L-Man: A Large Multi-modal Model Unifying Human-centric Tasks

Jialong Zuo[1], Ying Nie[2], Tianyu Guo[2], Huaxin Zhang[1], Jiahao Hong[1], Nong Sang[1], Changxin Gao[1]*, Kai Han[2]

[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2] Huawei Noah's Ark Lab

Abstract

Large language models (LLMs) have recently shown notable progress in unifying various visual tasks with an open-ended form. However, when transferred to human-centric tasks, despite their remarkable multi-modal understanding ability in general domains, they lack further human-related domain knowledge and show unsatisfactory performance. Meanwhile, current human-centric unified models are mostly restricted to a pre-defined form and lack open-ended task capability. Therefore, it is necessary to propose a large multi-modal model which utilizes LLMs to unify various human-centric tasks. We forge ahead along this path from the aspects of dataset and model. Specifically, we first construct a large-scale language-image instruction-following dataset named HumanIns based on existing 20 open datasets from 6 diverse downstream tasks, which provides sufficient and diverse data to implement multi-modal training. Then, a model named L-Man including a query adapter is designed to extract the multi-grained semantics of image and align the cross-modal information between image and text. In practice, we introduce a two-stage training strategy, where the first stage extracts generic text-relevant visual information, and the second stage maps the visual features to the embedding space of the LLM. By tuning on HumanIns, our model shows significant superiority on human-centric tasks compared with existing large multi-modal models, and also achieves better results on downstream datasets compared with respective task-specific models.

Introduction

Large language models (Touvron et al. 2023a,b; Brown et al. 2020; Raffel et al. 2020; Chowdhery et al. 2022; Ouyang et al. 2022) have shown remarkable performance as a general-purpose assistant in an unrestricted form. Motivated by their impressive generic potential across a range of applications, in the field of computer vision, some researchers (Liu et al. 2023b; Bai et al. 2023; Dai et al. 2023) have extended the ability of LLMs into vision-centric tasks by bridging the gap between visual and textual features. However, when turning into human-centric tasks, these general methods often suffer from limited domain-specific knowledge and lack sensitivity to human-related fine-grained characteristics, which results in unsatisfactory performance. Meanwhile, due to that there are obvious correlations among different human-centric tasks, some researchers (Ci et al. 2023; Zuo et al. 2024a; Tang et al. 2023) are dedicated to develop an unified visual model that can benefit diverse human-centric tasks. However, while these unified models have achieved generalist capability to a certain extent, they are mostly restricted to a pre-defined form and lack open-ended task capability.

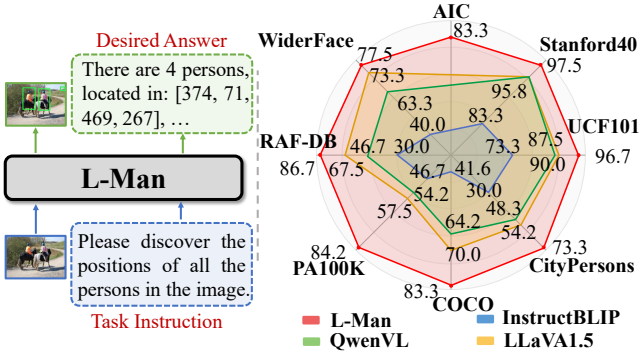

Figure 1: L-Man unifies various human-centric tasks in an instruction-following form and achieves SoTA performance on several datasets compared with InstructBLIP (Dai et al. 2023), Qwen-VL-Chat (Bai et al. 2023) and LLaVA1.5 (Liu et al. 2023a).

Therefore, this paper explores to propose a large multi-modal model which utilizes LLMs to unify various human-centric tasks and is not restricted to a pre-defined form, as shown in Figure 1. While this exploration is theoretically feasible, it is unavoidable to face two obstacles. First, although there are many human-related datasets, the annotations follow a pre-defined form, and there is currently no comprehensive language-image dataset that spans diverse human-centric tasks and provides ample multi-modal data for tuning. Second, most existing general large multi-modal models are trained by the objective of unified global consistencies and lack the perception of fine-grained details. However, in human-centric domain, multi-grained features are critical to achieving better performance.

To address these issues, we first build a large-scale

*Corresponding Author. (email: cgao@hust.edu.cn)

Figure 2: Overview of our HumanIns. It includes diversified scenes, tasks and instructions.

language-image instruction-following dataset, named **HumanIns**, based on existing human-related datasets. It provides sufficient and diverse data to implement multi-modal training. As shown in Figure 2, HumanIns has the following diversities. (1) Diversity of scenes. The images have a spanning scene distribution, ranging from day to night, from outdoors to indoors and from low resolution to high resolution. (2) Diversity of tasks. HumanIns covers diverse human-centric tasks for multi-modal tuning. (3) Diversity of instructions. HumanIns contains diverse and unrestrictive instructions. The rich and diversified data in HumanIns is indispensable for unifying human-centric tasks.

Meanwhile, we design a **L**arge **M**ulti-modal model for unifying hum**an**-centric tasks (**L-Man**). Specifically, our L-Man has the following characteristics. (1) Supporting multi-resolution image input. Downstream tasks have diverse image resolution requirements for optimal performance. General multi-modal methods lack a multi-resolution design, prompting our proposal for training with multi-resolution image input. (2) Perceiving multi-grained information. Human-centric tasks vary in granularity preferences. In contrast to many existing large multi-modal models (Liu et al. 2023b; Dai et al. 2023; Bai et al. 2023) with unified global consistencies, our framework is able to learn human-related features across diverse granularities. (3) Two-stage training strategy. The first stage pre-trains the query adapter to extracts text-relevant information from visual features at a generic representation level. The second stage combines various human-centric tasks in an instruction-following format, by utilizing the pre-trained query adapter to map visual tokens to the embedding space of the large language model.

