLMs distinguish between similar skeleton signals?** As we mentioned, daily activities are filled with many similar actions. When the appearance information is missing, it is difficult to distinguish whether we are drinking or eating.

To solve the above problems, we aim to make similar actions in daily activities towards more discrete representations, and to make LLMs understand such representations for downstream tasks. Therefore, we propose a novel paradigm named learning **Skeleton representation with visual-motion knowledge for Action Recognition (SUGAR)**. As shown in Fig. 1, most of the existing paradigms (Shi et al. 2019b; Yan, Xiong, and Lin 2018; Chen et al. 2021), including LLMs-based method (Qu, Cai, and Liu 2024), only emphasize how to design powerful recognizers to achieve excellent performance and then output the action targets. In contrast, we develop a new training scheme, which leverages generated visual and motion information to supervise the learning process in skeletons. We believe that this a priori knowledge is crucial for LLMs as a recognizer.

Specifically, our SUGAR mainly targets the mentioned problems **1)** and **2)**. To make LLMs understand the skeleton input, we first define a series of action lists and fine-tune the LLMs with a fixed instruction. Specifically, we leverage the pre-trained skeleton encoder to generate representations, and then perform low-rank adaptation (LoRA) (Hu et al. 2022) on the LLMs to make the model understand the input skeleton representations. Notably, we design a Temporal Query Projection (TQP) module to shorten the input representation sequence and maintain continuous temporal modeling.

To better distinguish similar fine-grained actions, we pro-

pose to introduce motion and visual information¹. During movement, each part of each action has a different trajectory, and describing this trajectory can lead to greater variability between different movements. Moreover, such high-level semantic information can provide fine-grained prior knowledge for skeleton representation learning. However, when motion trajectories are exactly similar (e.g., eating and drinking), motion information alone is not sufficient. Therefore, we use an off-the-shelf Visual Language Model (VLM) (OpenAI 2023) to automate the synthesis of visual information related to actions. Finally, we store visual and motion information as text and learn discrete skeleton representations in a way that mimics the CLIP-based pre-training of Image-Text (Radford et al. 2021). In this way, LLMs can learn a skeleton representation that is aligned with the corresponding description.

Related Work

Unimodal and Multimodal Action Recognition. Unimodal methods for action recognition can be categorized into RGB-based and skeleton-based (joint+bone), both of which perform inference from a single modality of information. RGB-based methods (Ye and Yu 2024; Nan et al. 2024; Rajendran et al. 2024; Li et al. 2025) mainly deal with a sequence of image information, and are good at capturing rich visual contexts. However, they usually suffer from excessive parameters and overfitting. On the other hand, the skeleton-based recognition performance has made great progress from the ST-GCN model proposed by Yan et al. (Yan, Xiong, and Lin 2018). Afterwards, models like 2s-AGCN (Shi et al. 2019b) and CTR-GCN (Chen et al. 2021) use dual-stream to dynamically learn the topology and achieve higher accuracy.

The combination of skeleton and visual can enhance the recognition of actions. Fabien et al. (Baradel, Wolf, and Mille 2017) proposed a novel multimodal approach that focuses on the hand pose. This method opens up new possibilities for future work but it ignores some details of the rest of the body. Some works (Das et al. 2019) attempt to weight fusion of RGB images by utilizing pose data, which is an innovative way of integrating different types of modality. Other works (Das et al. 2020; Bruce et al. 2022) introduced a spatial embedding to map pose data to visual features. (Das et al. 2021) developed a feature-level and attention-level distillation, which offers a practical solution for combining RGB and pose. However, the fusion of multimodal inputs requires massive computation. Moreover, we have no way to know whether heterogeneous visualizations and skeletons can be effectively combined in the learning process.

Vision-Language Models for Action Recognition. Besides the contrastive learning of images and language, some work has started to apply representation learning to the action recognition domain. Wang et al. (Wang, Xing, and Liu 2021) follow the CLIP training strategy to help with downstream action recognition tasks. They convert action labels into representations and perform representation learning by calcu-

¹Motion and visual information refer to the text knowledge extracted from the motion skeleton and the video, respectively.

