Generative Action Description Prompts for Skeleton-based Action Recognition - Supplementary Material -

Wangmeng Xiang[1,2] Chao Li[2] Yuxuan Zhou[2,3] Biao Wang[2] Lei Zhang[1*]
[1]The Hong Kong Polytechnic University [2]DAMO Academy, Alibaba Group [3]Mannheim University
{wangmeng.xwm,lllcho.lc,wb.wangbiao}@alibaba-inc.com, yuxuazho@mail.uni-mannheim.de
cslzhang@comp.polyu.edu.hk

The following materials are provided in our supplementary file:

- Implementation details of GAP framework on different datasets.
- Comparison of losses CL, KLD, and JSD.
- Ablation studies on ST-GCN backbone.
- Visualization of text feature similarity matrices on different actions.
- Action descriptions generated by different templates on NTU RGB-D 120. Prompts for NTU RGB-D and NW-UCLA can be found at https://github.com/MartinXM/GAP.

1. Implementation details

Part partition. The body part partition is slightly different in NW-UCLA as it only contains 20 joints, while NTU RGB-D and NTU RGB-D 120 contain 25 joints. However, the overall groupings remain the same, which contain four parts: head, hands, hip, legs.

Implementation. Our implementation is based on CTR-GCN [1]. We also adopt the data pre-processing in InfoGCN [2], where $K = 1$ for bone modality and $K = 8$ or 6 for joint modality in NTU RGB-D 60/120 and NW-UCLA, respectively. For the implementation of text encoder, two different pre-training schemes (image-text and pure text) are considered. For image-text pre-training, we adopt CLIP [3]. For pure text pre-training text encoder, we adopt Roberta[1].

We implement our framework with Pytorch[2]. All the models in our experiments are trained with 2 RTX 3090 GPUs and the seed is set to 1. Mixed-precision training is adopted to accelerate training speed and reduce memory footprint.

2. Comparison of losses

We compare 3 different losses: standard Contrastive Loss (CL), KL-Divergence (KLD) and Jenson-Shannon Divergence (JSD) on NTU120 X-sub, whose results are **85.4%**, **85.5%**, **85.7%**, respectively. Fig. 1 shows that CL converges faster than KLD and JSD but with lower performance. JSD provides the best performance, probably due to its symmetric and smooth property.

*Corresponding author
[1]https://github.com/huggingface/transformers
[2]https://pytorch.org

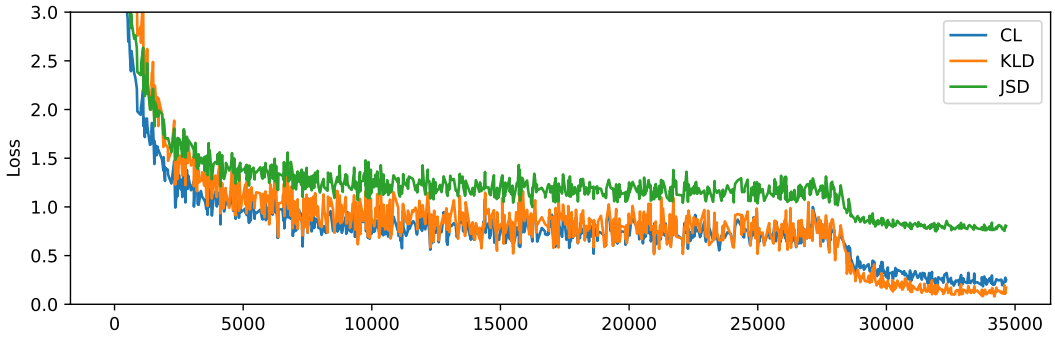


Figure 1: Convergence curves of CL, KLD and JSD losses.

3. Ablation studies on ST-GCN backbone

In Tab. 1, we reran two ablation studies with weaker ST-GCN backbone to show clearer trends of our design choices. We also ran NTU120 Xsub joint for **5** runs. The mean±std is **85.46±0.13**. The mean is the same as our reported result in the main paper, which indicates that our method is relatively stable.

(a) Text prompt type

Model Strategy	Acc(%)
ST-GCN	82.61±0.11
Label name	82.79±0.13
Syn./Para.	83.22±0.12
Body parts	83.54±0.13
Syn.+Body parts	83.84±0.14

(b) Description methods

Methods	Acc(%)
Part CLS	82.50±0.11
Manual	83.18±0.10
HAKE	83.21±0.15
GPT-3	83.84±0.14

Table 1: Ablation with weaker ST-GCN backbone.