Our L-Man trained on HumanIns shows significant superiority compared to some existing general models (Dai et al. 2023; Bai et al. 2023; Liu et al. 2023a) on a range of human-centric tasks, as shown in Figure 1. Also, it achieves even better performance compared with some strong task-specific baselines . In summary, our contributions are two folds: (1) we build HumanIns, a large-scale language-image instruction-following dataset covering diverse human-centric tasks. (2) We propose L-Man, a large multi-modal model for unifying human-centric tasks.

HumanIns Dataset

We build the HumanIns dataset to train our L-Man. In this section, we show the details of how to construct the dataset and its characteristic properties.

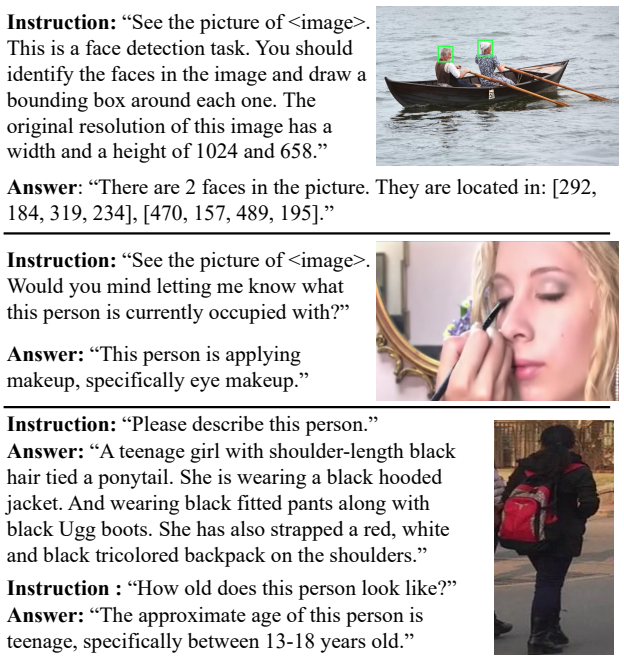

Figure 3: Some specific examples in HumanIns. From top to bottom are examples of face detection, action recognition, and pedestrian attribute recognition, respectively.

Dataset Construction

Image Collection and Process. In general, nonverbal communication in daily life includes action, visual appearance, and posture information. Following this common sense, we select action recognition (Yao et al. 2011; Soomro, Zamir, and Shah 2012) as the action task, pedestrian detection (Xiao et al. 2017; Zheng et al. 2017; Lin et al. 2014; Zhang, Benenson, and Schiele 2017; Dollar et al. 2011), pedestrian attribute recognition (Zheng et al. 2015; Ristani et al. 2016;

Liu et al. 2017; Li et al. 2017; Ding et al. 2021; Zuo et al. 2024c), face detection (Jain and Learned-Miller 2010; Yang et al. 2016), expression recognition (Li, Deng, and Du 2017; Zhang et al. 2018) as the visual appearance task, and human pose estimation (Wu et al. 2017; Andriluka et al. 2014) as the posture task in our HumanIns. Following the standard protocols, 20 datasets containing 908,587 images are collected as the samples in our dataset. The specific details are shown in the supplement. Worth noting, some samples (Soomro, Zamir, and Shah 2012) for the action recognition task are in video format, which contain large information redundancy between consecutive frames. In this case, we conduct frame sampling and classify all sampled frames in the same video as the same type.

Instruction-Answer Generation. We generate diverse instruction-answer pairs for different human-centric tasks in HumanIns. For the instructions of each task, we first manually design some task-related descriptions as a seed, then we enrich the descriptions by iteratively asking the LLM (Touvron et al. 2023b) to response with the following prompt: *"Please generate 5 sentences that have the same meaning as the following: [task description]"*. We will manually refine the generated responses to guarantee the correctness. For the answers, the datasets for different tasks can be classified to vision-only dataset and vision-language dataset. The former refers to datasets where there are no texts in the original annotations, while the latter has texts. For vision-only datasets of each task, we specify the desired answer format via manual rules. Taking face detection as an example, the original annotations only contain coordinates and a representative answer format via manual rules is shown in the top sample of Figure 3. Also, an example of the action recognition task is shown in the middle. For vision-language datasets including CUHK-PEDES (Li et al. 2017) and ICFG-PEDES (Ding et al. 2021) in pedestrian attribute recognition task, there are captions for describing the pedestrian's overall appearance of each image. To enrich the diversity, we ask the LLM (Touvron et al. 2023b) to response with the following instruction: *"This is an overall appearance of a pedestrian: <caption>. Please answer the following question according to it strictly: <question>."* By designing different targeted questions, such as *"How old does this person look like"*, we obtain more instruction-answer pairs for this task. Some examples are shown in the bottom of Figure 3. The complete details of task descriptions and output formats for each task are shown in the supplement.

Dataset Properties

The statistics of HumanIns are detailed in Table 1. Compared with existing human-related datasets (Tang et al. 2023; Ci et al. 2023; Zuo et al. 2024a,b), the properties of HumanIns are summarized as follows:

Diversified. Our dataset contains a wide range of variations in the scenes, tasks and instructions. Unlike the previous datasets with restricted pre-defined form, our dataset exhibits significant advantages in diversities due to the contained language-image instruction-following data.