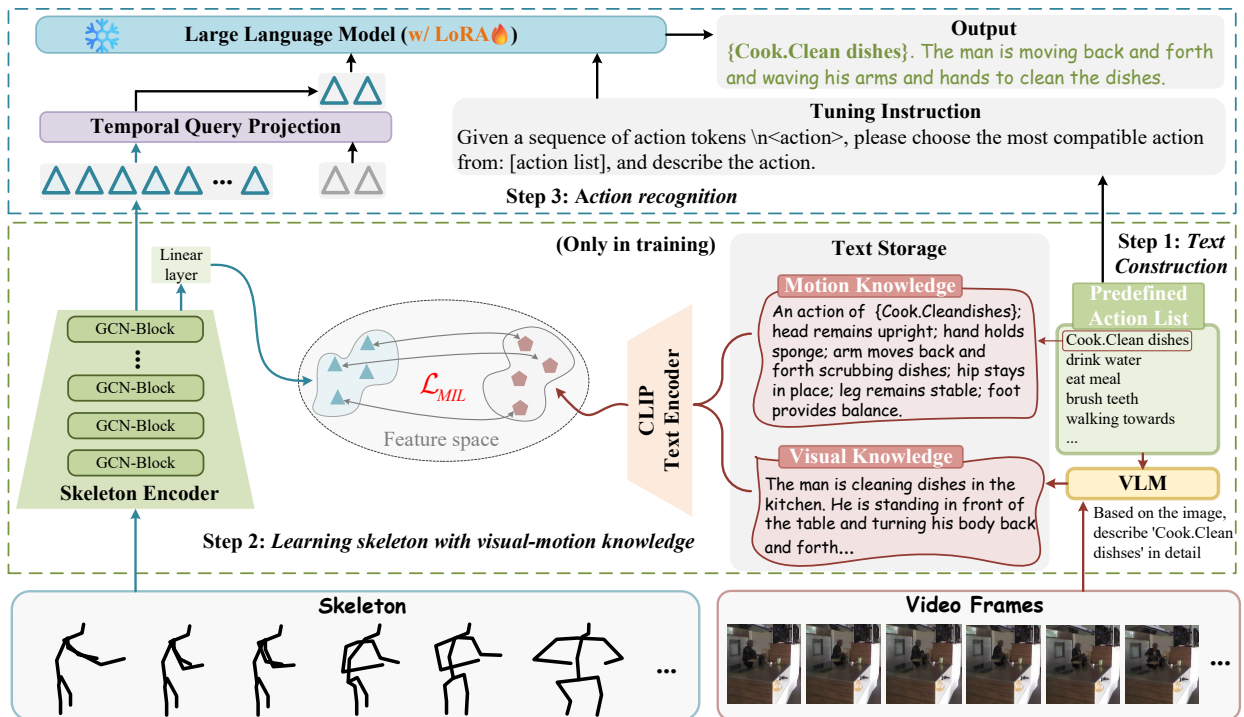


Figure 2: Overall framework of SUGAR. The complete training procedure is divided into three parts. We use the GPT-generated fine-grained action description and VLM-generated visual description as input for the text encoder to supervise the skeleton representation learning, where the linear layer maps the skeleton to the same feature space as the text. During inference, only the skeleton data needs to be input for action recognition.

lating the similarity with the video. However, we argue that labels like “A *action of* { }” lack substantial semantic information and do not provide exhaustive knowledge. Other efforts (Xie et al. 2024; Lin et al. 2025) to introduce representation learning into skeleton still do not provide rich linguistic knowledge to aid learning. To learn a more discrete representation, we introduce linguistic knowledge such as motion and vision to improve the performance.

Large Language Models for Action Recognition. A variety of studies (Ye et al. 2024, 2025; Li et al. 2023b) have shown that the powerful generalization ability of LLMs can help different downstream task reasoning. With the help of prompt learning (Zhou et al. 2022), LLMs can generate any form of text with any content. Recently, some work has started to focus on how to apply LLMs to action-centric tasks. Qu et al. (Qu, Cai, and Liu 2024) use VQ-VAE (Van Den Oord, Vinyals et al. 2017) to learn specific action tokens and fine-tune large models using LoRA (Hu et al. 2022). Drawing on their ideas, we find that utilizing rich linguistic knowledge to learn skeleton representations can bring better performance to the recogniser and performs well in zero-shot scenarios.

Methods

Overview of SUGAR

As illustrated in Fig. 2, the workflow of SUGAR starts with constructing sets of texts related to action and vision. We

collect a detailed description of each action and related visual information through a predefined list of actions (a dictionary that contains possible action labels). To avoid manual annotation, we have downloaded the powerful visual language model and GPT as generative tools. Then, we aim to pre-train a robust skeleton encoder that can be aligned to the text. Given a skeleton sequence as the input, the goal of the encoder is to make LLMs learn a more discrete skeleton representation. After that, we define a fixed instruction and fine-tune the LLM to understand the input skeleton representation. During inference, we only need to input the skeleton sequence to perform the action recognition task.

Step 1: Text Construction

Predefined Action List. We usually expect an ideal recognizer to be generalizable to different practical scenarios. Although existing LLMs are capable of open-ended answers, it remains a challenge to answer specific classes in the field of action recognition accurately. Therefore, we define a reasonable action dictionary, which includes all possible category names of the action dataset we used. Furthermore, we test our SUGAR in Sec. to show that: without fine-tuning, the model is able to achieve zero-shot inference with the predefined action list.

