Pedestrian Attribute Recognition: A New Benchmark Dataset and A Large Language Model Augmented Framework

Jiandong Jin[2,3], Xiao Wang[2*], Qian Zhu[2], Haiyang Wang[2], Chenglong Li[1,3*]

[1]School of Artificial Intelligence, Anhui University, Hefei 230601, China
[2]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[3]Anhui Provincial Key Laboratory of Security Artificial Intelligence, Hefei 230601, China
{jdjinahu, wangxiaocvpr, lcl1314}@foxmail.com, {zq542664, why2434961256}@163.com

Abstract

Pedestrian Attribute Recognition (PAR) is one of the indispensable tasks in human-centered research. However, existing datasets neglect different domains (e.g., environments, times, populations, and data sources), only conducting simple random splits, and the performance of these datasets has already approached saturation. In the past five years, no large-scale dataset has been opened to the public. To address this issue, this paper proposes a new large-scale, cross-domain pedestrian attribute recognition dataset to fill the data gap, termed MSP60K. It consists of 60,122 images and 57 attribute annotations across eight scenarios. Synthetic degradation is also conducted to further narrow the gap between the dataset and real-world challenging scenarios. To establish a more rigorous benchmark, we evaluate 17 representative PAR models under both random and cross-domain split protocols on our dataset. Additionally, we propose an innovative Large Language Model (LLM) augmented PAR framework, named LLM-PAR. This framework processes pedestrian images through a Vision Transformer (ViT) backbone to extract features and introduces a multi-embedding query Transformer to learn partial-aware features for attribute classification. Significantly, we enhance this framework with LLM for ensemble learning and visual feature augmentation. Comprehensive experiments across multiple PAR benchmark datasets have thoroughly validated the efficacy of our proposed framework.

Code&Dataset —
https://github.com/Event-AHU/OpenPAR

1 Introduction

Pedestrian Attribute Recognition (PAR) (Wang et al. 2022) has been widely exploited in the Computer Vision (CV) and Artificial Intelligence (AI) community. It aims to map the given pedestrian image into semantic labels, such as *gender*, *hairstyle*, and *wearings*, using deep neural networks and achieves high performance on current benchmark datasets. These models can be employed in practical scenarios and may work well in simple scenarios. It can also help other human-centric tasks, e.g., pedestrian detection and tracking (Li et al. 2024), person re-identification (Lin et al. 2019) and retrieval (Huang et al. 2024). However, the performance of the current PAR model is still significantly affected by challenging factors (e.g., low illumination, motion blur, and complex backgrounds); moreover, there is still much room for exploration in the relationship between pedestrian image perception and multi-label attributes.

Considering these issues, we meticulously review the existing works and datasets on PAR and find that the development in the PAR field has begun to enter a bottleneck period. As an effective driving force for promoting the development of PAR, benchmark datasets play a crucial role. However, we believe that the PAR community needs to address several core issues on the benchmark datasets as follows: **1).** The performance of existing pedestrian attribute recognition datasets is *close to saturation*, and the performance improvement of new algorithms has shown a trend of weakening. However, only one small-scale PAR-related dataset has been released in the past five years, thus, there is an urgent need for new large-scale datasets to support new research endeavors. **2).** Existing PAR datasets use random partitioning for model training and testing, which can measure the overall recognition capability of a PAR model. However, this partitioning mechanism overlooks the impact of *cross-domain* (e.g., different environments, times, populations, and data sources) on the PAR model. **3).** Existing PAR datasets do not prominently reflect challenge factors, thus, this may potentially result in neglecting the impact of *data corruption* during real-world application, thereby introducing safety hazards in practical settings. In conclusion, it is evident that the PAR community urgently requires a new large-scale dataset to bridge the existing data gap.