High-quality. Our dataset is collected from a series of manually annotated datasets, ensuring a high level of quality in the annotations. Researchers can use this dataset with confidence to conduct relevant studies.

Tasks	datasets	Images	Instructions
Action Recognition	4	69,672	139,344
Pedestrian Detection	5	78,650	157,300
Pedestrian Attribute	5	254,364	1,295,601
Expression Recognition	2	102,644	102,644
Face Detection	2	13,548	27,096
Pose Estimation	2	389,709	779,418
Total	**20**	**908,587**	**2,501,403**

Table 1: Statistics of HumanIns.

Large-scale. Our dataset contains 908,587 images and 2,501,403 image-instruction pairs of 6 human-centric tasks totally, which is the largest dataset with instruction-answer pairs in the domain of unifying human-centric tasks by far.

L-Man

The proposed L-Man is an unified vision-language model for human-centric tasks. Given an image-instruction input pair, a pre-trained vision encocder CLIP-ViT-L14 (Radford et al. 2021) is first adopted to encode the input image, and the large language model Vicuna 7B (Zheng et al. 2023) is then utilized to decode the corresponding instruction. Meanwhile, a module of query adapter is introduced to fuse the embeddings of image and text. It should be noted that the embeddings of image are extracted from multiple intermediate layers. In addition, we employ a two-stage training strategy to improve the efficiency and performance. Also, we adopt a task sampling strategy to support the inputs of multi-resolution and multi-batchsize for different tasks.

Overall Model Architecture

Figure 4 illuminates the overall architecture of the proposed L-Man. For an input image, we first feed it to the pre-trained vision encoder (Radford et al. 2021) and extract hierarchical visual features from multiple intermediate layers. Then, we tokenize the corresponding input instruction to obtain the instruction tokens. The hierarchical visual features, instruction tokens and learnable query tokens will be sent to the query adapter to capture the multi-grained information of images. Finally, a sequence of visual tokens are output by the query adapter, which will be sent to the LLM (Zheng et al. 2023) in conjunction with the paired instruction to output the final decoded response.

Mechanism of The Query Adapter

Multi-grained Visual Features Extraction. Different tasks have significant variations in their granularity requirements for visual features. Extracting features solely from the last transformer layer of the vision encoder, with a single granularity, often struggles to meet the requirements of diverse tasks. Therefore, we adopt a multi-grained visual features extraction strategy. We begin by partitioning the input image into fixed-size grids, and then send each grid through an embedding layer to obtain visual input embeddings $\mathbf{V}_0$. Then,

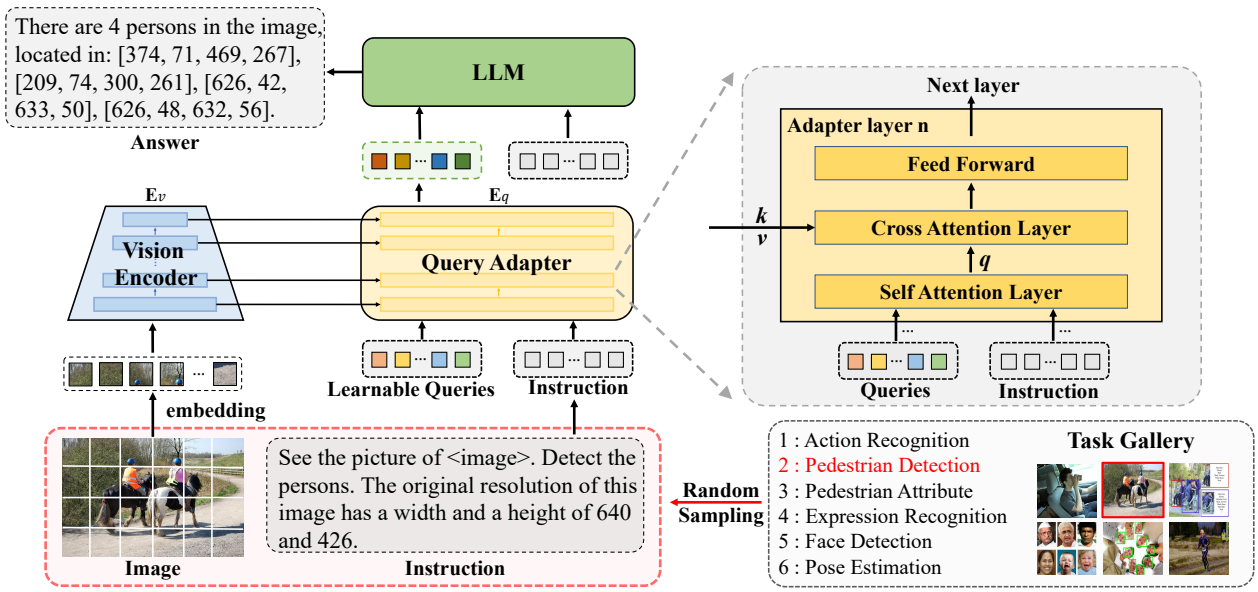

Figure 4: Overview of L-Man architecture. L-Man is consisted of a vision encoder, a query adapter and an LLM. We adopt a two-stage training strategy, while the first stage pre-trains the query adapter to extracts text-relevant information from visual features at a generic representation level, the second stage maps the visual tokens to the embedding space of the LLM.

for a vision encoder with $2m$ layers $\mathbf{E}_v = \{l_v^1, l_v^2, \ldots, l_v^{2m}\}$, we obtain each layer's outputs by the following:

$$\mathbf{V}_j = l_v^j(\mathbf{V}_{j-1}), \tag{1}$$

where $j \in \{1, 2, \ldots, 2m\}$, and the outputs at every alternate layer will be sent to the query adapter as the key and value of each according cross attention layer.