Generate Motion Knowledge. We find that each action has a corresponding unique description, such information can be used as vectors to supervise the learning of each action set. However, it is difficult to make an action specific when us-

ing instance-level description. Therefore, we utilize a fine-grained human body part pattern to represent an action. For example, an action of “Drink” can be represented as the head tilting back slightly, the hand grasping the cup, and so on.

Specifically, we follow the production process of the HAKE dataset (Li et al. 2020) and resolve each action into six body part movements: head, hand, arm, hip, leg, and foot. Moreover, to avoid manual annotation, we use the large language model GPT 3.5 turbo for action text generation. Given an action label from predefined action lists, we design the following prompts to generate motion descriptions \mathcal{T}_m :

Below are some examples to generate a short description.\n\n
 Example: A action of {drink water}: head tilts back slightly; hand grasps cup; arm lifts cup to mouth; hip remains stationary; leg remains stationary; foot remains stationary.\n\n
 Describing following body part movements when '{action}': head, hand, arm, hip, leg, foot."

With our human-written examples, the large-scale model can generalize them and produce the desired action description without any parameter updates.

Generate Action-related Visual Knowledge. We employ a powerful off-the-shelf vision-language model GPT-4V (OpenAI 2023) for generating action-related visual information. However, LLMs’ responses often contain redundancy and nonsense, which is not conducive to represent a particular class. For example, the VLM is inclined to generate a person’s appearance characteristics, the environment, etc. Therefore, we establish three rules to limit the generation of VLMs: - *Provides a description of the given image that correctly matches the action*, - *Only describe the scene related to the given action*, - *Do not provide any text or explanations unrelated to the action*.

Indoor activities generally involve long static movements and repetitive frames. To filter out the large number of duplicates in the generated frame descriptions, we transform each frame via the vision encoder from CLIP into features and collect the most dissimilar image set by counting the similarity scores between them. Then we perform text generation for each frame from the collected image set and obtain the visual description set \mathcal{T}_v .

Step 2: Skeleton Representation Learning

Skeleton Encoder. The skeleton is stored as coordinates, it constructs complex and diverse graphs with joint connections. Graph Convolution Network (GCN) (Kipf and Welling 2017) can be applied to graph signal processing due to its ability to aggregate node information. In this paper, we stack multiple GCN blocks as a skeleton encoder, and each block performs information aggregation on the input to learn the representation. Specifically, the input skeleton can be denoted as $G = \{V, \mathcal{E}\}$, a graph that contains V human joints and \mathcal{E} is the set of edges. We define the aggregated features of the input skeleton at layer l as $H^l \in \mathbb{R}^{D \times F}$, the F denotes the feature dimension. The graph convolution can be represented as:

$$H^{l+1} = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^lW^l), \quad (1)$$

where $D \in \mathbb{R}^{D \times D}$ is the degree matrix, A is the adjacency matrix of graph G , W^l is the l -layer’s weight parameter, and σ is the activation function.

Besides spatial aggregation, we leverage the multi-scale temporal modeling module designed by (Chen et al. 2021) to model the action in the temporal dimension. However, Maxpool in the original method loses fine-grained temporal information. We note that skeleton temporal sequences are critical for fine-grained action inference. Therefore, we discard the pooling operation in the time dimension and keep the complete time information for the next step.

Text Encoder. We employ a pre-trained CLIP-based (Radford et al. 2021) text encoder $E_t(\cdot)$ to transfer the collected information into embedding. Notably, we consider that motion description and visual description belong to two different categories, so we divide the output into two parts as follows:

$$m = E_t(\mathcal{T}_m), \quad v_i = E_t(\mathcal{T}_{v_i}), \quad (2)$$

where $i \in I_v$, I_v are indices of the collected visual knowledge. To diversify the collected linguistic knowledge, we randomly combine the motion description embedding and the visual description embedding to obtain $\mathbf{t} = \{m, v_i | i \in I_v\}$.

Contrastive Learning for Skeleton and Text. Contrary to one-to-one contrastive learning, we emphasize that the skeleton representation can be positively matched to multiple texts. Inspired by (Miech et al. 2020) that combines Multiple Instance Learning (MIL) and Noise Contrastive Estimation for contrastive learning, we further propose to encourage the learning of skeleton representations under multiple texts supervision. Specifically, the optimal goal is to contrast skeleton-text within the batch:

$$\mathcal{L}_{MIL} = -\frac{1}{|B|} \sum_i \log \frac{\sum_j \sum_n \exp(\mathbf{s}_i^\top \mathbf{t}_{j,n} / \tau)}{\sum_k \sum_n \exp(\mathbf{s}_i^\top \mathbf{t}_{k,n} / \tau)}, \quad (3)$$

where \mathbf{s} is the encoded feature of the skeleton, B is the batch size, $i, j, k \in B$, n is the index of the corresponding visual-motion description embedding in \mathbf{t} . τ is the temperature parameter.