In this paper, we propose a new benchmark dataset for pedestrian attribute recognition, termed **MSP60K**, as shown in Fig. 1. It contains 60,122 images, and over 5,000 person IDs, collected using smart surveillance systems and mobile phones. To make our dataset better reflect the challenges found in real-world scenarios, in addition to annotating as many complex images as possible, we also process these images using additional destructive operations, including blur, occlusion, illumination, adding noise, jpeg compression, etc. As these images belong to different domains and scenarios, such as supermarket, kitchen, construction site, ski resort, and various outdoor scenes, we split these images according to two protocols, i.e., *random split* and *cross-domain split*.

*Corresponding Author

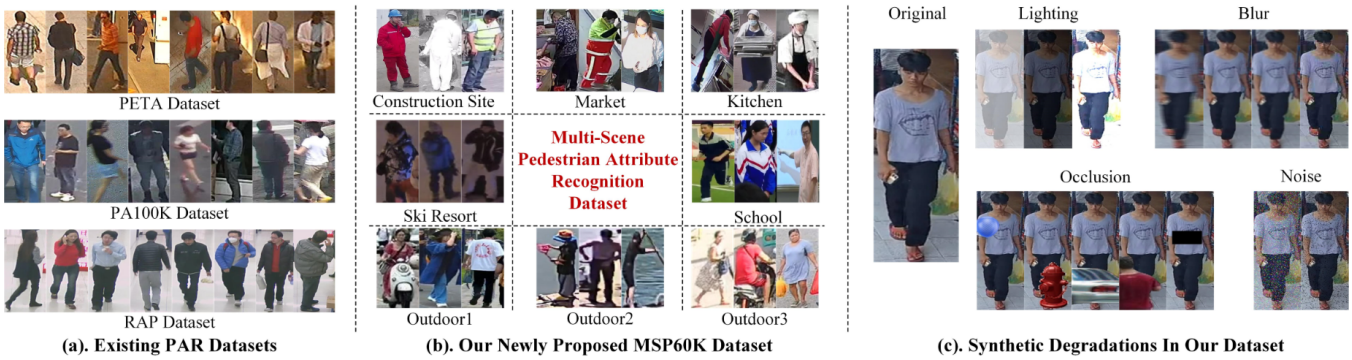

Figure 1: (a, b). Comparison between existing PAR datasets and our newly proposed MSP60K dataset. (c). Illustrates the synthetic degradation challenges we employed in our dataset to simulate the complex and dynamic real-world environment.

Therefore, the newly proposed benchmark dataset can better validate the performance of PAR models in real-world scenarios, especially under cross-domain settings. To create a more thorough benchmark for PAR, we assess 17 representative and recently developed PAR algorithms on our dataset using random and cross-domain protocols. These benchmark comparison methods can better facilitate the subsequent verification and experimentation of future PAR models.

Based on our newly proposed MSP60K PAR dataset, we also propose a novel large language model (LLM) augmented pedestrian attribute recognition framework, termed LLM-PAR. Based on the widely used multi-label classification framework, we rethink the relationship between pedestrian image perception and large language models as the key insight of this work. As we all know, large language models possess powerful abilities in text generation, comprehension, and reasoning. Therefore, we introduce a large language model, which generates textual descriptions of the image's attributes as an auxiliary task based on a multi-label classification framework. This LLM branch serves a dual purpose: on the one hand, it can assist in the learning of visual features through the generation of accurate textual descriptions, thereby achieving high-performance attribute recognition; on the other hand, the LLM can facilitate effective interaction between visual features and prompts. The output text tokens can also be integrated with the aforementioned multi-label classification framework for ensemble learning.

As shown in Fig. 3, our proposed LLM-PAR can be divided into two main modules, i.e., the standard multi-label classification branch and the large language model augmentation branch. Specifically, we first partition the given pedestrian image into patches and project them into visual embeddings. Then, a visual encoder with LoRA (Hu et al. 2022) is utilized for global feature learning and a **M**ulti-**E**mbedding **Q**uery Trans**Former** (MEQ-Former) is proposed for part-aware feature learning. After that, we adopt CBAM (Woo et al. 2018) attention modules to merge the output tokens and feed them into MLP (Multi-Layer Perceptron) layers for attribute classification. More importantly, we concatenate the part-aware visual tokens with the instruction prompt and feed them into the large language model for pedestrian attribute description. The text tokens are also fed into an attribute recognition head and ensembles with classification logits. Extensive experiments on our newly proposed MSP60K dataset and other widely used PAR benchmark datasets all validated the effectiveness of our proposed LLM-PAR.