Instruction-aware Adaptation. The query adapter is consisted of several transformer layers, denoted as $\mathbf{E}_q = \{l_q^1, l_q^2, \ldots, l_q^m\}$, where the number of layers in it is half the number of layers in vision encoder. To obtain the most discriminative features based on different instructions, we design learnable query tokens to guide the adapter attending to instruction-relevant information. Meanwhile, the instruction is tokenized to fixed-length instruction tokens. Then, we concatenate the query tokens and instruction tokens to form the adapter input tokens $\mathbf{X}_0$. Each layer l_q^k of the query adapter is consisted of a self attention layer SA_k, a cross attention layer CA_k and a forward layer FFN_k. We obtain each layer's outputs by the following:

$$\mathbf{X}_k = FFN_k(CA_k(SA_k(\mathbf{X}_{k-1}), \mathbf{V}_{2k}, \mathbf{V}_{2k})), \tag{2}$$

where $k \in \{1, 2, \ldots, m\}$, and $\mathbf{V}_{2k}$ is the output of the $2k$-th layer in the vision encoder and serves as the key and value in the cross attention layer CA_k. The last layer output of the query tokens will be served as the visual tokens, which has the same length as the query tokens and will be sent to the LLM with the paired instruction.

Two-Stage Training Strategy

In order to better bridge the modality gap and improve training efficiency, we employ a two-stage training strategy: (1) multi-modal representation learning stage and (2) language-image instruction tuning stage. The first stage involves training on a general image-text dataset (Lin et al. 2014) to initially equip the query adapter with the capability to extract highly text-relevant information from visual features. The second stage, on the other hand, involves fine-tuning on the HumanIns dataset to enable the entire model to follow instructions while also being able to perform various human-centric tasks. Detailed elaboration is as follows.

Multi-modal Representation Learning. In this stage, we only train the query adapter and connect it to a frozen vision encoder and utilize a frozen text encoder to perform pre-training using image-text pairs in (Lin et al. 2014). Our goal is to train the query adapter such that the query tokens learn to extract visual features that is most informative of the text. Inspired by (Dai et al. 2023), we jointly optimize two objectives.

The first is *language-image contrastive learning*, denoted as $\mathcal{L}_{lic}$. For an image-text pair, the image is firstly fed into the vision encoder and query adapter to obtain the visual tokens. Then, the visual tokens are pooled to get the visual global embedding $\mathbf{V}_g$. The paired text is fed into the text encoder to obtain the textual global embedding $\mathbf{T}_g$. Then, given a batch of N image-text pairs, for each visual global embedding $\mathbf{V}_g^i$, we construct a set of visual-textual embedding pairs as $\{(\mathbf{V}_g^i, \mathbf{T}_g^j), y_{i,j}\}_{j=1}^N$, where $y_{i,j} = 1$ means that the pair is matched and $y_{i,j} = 0$ indicates the unmatched pair. Let $sim(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z} / \|\mathbf{x}\| \|\mathbf{z}\|$ denotes the similarity of $\mathbf{x}$ and $\mathbf{z}$. Then, the similarity of each pair is calculated by:

$$p_{i,j} = \frac{\exp\left(sim\left(\mathbf{V}_g^i, \mathbf{T}_g^j\right)\right)}{\sum_{k=1}^N \exp\left(sim\left(\mathbf{V}_g^i, \mathbf{T}_g^k\right)\right)}. \tag{3}$$

Then, the contrastive loss from vision to language in a batch can be computed by:

$$\mathcal{L}_{v2l} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{y_{i,j}=1}\log\left(p_{i,j}\right), \tag{4}$$

Similarly, the contrastive loss from language to vision $\mathcal{L}_{l2v}$ can be be computed by exchanging $\mathbf{V}_g$ and $\mathbf{T}_g$ in above equations. Finally, this task can be optimized by:

$$\mathcal{L}_{lic} = \mathcal{L}_{v2l} + \mathcal{L}_{l2v}. \tag{5}$$

The second is *vision-grounded masked language modeling*, denoted as $\mathcal{L}_{vmlm}$. The task requires utilizing original unmasked images to predict the masked words in textual descriptions. By exploiting the visual representations, it enhances the perception of context and strengthens the interaction between vision and language modality. For a pair of an image and a masked textual description with masked words $\mathbf{w}_m = \{\mathbf{w}_{m_1}, \ldots, \mathbf{w}_{m_M}\}$ (M is the number of masked words), we feed them to respective encoders to extract the visual global embedding $\mathbf{V}_g$ and textual hidden-state outputs $\mathbf{h}_t$. Then, we concatenate each masked location output of $\mathbf{h}_t$ and $\mathbf{V}_g$ as the preliminary multi-modal embeddings $\{\mathbf{h}_{m_i}\}_{i=1}^{M}$. For each multi-modal embedding representing the masked word, we use a prediction head to realize the corresponding probability prediction. It can be optimized by minimizing the negative log-likelihood:

$$\mathcal{L}_{vmlm} = -\frac{1}{MN}\sum^{N}\sum_{k=1}^{M}\log P(\mathbf{w}_{m_k}|\mathbf{h}_{m_k}), \tag{6}$$

where N denotes the number of samples within a batch and P denotes the probability distribution mapping.