Step 3: Action Recognition

Temporal Query Projection. To map the skeleton representation into the LLMs’ embedding space, we have to design a multimodal projector that converts the embedding inputs into language tokens of a suitable length. Earlier projectors for processing long sequence embeddings (Luo et al. 2023; Li et al. 2023c) usually perform temporal compression or employ a linear layer to map to the corresponding dimension. However, we consider that the temporal signals of the skeleton are constructed from a continuous series of positional information, and that pooling or arbitrary extraction would disrupt the context of the entire topology. Therefore, we introduce a novel projection method named Temporal Query Projection (TQP), which can continuously query temporal signal representations and distill them into short language tokens.

As shown in Fig. 3, let $\mathbf{s} \in \mathbb{R}^{L_s \times d}$ be the skeleton representation of length L_s . We introduce the Q-Former function f_Q (Li et al. 2023a) to replace the previous direct

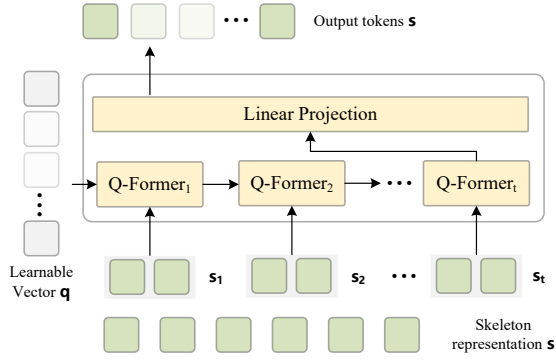


Figure 3: Temporal Query Projection consists of a number of Q-Formers (Li et al. 2023a) and a linear layer.

projection and cross-attention. Notably, we set a hyperparameter k to select t time points for a skeleton segment $\{s_i | i = 1, 2, \dots, t\}$, where $t = \frac{L_s}{k}$. Then we customize a learnable query vector $\mathbf{q} \in \mathbb{R}^{k \times d}$ in length k . Contrary to a single Q-Former, we utilize the previously queried skeleton representation as the next query in order to continuously model the entire skeleton signal, which can be denoted as:

$$\hat{s}_t = f_Q^t(s_{t-1}, s_t), \quad (4)$$

where f_Q^t denotes t -th f_Q , $\hat{s}_t \in \mathbb{R}^{L \times d}$, and all the parameters of Q-Former are shared.

Training Strategy. The training of SUGAR is mainly divided into two stages. The first stage is training a skeleton encoder supervised by visual-motion knowledge with the learning objective \mathcal{L}_{MIL} . The second stage is to fine-tune the LLMs with LoRA to adapt the non-human language tokens with the learning objective \mathcal{L}_{Lora} . Specifically, we define the LLM recognizer as $f_{LLM}()$, during the fine-tuning phase, we input the instructions shown in Fig. 2 and the action tokens \hat{s} after TQP into the LLM to predict the action class. \mathcal{L}_{Lora} can be expressed as:

$$\mathcal{L}_{Lora} = CrossEntropy(f_{LLM}(\hat{s}), y), \quad (5)$$

where we use cross-entropy loss to supervise the fine-tuning of the LLM, and y are tokens containing the corresponding ground truth and a description of the action. Furthermore, to make the output of the LLM more than just a single category, we collect a small number of brief descriptions of actions as instructions to fine-tune the model, where we use GPT to generate such brief descriptions based on the collected motion knowledge.

Inference. During inference, we only need to input the skeleton signal, and then use the pre-trained language-enhanced skeleton encoder to transform them into a discrete skeleton representation. After the projection process with TQP, LLMs predict the most appropriate one based on the predefined list of actions and output a brief description of the action.

Experiments

Datasets

Toyota Smarthome (SH) (Das et al. 2019) dataset contains 31 activities and generates a total of 16,115 video samples. The official proposal to get the skeleton sequence by (Yang et al. 2021), and the dataset provide three evaluation protocols: cross-subject ($X-sub$) and cross-view ($X-view1$ and $X-view2$). Results on Toyota Smarthome are mean per-class accuracy.

NTU RGB+D (NTU 60) (Shahroudy et al. 2016) contains 60 action categories with 56,880 video samples, and it provides the original skeleton sequences. This dataset provides two evaluation protocols: cross-subject ($X-sub$) and cross-view ($X-view$).

NTU RGB+D (NTU 120) (Liu et al. 2020) is an upgraded version of NTU 60, it contains more than 114k skeleton sequences and it has 120 action categories with 57,600 video samples. This dataset provides two evaluation protocols: cross-subject ($X-sub$) and cross-view ($X-view$).

PKU-MMD (PKU) (Liu et al. 2017) contains a total of 21,545 samples with skeleton data and 51 action categories. This dataset provides two evaluation protocols: cross-subject ($X-sub$) and cross-view ($X-view$).