To sum up, we draw the main contributions of this paper as the following three aspects:

1). We propose a new benchmark dataset for pedestrian attribute recognition, termed MSP60K, which contains 60122 images, over 5,000 IDs, and fully reflects the key challenges in real-world scenarios. We benchmark 17 PAR algorithms on the MSP60K dataset and hope that the introduction of this benchmark dataset can better promote the development and practical deployment of PAR models.

2). We propose a novel large language model (LLM) augmented PAR algorithm, termed LLM-PAR, based on the standard multi-label classification framework. The introduction of the LLM branch enables PAR to better leverage its reasoning capabilities, achieving enhanced visual feature representation and model integration.

3). Extensive experiments conducted on our newly proposed MSP60K dataset and other PAR datasets fully demonstrate the effectiveness of our proposed PAR model. New state-of-the-art performances are achieved on multiple PAR datasets, e.g., 92.25/90.39 on mA/F1 metric on the PETA (Deng et al. 2014) dataset, 91.09/90.41 on PA100K (Liu et al. 2017).

2 Related Work

Pedestrian Attribute Recognition

Pedestrian attribute recognition (Wang et al. 2022) aims to classify pedestrian images based on a predefined set of attributes. Current methods can be broadly categorized into attention-based, viewpoint-guidance, and visual-language modeling approaches. Due to the strong correlation between pedestrian attributes and specific body components, various methods, such as HPNet (Liu et al. 2017) and DAHAR (Liu et al. 2017; Jia et al. 2022) focused on localizing

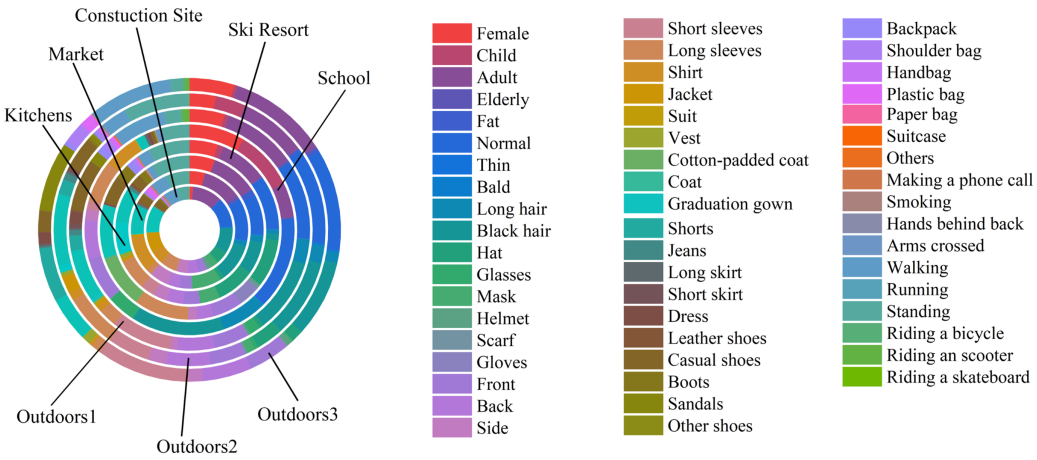

Figure 2: Attributes Distribution in Different Scenes. Circular chart illustrating attribute distribution across scenes.

attribute-relevant regions via attention mechanisms. Some researchers (Lu et al. 2023; Shen et al. 2024) model the connections of attribution in different views to address the challenge of changing posture and view. Furthermore, pedestrian attributes are closely interconnected. Consequently, JLAC (Tan et al. 2020) and PromptPAR (Wang et al. 2024a) jointly model attribute context and image-attribute relationships. While current methods recognize the importance of exploring contextual relationships in the PAR task, leveraging models like Transformers to capture attribute relationships within datasets often struggles to represent connections involving rare attributes.