Denoted λ as a hyper-parameter, the overall training objective is the weighted sum of the above objectives:

$$\mathcal{L} = \mathcal{L}_{lic} + \lambda\mathcal{L}_{vmlm}. \tag{7}$$

Language-Image Instruction Tuning. In this stage, we unfreeze the whole weights of the vision encoder (Radford et al. 2021), and continue to update the pre-trained weights of the query adapter and partial LLM (Zheng et al. 2023). Given a group of an image, an instruction and an answer, we perform instruction-tuning of the LLM on the answer tokens, using its original auto-regressive training objective. Specifically, we first concatenate the visual tokens and instruction tokens as the multi-modal instruction tokens $\mathbf{Z}_{ins}$. Denoting the answer tokens as $\mathbf{Z}_a = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\}$ with length L, we compute the probability of generating the target answer by:

$$p(\mathbf{Z}_a|\mathbf{Z}_{ins}) = \prod_{i=1}^{L} p_\theta\left(\mathbf{z}_i \mid \mathbf{Z}_{ins}, \mathbf{Z}_{a,<i}\right), \tag{8}$$

where θ is the trainable parameters in L-Man. Through this multi-modal instruction tuning , we can uniformly combine various human-centric tasks in the form of open-format instructions and answers.

Experiments

Implementation Details

Training Setups. L-Man consists of a pre-trained vision encoder, *i.e.*, CLIP-ViT-L/14 (Radford et al. 2021), a query adapter initialized by CLIP-Xformer (Radford et al. 2021) and an LLM Vicuna-7B (Zheng et al. 2023). The number of layers in the query adapter is half of the number of layers in the vision encoder. The number of query tokens is set to 128. More training details can be found in the supplement.

Evaluation. We have partitioned the training and testing sets for HumanIns to avoid data leakage issues. In the experiments, we investigate the general capability of L-Man on six human-centric tasks. More evaluation details are elaborated in the supplement.

Quantitative Results

Comparison With Other LMMs. We conduct a quantitative experiment to compare our model's instruction-following capability with other popular large multi-modal models (LMMs) (Dai et al. 2023; Liu et al. 2023b,a; Bai et al. 2023) in human-centric tasks. Due to the free form of responses in existing LMMs, there are currently no effective metrics to directly measure the performance of each model in downstream tasks. Therefore, we randomly select 120 images from each dataset's test split, and require each model to predict the answers based on the input images and task-relevant instructions. We utilize the ground truths to manually evaluate the quality of the answers. The details of chosen images, instructions and manual evaluation method are shown in the supplement. As the comparison results shown in Table 2, our L-Man has achieved significantly leading performance on all ten downstream human-centric datasets.

Comparison With Specialists. We also compare the performance of L-Man with the specialists on some downstream tasks (Li, Deng, and Du 2017; Soomro, Zamir, and Shah 2012; Yao et al. 2011; Liu et al. 2017). For expression recognition, as the results shown in Table 3, compared with some popular specialists, our model achieve competitive performance on RAF-DB dataset even without further fine-tuning. For action recognition, we compare the performance on UCF101 (Soomro, Zamir, and Shah 2012) and Stanford40 (Yao et al. 2011). As the results shown in Table 4, compared with all other specialists, L-Man achieves state-of-the-art performance on these two datasets even without further fine-tuning. For pedestrian attribute recognition, we compare the performance on PA100K (Liu et al. 2017). We adopt label-based metric mean accuracy (mA) and four instance-based metrics including accuracy, precision, racall rate and F1 value for evaluation. As the results shown in Table 5, L-Man achieves competitive performance on this dataset even without further fine-tuning. All these results demonstrate that L-Man shows remarkable capabilities in unifying human-centric tasks.

Qualitative Examples

Meanwhile, we have also conducted some qualitative experiments to intuitively evaluate the performance of our L-Man. We select two popular LMMs (Liu et al. 2023a; Bai

Method	LLM	Stanford40	UCF101	CityPersons	COCO	PA100K	RAF-DB	WiderFace	AIC	MPII
InstructBLIP	Vicuna-7B	83.3	73.3	30.0	41.6	46.7	30.0	40.0	-	-
InstructBLIP	Vicuna-13B	87.5	73.3	33.3	46.7	43.3	34.2	46.7	-	-
Qwen-VL	Qwen-7B	95.8	80.0	44.2	60.0	50.0	43.3	57.5	-	-
Qwen-VL-Chat	Qwen-7B	95.8	87.5	48.3	64.2	54.2	46.7	63.3	-	-
LLaVA	Vicuna-7B	93.3	83.3	48.3	60.0	54.2	70.0	67.5	-	-
LLaVA1.5	Vicuna-7B	95.8	90.0	54.2	70.0	57.5	67.5	73.3	-	-
LLaVA1.5	Vicuna-13B	95.8	93.3	56.7	76.7	67.5	70.0	70.0	-	-
L-Man	**Vicuna-7B**	**97.5**	**96.7**	**73.3**	**83.3**	**84.2**	**86.7**	**77.5**	**83.3**	**80.0**

Table 2: Comparison between L-Man and other popular LMMs on human-centric datasets. We show the best score in bold. For the human pose estimation task on AIC (Wu et al. 2017) and MPII (Andriluka et al. 2014), other LMMs are unable to accomplish this task due to a lack of relevant knowledge, which is represented by "-" in the table.