Implementation Details

For Step 1, we use the labels of the corresponding dataset as the predefined action list, and we use GPT 3.5 turbo for action text generation and GPT-4V (OpenAI 2023) for action-related visual text generation. For Step 2, the skeleton data is pre-processed by the code of (Zhang et al. 2020). We use CTR-GCN (Chen et al. 2021) as the skeleton encoder backbone and CLIP (Radford et al. 2021) as the text encoder. Then, we train the skeleton encoder using the SGD optimizer with an initial learning rate of 0.01 for a total number of 200 epochs with batch size 200 (100 for Toyota Smarthome) and reduced by a factor of 0.1. For Step 3, we freeze all parameters of the skeleton encoder and we use the LLaMA2 7B (Touvron et al. 2023) as the LLM. We set $r = 64$ and $\alpha = 16$ for the LoRA parameters, and the total batch size is set to 128 for training 1 epoch with a learning rate of $2e^{-5}$. All experiments are conducted on two NVIDIA A6000 GPUs.

Comparisons to the State of the Arts

We compare our SUGAR with previous linear-based methods and LLMs-based methods in Table 1. Different from most methods that use the ensemble strategies (Joint with Motion, Bone with Motion, Joint with Bone), we train the entire framework using only the joints of the skeleton data as inputs, in a way that drastically reduces the computational cost. It is clear that our SUGAR dramatically outperforms all state-of-the-art results of linear-based classification methods in the Toyota Smarthome dataset (Das et al. 2019). Comparing to the same type of LLMs-based method, LLM-AR (Qu, Cai, and Liu 2024), we outperform on $X-sub$ and $X-view1$ by 3.2% and 14.8%, respectively. We analyze that: the activities recorded in the Toyota Smarthome cannot be distinguished solely by their motion trajectories. Benefiting from

| Methods | Recognizer | Skeleton | Toyota Smarthome | | | PKU-MMD | | NTU RGB+D | | NTU RGB+D 120 | |
|---------------------|------------|--------------|------------------|-----------------|-----------------|--------------|---------------|--------------|---------------|---------------|---------------|
| | | | <i>X-sub</i> | <i>X-view</i> 1 | <i>X-view</i> 2 | <i>X-sub</i> | <i>X-view</i> | <i>X-sub</i> | <i>X-view</i> | <i>X-sub</i> | <i>X-view</i> |
| 2s-AGCN | FC | Joint + Bone | 55.7 | 21.6 | 53.3 | 84.5 | 92.1 | 84.2 | 93.0 | 78.2 | 82.9 |
| MS-G3D | FC | Joint + Bone | 55.9 | 17.4 | 56.7 | - | - | 86.0 | 94.1 | 80.2 | 86.1 |
| ST-GCN | FC | Joint + Bone | 62.9 | 40.6 | 51.4 | 86.7 | 92.7 | 81.5 | 88.3 | 82.1 | 84.5 |
| Shift-GCN | FC | Joint + Bone | - | - | - | - | - | 87.8 | 95.1 | 80.9 | 83.2 |
| SSTA-PRS | FC | Joint + Bone | 62.1 | 22.8 | 54.0 | - | - | - | - | - | - |
| ML-STGNet | FC | Joint + Bone | 64.6 | 29.9 | 63.5 | - | - | - | - | - | - |
| UNIK | FC | Joint + Bone | 62.1 | 33.4 | 63.6 | - | - | 86.8 | 94.4 | 80.8 | 86.5 |
| LLM-AR | LLM | Joint | 67.0 | 36.1 | 66.6 | - | - | 95.0 | 98.4 | 88.7 | 91.5 |
| SUGAR (Ours) | LLM | Joint | 70.2 | 50.9 | 67.1 | 89.0 | 94.3 | 95.2 | <u>97.8</u> | 90.1 | <u>89.7</u> |

Table 1: Comparison result (%) of skeleton-based methods on different action classification benchmarks. FC denotes linear-based methods, and LLM denotes LLMs-based methods.

| Methods | Protocol 1 | | Protocol 2 | |
|---------------------|-------------|-------------|-------------|-------------|
| | Top-1 | Top-5 | Top-1 | Top-5 |
| <i>Linear-based</i> | | | | |
| ST-GCN | 30.1 | 45.2 | 36.9 | 55.2 |
| 2S-AGCN | 32.3 | 45.5 | 35.6 | 54.5 |
| CTR-GCN | 34.5 | 46.7 | 34.2 | 55.1 |
| <i>LLM-based</i> | | | | |
| LLM-AR | 59.7 | 84.1 | 49.4 | 74.2 |
| SUGAR | 65.3 | 89.8 | 53.4 | 77.6 |

Table 2: Evaluation on unseen activities. We report the Top-1 and Top-5 overall accuracy.

| Language Knowledge | Acc(%) |
|--------------------|-------------|
| w/o visual, motion | 69.2 |
| w/ visual | 69.4 |
| w/ motion | 72.1 |
| w/ visual, motion | 73.4 |

Table 3: Results on the overall accuracy of incorporating different language knowledge in SH.

the introduction of visual knowledge, the skeleton can learn more diverse representations from high-level visual semantics, which makes it easier to recognize such composite actions compared to skeleton-only methods.