4. Visualization

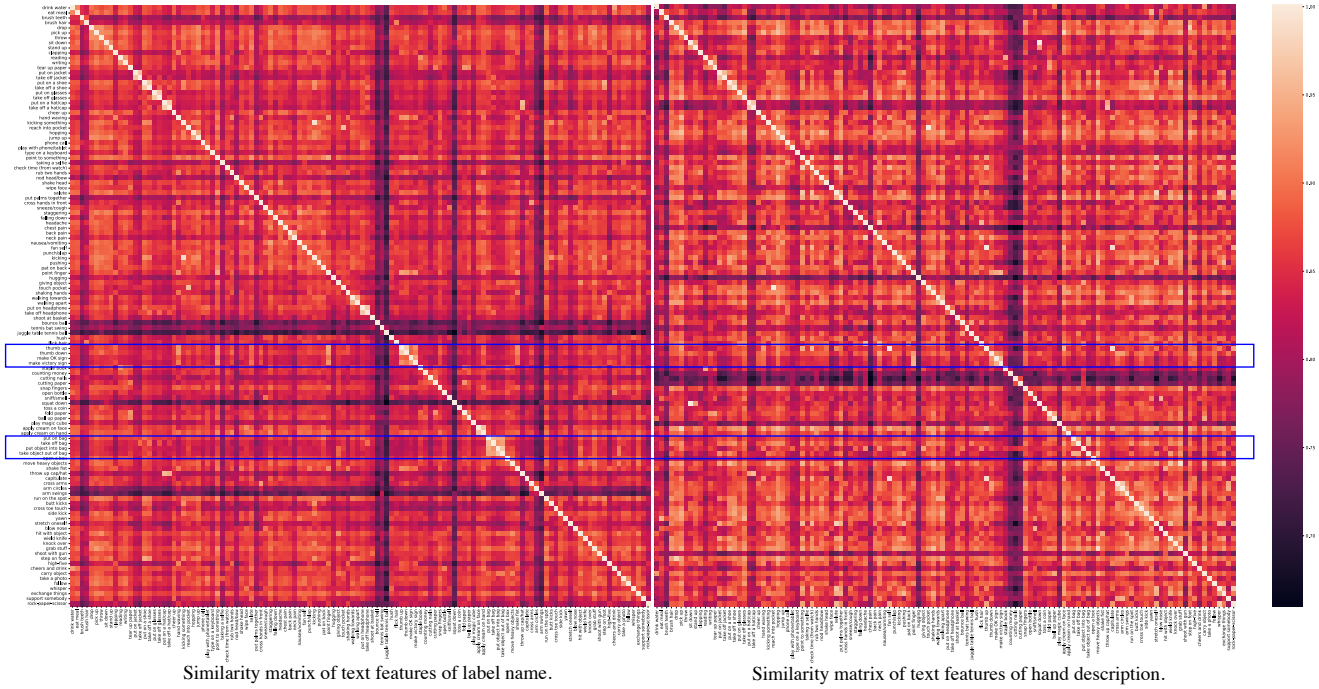
In Figure 2, we display similarity matrices of text features derived from both the label name and the description of hand parts. It becomes evident that the descriptions of hand parts demonstrate a greater discriminative capacity for actions that primarily involve the hands, such as "thumb up", "thumb down", "make OK sign", and "make victory sign". Furthermore, these descriptions are more effective for actions centered on arm movements, such as "put on bag", "take off bag", "put object into bag", and "take object out of bag". Analogously, as illustrated in Figure 3, text features from foot descriptions are more applicable for actions that predominantly focus on the feet, such as "put on a shoe" and "take off a shoe". However, foot descriptions perform poorly on actions that primarily involve the hands. Consequently, the best results are achieved by combining these descriptions, as demonstrated in our paper.