Benchmark Datasets for PAR

The most commonly used datasets of PAR are PETA (Deng et al. 2014), WIDER (Li et al. 2016b), RAP (Li et al. 2016a; Li et al. 2019), and PA100K (Liu et al. 2017). To enhance the ability to recognize pedestrian attributes at a long distance, Deng et al. (Deng et al. 2014) introduced a new pedestrian attribute dataset named PETA, labeling over 60 attributes. Unlike PETA's identity-level annotation, the RAP dataset captures an indoor shopping mall and employs instance-level annotation for the pedestrian images. Liu et al. (Liu et al. 2017) proposed a large pedestrian attribute recognition dataset, PA100K, with 100,000 images and 26 attributes, addressing the issue of information leakage by avoiding overlap between training and test sets. However, these datasets only contain simple scenes with limited background variation and lack significant style changes among pedestrians.

Vision-Language Models

With the rapid development of the natural language processing field, many large language models (LLMs) such as Flan-T5 (Longpre et al. 2023), and LLaMA (Touvron et al. 2023) have emerged. Although notable foundational models like SAM (Kirillov et al. 2023) and CLIP (Radford et al. 2021) have been introduced in the vision domain, the complexity of visual tasks has hindered the development of generalized multi-domain visual models. Some researchers have begun to view LLMs as world models, leveraging them as the cognitive core to enhance various multi-modal tasks. Recognizing the high cost of training a large multi-modal model from scratch, BLIP series (Li et al. 2022a, 2023), MiniGPT-4 (Zhu et al. 2024), bridge existing pre-trained visual models and large language models. Although these models have significant improvements in the field of vision understanding and text generation, there are many challenges, such as low-resolution image recognition, fine-grained image caption, and the hallucination of LLMs.

3 MSP60K Benchmark Dataset

Protocols

To provide a robust platform for training and evaluating pedestrian attribute recognition (PAR) in real-world conditions, we adhere to these guidelines while constructing the MSP60K benchmark dataset: *1). Large Scale:* We annotate 60,122 pedestrian images, each with 57 attributes, comprehensively analyzing pedestrian characteristics in various conditions. *2). Multiple Distances and Viewpoints:* Images are captured from different angles and distances using various cameras and handheld devices, covering the front, back, and side views. The resolution of pedestrian images in our dataset is from 30×80 to 2005×3008. *3). Complex and Varied Scenes:* Unlike existing datasets with uniform backgrounds, our dataset includes images from eight different environments with diverse backgrounds and attribute distributions, helping evaluate recognition methods in varied settings. *4). Rich Source of Pedestrian Identity:* We gather data on pedestrians from different scenarios, nationalities, and seasonal variations, enhancing the dataset with diverse styles and characteristics. *5). Simulated Complex Real-world Environments:* The dataset includes variations in lighting, motion blur, occlusions, and adverse weather conditions, simulating real-world challenges in pedestrian attribute recognition.

Attribute Groups and Details

To effectively evaluate the performance of existing PAR methods in complex scenarios, each image in our dataset is labeled with 57 attributes, which are categorized into 11 groups: gender, age, body size, viewpoint, head, upper body, lower body, shoes, bag, body movement, and sports information. The complete list of the defined attributes can be found in Supplementary Material.

Statistical Analysis

Our MSP60K offers 8 distinct scenes and 57 attributes, providing richer annotations than datasets like PA100K (26 attributes) and WIDER (14 attributes). The dataset comprises 60,122 images of over 5,000 unique individuals. It includes varied environments such as markets, schools, kitchens, ski resorts, and various outdoor and construction sites, offering a broader scope than other datasets.

In our benchmark dataset, we split the data using the random and cross-domain partitioning strategies:

• **Random Partitioning**: 30,298 images for training, 6,002 for validation, and 23,822 for testing, ensuring a random distribution of scenes like other PAR benchmark datasets.