Method	RAF-DB							
	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Average
ImNet (2021a)	78.00	54.00	57.00	96.00	88.00	85.00	85.00	77.57
MVT (2021)	78.40	63.75	60.81	95.61	89.12	87.45	87.54	80.38
VTFF (2021)	85.80	68.12	64.86	94.09	87.50	87.24	85.41	81.86
SCAN-CCI (2021b)	81.00	70.00	66.00	**96.00**	89.00	86.00	88.00	82.29
Transfer (2021)	88.89	79.37	68.92	95.95	90.15	88.70	**89.06**	85.86
L-Man	**89.19**	**81.88**	**69.76**	95.17	**91.89**	**89.03**	87.23	**86.31**

Table 3: Comparison between L-Man and some SoTA specialists on RAF-DB dataset. We show the best score in bold.

Method	UCF101	Method	Stanford40
SpeedNet (2020)	81.1	PoseG (2022)	89.5
VTHCL (2020)	82.1	LossG (2019)	91.1
MemDPC (2020a)	86.1	SCLAR (2022)	92.9
CoCLR (2020b)	87.9	SAAM (2020)	93.0
RSPNet (2021)	93.7	Relation (2020)	93.1
CVRL (2021)	94.4	PoseE (2022)	93.2
GDT (2020)	95.2	BodyS (2020)	93.8
VideoMAE V1 (2022)	96.1	MultiA (2021)	94.2
VideoMAE V2 (2023)	99.6	Interaction (2023)	94.8
L-Man	**99.6**	**L-Man**	**94.9**

Table 4: Comparison between L-Man and some state-of-the-art specialists on action recognition datasets.

Method	PA-100K				
	mA	Accu	Prec	Recall	F1
LGNet (2018)	76.96	75.55	86.99	83.17	85.04
PGDM (2018)	74.95	73.08	84.36	82.24	85.04
ALM (2019)	80.68	77.08	84.24	88.84	86.46
PDNet (2020)	80.40	78.80	87.50	86.91	87.20
JLAC (2020)	82.31	79.47	87.45	87.77	87.61
DAFL (2022)	83.54	80.13	87.01	89.19	88.09
VTB (2022)	**83.72**	**80.89**	87.88	89.30	88.21
L-Man	82.26	79.98	**88.26**	**89.74**	**88.71**

Table 5: Comparison between L-Man and some popular specialists on PA100K dataset.

et al. 2023). Each model is required to respond according to the input image and instruction. As the qualitative example shown in the left part of Figure 5, L-Man is capable of better perceiving fine-grained appearance features related to pedestrians and generating more accurate descriptions. In contrast, other models tend to produce irrelevant descriptions, such as detailing background information. Moreover, for the popular two tasks, pedestrian detection and human pose estimation, the qualitative examples of L-Man are presented in the right part of Figure 5. It can be observed that our L-Man can effectively accomplish the responding different tasks based on different instructions and provide accurate answers.

Ablation Study

Number of Query Tokens. In our method, query tokens are utilized to map the extracted visual information into the textual space of LLM. The number of query tokens significantly affects not only the effectiveness of this multi-modal mapping but also the training costs. Too few query tokens cannot effectively extract rich visual information and results in poor performance, while too many query tokens can easily introduce unnecessary noises and increase training costs. Therefore, we conduct ablation studies to explore the impact of different numbers of query tokens on the model's performance. As the results shown in Figure 6, the model obtains best performance on both RAF-DB (Li, Deng, and Du 2017) and PA-100K (Liu et al. 2017) with a query token number of 128, achieving 86.3% and 80.0% accuracy, respectively. Therefore, this setup, which effectively balances the model performance and training costs, is used as the default setting in all experiments.

Effectiveness of Each Component. We train the models under different settings and directly evaluate their performance on the PA-100K dataset (Liu et al. 2017). The results are reported in Table 6. $\mathcal{L}_{lic}$ denotes the language-image contrastive learning objective utilized in the multi-modal representation learning stage. $\mathcal{L}_{vmlm}$ is vision-grounded masked language modeling objective utilized in the multi-modal representation learning stage. mg denotes the multi-grained visual feature extraction module designed in the mechanism of