For the results in NTU RGB+D (Shahroudy et al. 2016), NTU RGB+D 120 (Liu et al. 2020), and PKU-MMD (Liu et al. 2017), compared to previous state-of-the-art methods, we consistently achieve competitive results across all the evaluation protocols. Furthermore, our paradigm of introducing motion knowledge to supervise the learning of skeletons makes the LLM obtain a more discrete representation. We strongly believe that this way has greater generality in action recognition, especially in zero-shot reasoning (which is discussed in Sec.).

| Methods | Acc (%) |
|---------------------------|-------------|
| X-Attn. | 52.1 |
| One Q-Former | 70.7 |
| One linear layer | 70.4 |
| Temporal query projection | 73.4 |

Table 4: Results on the overall accuracy of different bridging modules in SH (Das et al. 2019). X-Attn. denotes mimicking cross-attention in (Ye et al. 2024).

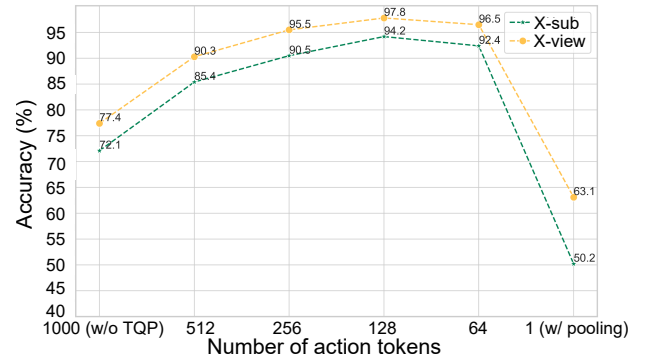


Figure 4: The effect of the length of action tokens on NTU RGB+D, where 1000 denotes the length of the complete time series output by the skeleton encoder and 1 represents action tokens after temporal compression.

Zero-shot on Action Recognition

Compared to traditional linear-based recognition approaches, we argue that the rich prior knowledge contained in large language models might benefit models' generalization capacities for recognizing unseen activities. To verify this conjecture, we build two evaluation protocols: the first is to pre-train on NTU 60 and then infer on 10 unseen action classes of NTU 120 cross-subject. The second one is cross-dataset testing, i.e., pre-training on NTU 60 and then inference on the test set of PKU-MMD cross-subject. All evaluation protocols report overall accuracy, and to explore performance in more detail, we instruct SUGAR to output

what it considers to be the top five most compatible action categories as: “Given a sequence of action tokens [action], please choose the top five most compatible action from [action list].”

To better represent the superiority of SUGAR for zero-shot inference, we design a linear-based classifier with ST-GCN (Yan, Xiong, and Lin 2018) as the backbone and pre-train it on NTU 60. As shown in Table 2, we find an interesting phenomenon: SUGAR is highly adaptable to both unseen action categories and even action across datasets. Compared to linear-based methods, the logistic distribution of the classifier output cannot be directly adapted to action categories from other datasets, which is a limitation compared to LLMs-based methods. In summary, SUGAR holds a promising recognition ability for unseen activities, and this experiment justifies the need for large language models as an alternative to linear-based approaches.

Ablation Studies

Impact of Visual-Motion Knowledge. We test the impact of the encoder on the classification results before and after adding this knowledge in Table 3. Notably, we train a skeleton encoder without adding visual knowledge or motion knowledge using the cross-entropy loss. The results show that using language prior knowledge to guide encoder learning leads to a stable performance improvement of 4.2% in overall accuracy.

Impact of TQP. To validate the effectiveness of our core module TQP, we report in Table 4 the impact of comparing different bridging strategies on the Toyota Smarthome. Initially, we attempt to transform skeleton representations into action tokens using the three currently dominant bridging modules: linear layer, cross-attention, and Q-Former. As the results show, we are surprised that one linear layer can bring results that are already able to approach the state-of-the-art classifier-based methods. However, such a way poses the problem of generating excessively long tokens, resulting in a longer overall training time. On the contrary, Q-Former’s query mechanism can effectively reduce the length of input tokens to minimize the computational cost. Further, our proposed TQP can query the temporal information of skeleton representation in a sequential manner, which improves our performance by 2.7% compared to a single Q-Former.