• **Cross-domain Partitioning:** To validate domain generalization and zero-shot performance of PAR models, we divide our dataset based on scenarios, i.e., five scenarios (*Construction Site, Market, Kitchens, School, Ski Resort*) with 34,128 images are used for training, while three scenarios (*Outdoors1, Outdoors2, Outdoors3*) with 24,994 images are used for testing.

To assess the robustness of the model, we intentionally degrade 1/3 of the images in each subset by introducing variations such as changes in *lighting*, *random occlusions*, *blurring*, and *noise*. With its extensive size and diverse conditions, MSP60K offers a comprehensive platform for evaluating PAR methods.

4 Methodology

Overview

This paper introduces a method for improving pedestrian attribute recognition (LLM-PAR) using multi-modal large language models (MLLMs) that describe the image in detail. As shown in Fig. 3, we leverage MLLMs to explore the contextual relationships between attributes, generating descriptions that assist attribute recognition. The approach consists of three main modules: 1) a multi-label classification branch, 2) a large language model branch, and 3) model aggregation. Specifically, we first extract the visual features of pedestrians using a visual encoder. Then, we design MEQ-Former to extract specific features for different attribute groups and translate to the latent space of MLLMs, improving the ability of MLLMs to capture fine details of pedestrians. The attribute group features are integrated into instruction embedding via a projection layer, the features feed into the large language model to generate pedestrian captions. Finally, the classification results from the visual features of each group are aggregated with the results from the language branch to produce the final classification results. The following sections will provide a detailed introduction to these modules.

Multi-Label Classification Branch

Given an input pedestrian image $I \in \mathbb{R}^{H\times W\times 3}$, as shown in Fig. 3, we first partition it into patches and project them into visual tokens. The visual tokens are added with Position Embedding (P.E.) which encodes the spatial information. The output will be fed into a visual encoder (EVA-ViT-G (Fang et al. 2023) is adopted for default) to extract the global visual representation F_V. In our implementation, we freeze the parameters of the pre-trained visual encoder and adopt LoRA (Hu et al. 2022) to achieve efficient tuning. Then, a newly designed Multi-Embedding Query Transformer (MEQ-Former) which extracts specific features from different attribute groups derived from primary visual features. Here, the attribute groups are obtained by categorizing the attributes into groups $\{A^j \mid j = 0, 1, \ldots, K\}$, based on their type, such as *head*, *upper body clothing*, *actions*, where K denotes the number of attribute groups.

As shown in Fig. 3, we create K sets of Partial Query (PartQ) $Q_p \in \mathbb{R}^{K\times L\times D}$, where L and D are the number and dimension of the queries, respectively. These embeddings are fed into the Attributes Group Features Aggregate (AGFA) module to extract specific features $F_g = \{F_g^1, F_g^2, ..., F_g^K\}$ for different attribute groups. The AGFA module consists of stacked Feed-Forward Networks (FFN) and Cross-Attention (CrossAttn) layers. This process can be formulated as:

$$F_g = FFN(CrossAttn(Q = Q_p, K = F_V, V = F_V)) \tag{1}$$

The F_g is fed into the Q-Former E_Q, which serves as a bridge between the visual and language modalities, to generate text-related information F_q^j. Q-Former comprises stacked self-attention and cross-attention layers, and aggregates image information through cross-attention mechanisms. Then, we introduce the Convolutional Block Attention Modules (CBAM) (Woo et al. 2018) to capture fine-grained features for each attribute from the F_g to produce attribute-specific predictions in the attribute-level classifiers, and we propose instance-level classifiers that share CBAM to aggregate features within groups to allow rare attributes to benefit from common ones.