5. Text descriptions

We provide text descriptions generated by GPT-3 (text-davinci-002) for different text prompts on NTU RGB-D 120 on https://github.com/MartinXM/GAP in *paragraph-GPT3.txt*, *synonym-GPT3.txt* and *part-GPT3.txt*, respectively. The text prompts used for generating text descriptions are as follows:

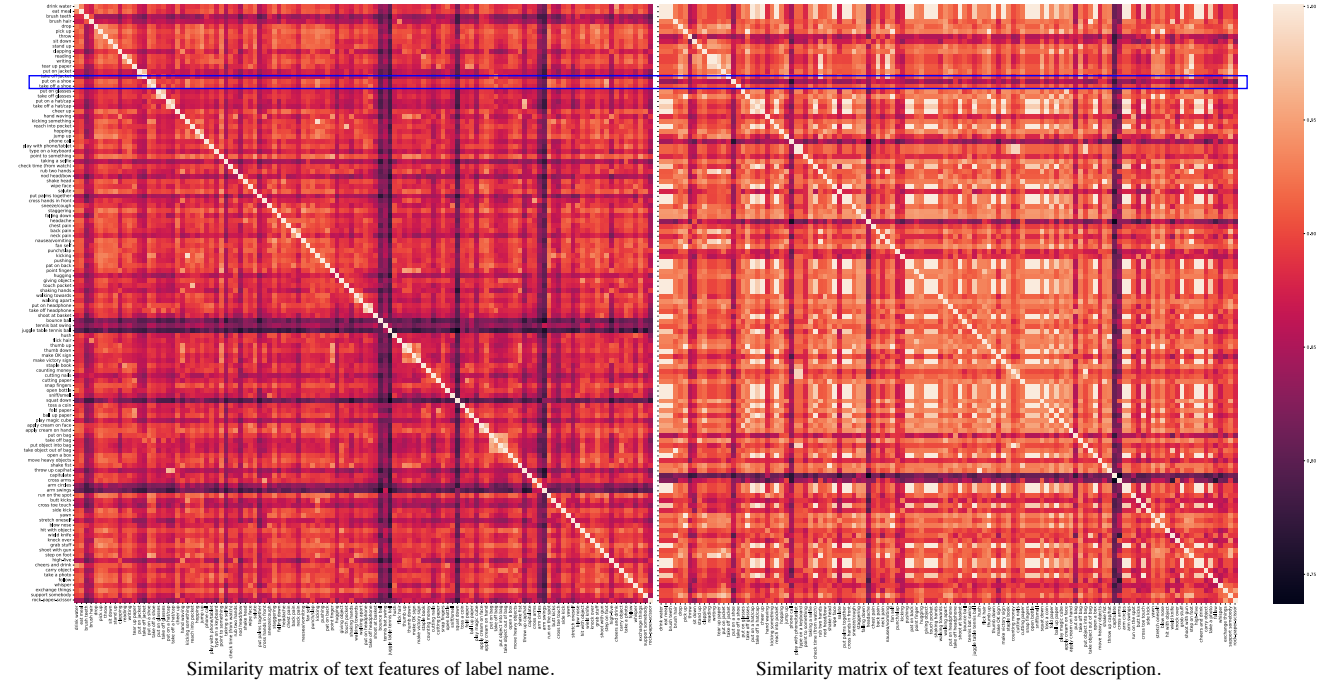
- Describe a person [action] in details.
- Suggest 10 synonyms for [action].
- Describing following body parts actions when [action]: head, hand, arm, hip, leg, foot.

The generated descriptions are only edited to correct format error of GPT-3 without changing the semantic contents. We found that **paragraph** provides rich descriptions but sometimes with unnecessary details. **Synonym**



Similarity matrix of text features of label name. Similarity matrix of text features of hand description.

Figure 2: Similarity matrices of text features using label name and hand part description.



Similarity matrix of text features of label name. Similarity matrix of text features of foot description.

Figure 3: Similarity matrices of text features using label name and foot part description.

enriches the label name meanings with similar words. **Part description** provides the most precise description of actions.

Several text descriptions are shown below:

drink water

- **Paragraph**: The man is sitting at a table with a plate of food in front of him. He is using a fork and knife to eat the food on his plate. He is chewing each bite of food before swallowing. He is also taking breaks in between bites to drink from a glass of water.
- **Synonym**: dine, sup, lunch, feast, banquet, picnic, breakfast, brunch, dinner, lunch.
- **Part description**: head tilts slightly forward; hand brings food to mouth; arm supports hand; hip remains stationary; leg remains stationary; foot remains stationary.

eat meal

- **Paragraph**: The man is drinking water from a glass. He is holding the glass in his hand and bringing it to his lips. He is tilting the glass back and forth, allowing the water to flow into his mouth. He is swallowing the water and then setting the glass down on the table.
- **Synonym**: sip, guzzle, gulp, swig, chug, quaff, swill, slug, chug down, toss back.
- **Part description**: head tilts back slightly; hand grasps cup; arm lifts cup to mouth; hip remains stationary; leg remains stationary; foot remains stationary.