Length of Action Tokens. Another matter worth exploring is the length of the action tokens that feed into the LLM. We evaluate the impact of different input lengths on NTU 60 in Fig. 4. In this study, We test 6 different input lengths: 1000 (full length), 512, 256, 128, 64 and 1 (time-compressed) to find the most suitable one. The results show that by continuously reducing the input length, SUGAR shows signs of incremental increase. This further suggests that LLM may not excel at modeling non-language tokens with long temporal information. Furthermore, an interesting finding is that SUGAR has the worst results when we try to compress the length of action tokens to 1. We argue that many of the skeleton signals that represent different actions are too similar, and that performing compression makes such data more homogeneous to each other, leading to LLM’s inability to discriminate between these actions. Therefore, we define the

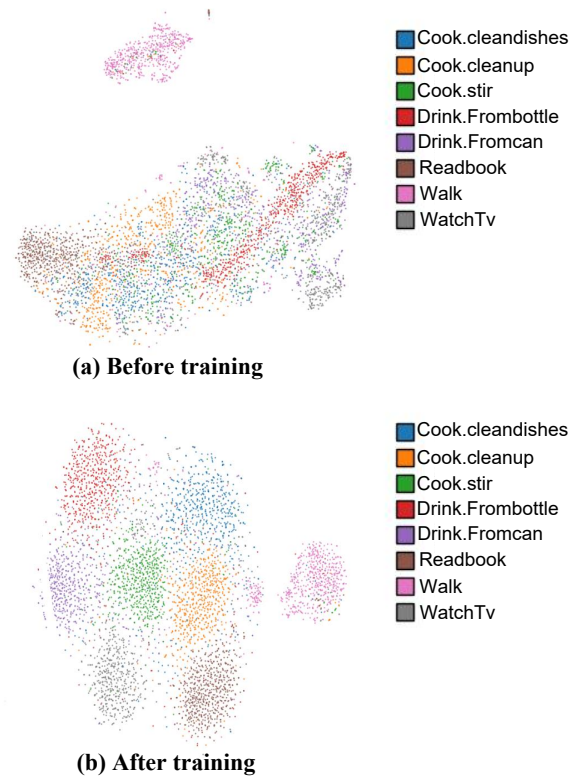


Figure 5: t-SNE (Van der Maaten and Hinton 2008) visualization of the learned skeleton representation. We show 8 action classes of the Toyota Smarthome, indicated by colors.

length L of the query vector q in TQP as 128.

Visualization of Learned Representations

We present the t-SNE (Van der Maaten and Hinton 2008) visualization of the skeleton representation before and after the supervised training with language knowledge in Fig. 5. Notably, we select visual distributions of similar action categories, e.g., “Drink From bottle” and “Drink From can”, “clean up”, and “clean dishes”. We find that the representations acquired by introducing visual-motion knowledge to supervise during training are more discrete compared to the initial, and even similar actions are largely separated. We believe that in this way it allows LLM to better discriminate between similar actions.

Conclusion

In this paper, we develop a novel paradigm named SUGAR for skeleton-based action recognition, which introduces rich visual-motion knowledge to supervise the learning of skeleton representations and utilizes LLM as a recognizer. Experiments demonstrate that SUGAR is superior to traditional linear-based methods and has a strong generalization ability.

Acknowledgments

This research was funded by the Beijing Zhongguancun Academy (Grant No. 20240306), the National Natu-

ral Science Foundation of China (Grant No. 62376266 and 62406318), the CCF-Tencent Rhino-Bird Open Research Fund, Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047), SongShan Lake HPC Center (SSL-HPC) in Great Bay University. and the National Natural Science Foundation of China (Grant No. 62576076).