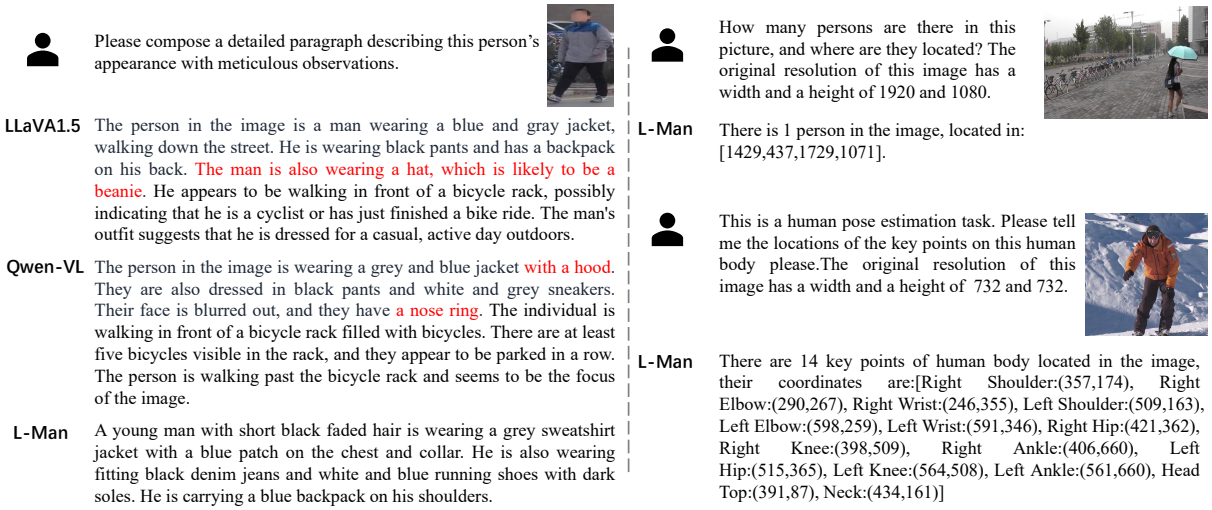

Figure 5: Qualitative examples demonstrating L-Man's capability on person attribute recognition, pedestrian detection and human pose estimation task. L-Man generates more fine-grained and accurate responses.

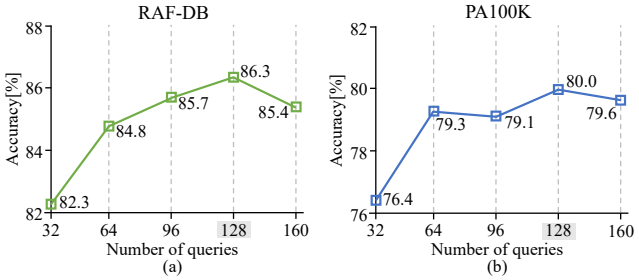

Figure 6: Ablation studies on the number of query tokens.

No.	Components				PA-100K				
	$\mathcal{L}_{lic}$	$\mathcal{L}_{vmlm}$	mg	ins	mA	Accu	Prec	Recall	F1
0					76.62	76.38	85.21	82.19	82.32
1			✓		78.76	78.82	86.84	85.43	85.67
2				✓	78.12	78.91	86.64	84.98	85.13
3			✓	✓	79.82	79.10	87.67	87.15	86.63
4	✓				77.13	77.87	86.46	84.68	84.55
5	✓		✓		79.64	78.88	87.63	86.72	86.27
6	✓			✓	78.56	79.12	87.24	85.88	86.23
7	✓		✓	✓	81.34	79.62	87.96	88.78	87.97
8		✓			76.93	76.52	85.44	83.31	83.88
9		✓	✓		78.96	78.45	86.72	85.65	85.33
10		✓		✓	78.55	79.32	86.66	85.27	85.63
11		✓	✓	✓	80.24	79.27	87.63	88.27	87.32
12	✓	✓			79.68	78.96	87.41	86.42	85.78
13	✓	✓	✓		81.42	79.31	87.80	88.81	87.84
14	✓	✓		✓	81.37	79.45	87.68	88.92	88.21
15	✓	✓	✓	✓	**82.26**	**79.98**	**88.26**	**89.74**	**88.71**

Table 6: Ablation study on each component of L-Man. $\mathcal{L}_{lic}$ and $\mathcal{L}_{vmlm}$ are the two objectives in the first training stage. mg and ins denote the multi-grained extraction and instruction-aware adaptation mechanism in query adapter, respectively.

the query adapter. ins denotes the instruction-aware adaption mechanism, which means whether to concatenate the instruction tokens to the query tokens. The baseline method No.0 does not employ the four strategies mentioned above. This implies that its query adapter does not undergo pre-training in the first stage; instead, it is initialized randomly and used for instruction-tuning in the second stage. Additionally, it does not utilize features from different layers of the vision encoder, relying solely on the last layer's features. Furthermore, the instruction tokens of No.0 are not input into the query adapter to extract instruction-aware visual information. From the table, it can be observed that each individual component significantly contributes to the model's performance in the downstream task. Combining all components together (No.15) achieves optimal performance, reaching 82.26% mAP, 79.98% accuracy, 88.26% precision, 89.74% recall, and 88.71% F1 value on the PA100K dataset.

Conclusion

In this paper, we investigate the opportunities and challenges in utilizing large language models to unify various human-centric tasks, and attempt to propose a large multi-modal model unifying human-centric tasks with an open-ended form. Firstly, we construct a new language-image instruction-following dataset named HumanIns with the existing publicly available datasets. Then, we propose a novel large multi-modal model for unifying various human-centric tasks in an unrestricted form of language instructions. A range of experimental results demonstrate the superiority of our proposed dataset and method in related research domains. We hope our work can facilitate the researches on unifying various human-centric tasks in an unrestricted instruction-following form.

Acknowledgements

This work was supported by the National Natural Science Foundation of China No.62176097, and the Hubei Provincial Natural Science Foundation of China No.2022CFA055.

We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis.

Ashrafi, S. S.; Shokouhi, S. B.; and Ayatollahi, A. 2021. Action recognition in still images using a multi-attention guided network with weakly supervised saliency detection. *Multimedia Tools and Applications*, 80: 32567–32593.

Ashrafi, S. S.; Shokouhi, S. B.; and Ayatollahi, A. 2023. Still image action recognition based on interactions between joints and objects. *Multimedia Tools and Applications*, 82(17): 25945–25971.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Benaim, S.; Ephrat, A.; Lang, O.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; Irani, M.; and Dekel, T. 2020. Speednet: Learning the speediness in videos. In *CVPR*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*.