take off jacket

- **Paragraph**: The man is taking off his jacket. He is standing up straight and reaching his arms up above his head. His jacket is coming off easily and he is taking it off quickly. He is not having any trouble taking his jacket off.
- **Synonym**: remove jacket, take jacket off, divest oneself of jacket, unburden oneself of jacket, get rid of jacket, be rid of jacket, dispose of jacket, get jacket off, shed jacket, slip off jacket.
- **Part description**: head tilts back slightly; grabs the bottom of jacket with both hands, brings hands up the jacket; arms straighten as the jacket falls downs;hip steps out of the jacket;legs straight;feet on the ground.

make victory sign

- **Paragraph**: He is standing with his feet apart and his arms raised in the air, making a V sign with his fingers. He has a triumphant look on his face and is clearly enjoying himself.
- **Synonym**: give the thumbs up, give a thumbs up, give the okay sign, give the A-okay sign, give the victory sign, give a V sign, give a peace sign, give the finger, give the bird, give the one-finger salute
- **Part description**: head tilts slightly forward; hand forms a V shape with the index and middle fingers; arm extends fully; hip remains stationary; leg remains stationary; foot remains stationary.

make ok sign

- **Paragraph**: The man is making an OK sign with his hand. His thumb and index finger are touching to form a circle, and his other three fingers are extended. He may be doing this to indicate that everything is okay, or to give someone the okay sign.
- **Synonym**: A-OK, all right, alright, copacetic, dandy, fine, good, hunky-dory, okay, swell
- **Part description**: head tilts slightly to the side; hand forms a circle with the thumb and first two fingers; arm extends straight out from the shoulder; hip remains at a neutral position; leg remains at a neutral position; foot remains at a neutral position.

move heavy objects

- **Paragraph**: The man is muscular with broad shoulders. He has a deep tan, and his arms are covered in a light sheen of sweat. He grunts with effort as he lifts a heavy object, his biceps straining. His face is set in a determined expression, and his eyes are focused on the task at hand. He moves with purpose, each step sure and confident. He makes the task look easy, despite the obvious strain it is causing him.
- **Synonym**: drag, haul, lug, tow, carry, convey, transport, shift, relocate, remove
- **Part description**: head turns to look at the object; hand reaches out and grasps the object; arm pulls the object towards the body; hip moves forward to create leverage; leg pushes against the ground to create power; foot stabilizes the body.

support somebody

- **Paragraph**: He is standing behind the person in front of him, with his hands on their shoulders. He is looking down at them with a concerned expression on his face. He is speaking quietly to them, offering words of encouragement.
- **Synonym**: help, assist, back, prop, buttress, shore up, strengthen, reinforce, hold up, lift up
- **Part description**: head tilts slightly forward; hand grasps the other person's arm just above the elbow; arm supports the other person's arm; hip stands upright; leg stands upright; foot stands flat on the ground.

open a box

- **Paragraph**: The man is standing in front of a box. He is reaching for the lid of the box. He is opening the lid of the box. He is looking inside the box.
- **Synonym**: unpack, unseal, unbox, decant, disentangle, extricate, liberate, release, remove, untie
- **Part description**: head tilts slightly forward; hand reaches out and grasps the edge of the lid; arm extends forward; hip remains stationary; leg remains stationary; foot remains stationary.

open bottle

- **Paragraph**: He unscrews the cap of the bottle with one hand, while holding the base of the bottle in the other. He twists the cap until it comes off with a small pop. He brings the bottle to his lips and takes a long drink, savoring the flavor of the liquid inside.
- **Synonym**: jar, can, container, vessel, receptacle, flask, decanter, urn, cruet, carafe
- **Part description**: head tilts back slightly; hand grasps the neck of the bottle; arm extends the arm holding the bottle; hip remains stationary; leg remains stationary; foot remains stationary.

jump up

- **Paragraph**: He is a man who is of average height and build. He has dark hair and eyes, and is wearing a pair of jeans and a t-shirt. He is jump up in the air, and his arms and legs are outstretched. He has a look of concentration on his face, and he is landing on his feet.
- **Synonym**: leap, bound, spring, vault, hop, skip, caper, gambol, frisk, frolic
- **Part description**: head tilts back and the chin points up; hands come up to the chest; arms bend at the elbows and the forearms come up; hips push forward and the legs bend at the knees; legs push off the ground and the feet come up; feet land on the ground and the legs bend at the knees.

References

[1] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 1

[2] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20186–20196, June 2022. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1