LocLLM: Exploiting Generalizable Human Keypoint Localization via Large Language Model

Supplementary Material

Dongkai Wang Shiyu Xuan Shiliang Zhang
National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
{dongkai.wang, slzhang.jdl}@pku.edu.cn, shiyu_xuan@stu.pku.edu.cn

1. Keypoint Location Description

In supplemental material we provide the detailed keypoint location description used by LocLLM. Table 1 shows the all keypoint description, which is generated by ChatGPT and we manually check them with wikipedia to ensure correctness.

Table 1. **The keypoint location description.** These descriptions are generated by ChatGPT and we manually check it with wikipedia.

- *nose*: The *nose* is the central, protruding feature on their face, located just above the upper lip.
- *left eye*: The *left eye* is the visual organ on the left side of their face, typically located above the left cheek and beside the nose.
- *right eye*: The *right eye* is the visual organ on the right side of their face, typically located above the right cheek and beside the nose.
- *left ear*: The *left ear* is the auditory organ on the left side of their head, typically located to the side of the left temple.
- *right ear*: The *right ear* is the auditory organ on the right side of their head, typically located to the side of the right temple.
- *left shoulder*: The *left shoulder* is the joint connecting the left arm and the torso, typically situated on the upper left side of the chest.
- *right shoulder*: The *right shoulder* is the joint connecting the right arm and the torso, typically situated on the upper right side of the chest.
- *left elbow*: The *left elbow* is the joint connecting the left upper arm and the left forearm, typically situated in the middle of the left arm, between left shoulder and left wrist.
- *right elbow*: The *right elbow* is the joint connecting the right upper arm and the right forearm, typically situated in the middle of the right arm, between right shoulder and right wrist.
- *left wrist*: The *left wrist* is the joint connecting the left forearm and the left hand, typically located at the base of the left hand.
- *right wrist*: The *right wrist* is the joint connecting the right forearm and the right hand, typically located at the base of the right hand.
- *left hip*: The *left hip* is the joint connecting the left thigh to the pelvis, typically located on the left side of the lower torso.
- *right hip*: The *right hip* is the joint connecting the right thigh to the pelvis, typically located on the right side of the lower torso.
- *left knee*: The *left knee* is the joint connecting the left thigh and the left lower leg, typically situated in the middle of the left leg, it is located between the left hip and left ankle.
- *right knee*: The *right knee* is the joint connecting the upper leg and lower leg on the right side, it is located between the right hip and right ankle.
- *left ankle*: The *left ankle* is the joint connecting the left lower leg and the left foot, typically located at the base of the left leg.
- *right ankle*: The *right ankle* is the joint connecting the right lower leg and the right foot, typically located at the base of the right leg.
- *neck*: The *neck* is the part of the body connecting the head to the torso, typically situated between the shoulders.
- *pelvis*: The *pelvis* is the bony structure that forms the base of the spine and connects the torso to the lower body, typically located between the left hip and right hip.

2. Details of Inference Procedure

LocLLM is designed to locate only one type of keypoint in a single forward, and a batch inference is needed to locate multiple type of keypoints (*e.g.*, stacking nose, knee *etc* inputs at batch dimension) for an image. This single keypoint inference design allows us to flexibly locate novel type of keypoint by providing corresponding novel descriptions. How to accelerate the inference process is a valuable direction for future research.

3. Analysis on Numeric Value Prediction with Cross Entropy Loss

Background: How does the Cross Entropy loss handle textual predictions that are slightly off in numeric value, but significantly off in 'text'? For example, 0.399 vs. 0.401. It's not too clear how LocLLM handles this scenario.

Cross Entropy (CE) loss is the standard training loss for existing autoregressive LLMs, *e.g.*, Llama [3], Vicuna [1], LLaVA [2] and Mini-GPT4 [4]. Therefore, it is natural to adopt it to train our LocLLM and align the objective function with pretrained LLMs. Adopting CE loss also allows us to utilize the pretrained classifier in LLM, making it more effective for following visual instruction tuning.

To investigate the capability of handling boundary cases in LocLLM, we count the predictions of LocLLM on keypoints with ground truth in $[0.395, 0.405]$ on COCO `val` set (with 111 samples in total). All predictions of above samples are in range $[0.361, 0.434]$.

	Gt$\in [0.395, 0.400)$	Gt$\in [0.400, 0.405]$
Pred$\in [0.361, 0.434]$	49	62
Pred$\in [0.361, 0.399]$	29 (59%)	25 (40%)
Pred$\in [0.400, 0.434]$	20 (41%)	37 (60%)

It can be observed that for keypoints in $[0.395, 0.399]$(2nd column), LocLLM can output coordinates both start from 0.3xx (59%) and 0.4xx (41%), indicating LocLLM with CE is not biased to one direction in boundary cases. We think it is because LLM is not only focusing on a single token but can understand the context of a set of tokens, making it could handle boundary cases of numeric value well. This contextual understanding capability of LLM is also verified in many NLP and vision-language tasks.

References

[1] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2