Chen, P.; Huang, D.; He, D.; Long, X.; Zeng, R.; Wen, S.; Tan, M.; and Gan, C. 2021. Rspnet: Relative speed perception for unsupervised video representation learning. In *AAAI*.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE TCSVT*, 32(10): 6994–7004.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Ci, Y.; Wang, Y.; Chen, M.; Tang, S.; Bai, L.; Zhu, F.; Zhao, R.; Yu, F.; Qi, D.; and Ouyang, W. 2023. UniHCP: A Unified Model for Human-Centric Perceptions. In *CVPR*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dehkordi, H. A.; Nezhad, A. S.; Kashiani, H.; Shokouhi, S. B.; and Ayatollahi, A. 2022. Multi-expert human action recognition with hierarchical super-class learning. *Knowledge-Based Systems*, 250: 109091.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Dollar, P.; Wojek, C.; Schiele, B.; and Perona, P. 2011. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI*, 34(4): 743–761.

Gera, D.; and Balasubramanian, S. 2021a. Imponderous net for facial expression recognition in the wild. *arXiv preprint arXiv:2103.15136*.

Gera, D.; and Balasubramanian, S. 2021b. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *Pattern Recognition Letters*, 145: 58–66.

Han, T.; Xie, W.; and Zisserman, A. 2020a. Memory-augmented dense predictive coding for video representation learning. In *ECCV*.

Han, T.; Xie, W.; and Zisserman, A. 2020b. Self-supervised co-training for video representation learning. In *NeurIPS*.

Jain, V.; and Learned-Miller, E. 2010. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report.

Jia, J.; Gao, N.; He, F.; Chen, X.; and Huang, K. 2022. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *AAAI*.

Li, D.; Chen, X.; Zhang, Z.; and Huang, K. 2018. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios. In *ICME*.

Li, H.; Sui, M.; Zhao, F.; Zha, Z.; and Wu, F. 2021. MVT: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*.

Li, S.; Deng, W.; and Du, J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person search with natural language description. In *CVPR*.

Li, Y.; Li, K.; and Wang, X. 2020. Recognizing actions in images by fusing multiple body structure cues. *PR*, 104: 107341.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, L.; Tan, R. T.; and You, S. 2019. Loss guided activation for action recognition in still images. In *ACCV*.

Liu, P.; Liu, X.; Yan, J.; and Shao, J. 2018. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*.

Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; and Wang, X. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*.

Liu, Y.; Tian, M.; Hou, J.; Yi, S.; and Lin, Z. 2020. Pentadent-net: Pedestrian attribute recognition with distance refinement and correlation mining. In *ICIP*.

Ma, F.; Sun, B.; and Li, S. 2021. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*.

Ma, W.; and Liang, S. 2020. Human-object relation network for action recognition in still images. In *ICME*.

Mi, S.; and Zhang, Y. 2022. Pose-guided action recognition in static images using lie-group. *Applied Intelligence*, 1–9.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.

Patrick, M.; Asano, Y.; Kuznetsova, P.; Fong, R.; Henriques, J. F.; Zweig, G.; and Vedaldi, A. 2020. Multi-modal self-supervision from generalized data transformations.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal contrastive video representation learning. In *CVPR*.

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 *ICML*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; and Tomasi, C. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; and Li, S. Z. 2020. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *AAAI*.

Tang, C.; Sheng, L.; Zhang, Z.; and Hu, X. 2019. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *ICCV*.

Tang, S.; Chen, C.; Xie, Q.; Chen, M.; Wang, Y.; Ci, Y.; Bai, L.; Zhu, F.; Yang, H.; Yi, L.; et al. 2023. HumanBench: Towards General Human-centric Perception with Projector Assisted Pretraining. In *CVPR*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, J.; and Liang, S. 2022. Pose-enhanced relation feature for action recognition in still images. In *ICMM*.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *CVPR*.

Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.; et al. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *CVPR*.

Xue, F.; Wang, Q.; and Guo, G. 2021. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*.

Yang, C.; Xu, Y.; Dai, B.; and Zhou, B. 2020. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*.

Yang, S.; Luo, P.; Loy, C.-C.; and Tang, X. 2016. Wider face: A face detection benchmark. In *CVPR*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *ICCV*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2018. From facial expression recognition to interpersonal relation prediction. *IJCV*, 126: 550–569.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *ICCV*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *CVPR*.

Zheng, Y.; Zheng, X.; Lu, X.; and Wu, S. 2020. Spatial attention based visual semantic learning for action recognition in still images. *Neurocomputing*, 413: 383–396.

Zuo, J.; Hong, J.; Zhang, F.; Yu, C.; Zhou, H.; Gao, C.; Sang, N.; and Wang, J. 2024a. PLIP: Language-Image Pre-training for Person Representation Learning. In *NeurIPS*.

Zuo, J.; Nie, Y.; Zhou, H.; Zhang, H.; Wang, H.; Guo, T.; Sang, N.; and Gao, C. 2024b. Cross-video Identity Correlating for Person Re-identification Pre-training. In *NeurIPS*.

Zuo, J.; Zhou, H.; Nie, Y.; Zhang, F.; Guo, T.; Sang, N.; Wang, Y.; and Gao, C. 2024c. UFineBench: Towards Text-based Person Retrieval with Ultra-fine Granularity. In *CVPR*.