References

- Baradel, F.; Wolf, C.; and Mille, J. 2017. Human action recognition: Pose-based attention draws focus to hands. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 604–613.
- Bruce, X.; Liu, Y.; Zhang, X.; Zhong, S.-h.; and Chan, K. C. 2022. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3): 3522–3538.
- Cai, D.; Kang, Y.; Yao, A.; and Chen, Y. 2023. Ske2Grid: Skeleton-to-Grid Representation Learning for Action Recognition. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Int. Conf. Mach. Learn.*, volume 202, 3431–3441.
- Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *Int. Conf. Comput. Vis.*, 13339–13348.
- Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; and Lu, H. 2020. Decoupling GCN with DropGraph Module for Skeleton-Based Action Recognition. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Eur. Conf. Comput. Vis.*, 536–553.
- Cheng, K.; Zhang, Y.; He, X.; Cheng, J.; and Lu, H. 2021. Extremely Lightweight Skeleton-Based Action Recognition With ShiftGCN++. *IEEE Trans. Image Process.*, 30: 7333–7348.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.
- Das, S.; Dai, R.; Koperski, M.; Minciullo, L.; Garattoni, L.; Brémond, F.; and Francesca, G. 2019. Toyota Smarthome: Real-World Activities of Daily Living. In *Int. Conf. Comput. Vis.*, 833–842. IEEE.
- Das, S.; Dai, R.; Yang, D.; and Bremond, F. 2021. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 9703–9717.
- Das, S.; Sharma, S.; Dai, R.; Bremond, F.; and Thonnat, M. 2020. Vpn: Learning video-pose embedding for activities of daily living. In *Eur. Conf. Comput. Vis.*, 72–90. Springer.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR*.
- Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2023. VTimeLLM: Empower LLM to Grasp Video Moments. *CoRR*, abs/2311.18445.
- Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Int. Conf. Learn. Represent.* OpenReview.net.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. C. H. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Int. Conf. Mach. Learn.*, 19730–19742. PMLR.
- Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023b. VideoChat: Chat-Centric Video Understanding. *CoRR*, abs/2305.06355.
- Li, S.; Zhang, Y.; Zhao, Y.; Wang, Q.; Jia, F.; Liu, Y.; and Wang, T. 2023c. VLM-Eval: A General Evaluation on Video Large Language Models. *CoRR*, abs/2311.11865.
- Li, W.; Luo, D.; Yang, D.; Li, Z.; Wang, W.; and Zhou, Y. 2025. The Role of Video Generation in Enhancing Data-Limited Action Understanding. *arXiv preprint arXiv:2505.19495*.
- Li, Y.; Xu, L.; Liu, X.; Huang, X.; Xu, Y.; Wang, S.; Fang, H.; Ma, Z.; Chen, M.; and Lu, C. 2020. PaStaNet: Toward Human Activity Knowledge Engine. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 379–388. Computer Vision Foundation / IEEE.
- Lin, X.; Liu, A.; Yu, Z.; Cai, R.; Wang, S.; Yu, Y.; Wan, J.; Lei, Z.; Cao, X.; and Kot, A. 2025. Reliable and Balanced Transfer Learning for Generalized Multimodal Face Anti-Spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017. PKU-MMD: A Large Scale Benchmark for Skeleton-Based Human Action Understanding. In Liu, X.; Mu, Y.; Jiang, Y.; and Luo, J., eds., *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities, VSCC@MM 2017, Mountain View, CA, USA, October 23, 2017*, 1–8. ACM.
- Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.; and Kot, A. C. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10): 2684–2701.
- Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.; Lu, P.; Wang, T.; and Wei, Z. 2023. Valley: Video Assistant with Large Language model Enhanced ability. *CoRR*, abs/2306.07207.
- Miech, A.; Alayrac, J.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 9876–9886. Computer Vision Foundation / IEEE.
- Nan, H.; Ye, Q.; Yu, Z.; and An, K. 2024. 3sG: Three-stage guidance for indoor human action recognition. *IET Image Processing*, 18(8): 2000–2010.
- Noor, N.; and Park, I. K. 2023. A Lightweight Skeleton-Based 3D-CNN for Real-Time Fall Detection and Action Recognition. In *Int. Conf. Comput. Vis.*, 2171–2180. IEEE.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

- Qu, H.; Cai, Y.; and Liu, J. 2024. Llms are good action recognizers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 18395–18406.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 8748–8763. PMLR.
- Rajendran, M.; Tan, C. T.; Atmosukarto, I.; Ng, A. B.; and See, S. 2024. Review on synergizing the Metaverse and AI-driven synthetic data: enhancing virtual realms and activity recognition in computer vision. *Visual Intelligence*, 2(1): 27.
- Shahroudy, A.; Liu, J.; Ng, T.; and Wang, G. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1010–1019. IEEE Computer Society.
- Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019a. Skeleton-Based Action Recognition With Directed Graph Neural Networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 7912–7921. Computer Vision Foundation / IEEE.
- Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019b. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 12026–12035.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.
- Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- Wang, L.; Boddeti, V.; and Lim, S. 2024. Action Reimagined: Text-to-Pose Video Editing for Dynamic Human Actions. *arXiv preprint arXiv:2403.07198*.
- Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.
- Xie, X.; Cui, Y.; Tan, T.; Zheng, X.; and Yu, Z. 2024. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37.
- Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In McIlraith, S. A.; and Weinberger, K. Q., eds., *AAAI*, 7444–7452. AAAI Press.
- Yang, D.; Dai, R.; Wang, Y.; Mallick, R.; Minciullo, L.; Francesca, G.; and Brémond, F. 2021. Selective Spatio-Temporal Aggregation Based Pose Refinement System: Towards Understanding Human Activities in Real-World Videos. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 2362–2371. IEEE.
- Ye, Q.; and Yu, Z. 2024. Pose-Promote: Progressive Visual Perception for Activities of Daily Living. *IEEE Signal Processing Letters*.
- Ye, Q.; Yu, Z.; Shao, R.; Cui, Y.; Kang, X.; Liu, X.; Torr, P.; and Cao, X. 2025. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Ye, Q.; Yu, Z.; Shao, R.; Xie, X.; Torr, P.; and Cao, X. 2024. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, 146–164. Springer.
- Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP*, 543–553.
- Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023. Generating human motion from textual descriptions with discrete representations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 14730–14740.
- Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1109–1118. Computer Vision Foundation / IEEE.
- Zhou, H.; Liu, Q.; and Wang, Y. 2023. Learning Discriminative Representations for Skeleton Based Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 10608–10617. IEEE.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9): 2337–2348.