Large Language Model Branch

Although this multi-label classification framework can achieve decent accuracy, it still fails to consider the logical reasoning of large language models, which is evident in the image-text domain. Therefore, this paper attempts to use LLM as an auxiliary branch to enhance pedestrian attribute recognition. As shown in Fig. 3, we first build the instructions based on each attribute group A^j. Then, we adopt the *Tokenizer* (Zheng et al. 2023b) to get the instruction embeddings $T_E = \{T_E^1, T_E^2, \ldots, T_E^{k+1}\}$ and concatenate them with visual features F_q of the human image as the instruction features F_I. Note that, we embed the ground truth and concatenate it with F_I as the initial input of the LLM during the training phase. The Vicuna-7B (Zheng et al. 2023b) and OPT-6.7B (Liu et al. 2021) are exploited as the LLM and also tuned using LoRA in our experiments. Finally, we

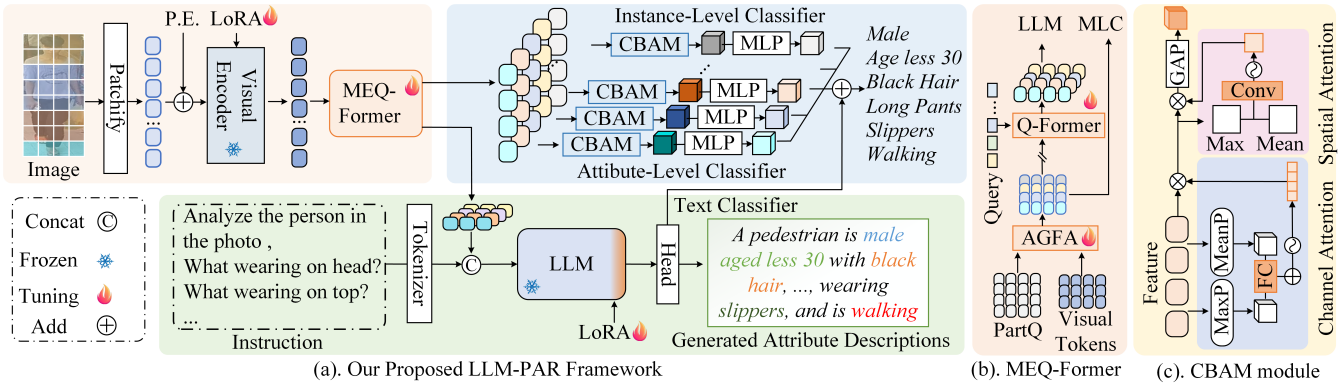

Figure 3: An illustration of our proposed LLM-PAR framework illustrates how we use Multimodal Large Language Models (MLLMs) for deep semantic reasoning, combining images and descriptive text to provide more interpretable visual understanding. Through this framework, we can recognize pedestrian attributes and generate natural language descriptions, thereby offering more intuitive explanations. Our framework consists of three parts: visual feature extraction, language description generation, and language-enhanced classification.

get the last hidden state from MLLM and the corresponding image captions through the Language Model Head.

Model Aggregation for PAR

After being equipped with the LLM, our algorithmic framework can simultaneously output pedestrian attribute results and complete text passages to describe the attributes of a given pedestrian. To leverage the strengths of these two branches, we have designed an algorithm integration module to achieve enhanced prediction results. As shown in Fig. 3, we define two visual classifiers for attribute recognition, i.e., the attribute-level and instance-level classifiers. We also get the classifier for recognition using tokens from the large language model branch.

In our implementation, we exploit the following three strategies to fuse these three results as ours. Specifically, *1). Attributes-Specific Aggregation (ASA)*: we adaptively weight and sum the attribute predictions of each classifier based on the weights learned from the training subset. *2). Mean Pooling*: We directly take the average of the results from the two branches as the final model output. *3). Max Pooling*: We take the maximum value of the logits from the three prediction branches as the final prediction result. Note that, we adopt the *Mean Pooling* strategy as the default setting in our experiments if not otherwise specified. More detailed results can be found in our experiments.

Loss Function

In the training phase, we adopt the widely used weighted cross-entropy loss (WCE Loss) $\mathcal{L}_{wce}(\cdot)$ (Li, Chen, and Huang 2015) for attribute prediction branches, i.e.,

$$\mathcal{L}_{MLC} = \mathcal{L}_{wce}(\hat{y}, P_{attr}) + \mathcal{L}_{wce}(\hat{y}, P_{in}) \quad (2)$$

We also adopt cross-entropy loss $\mathcal{L}_{ce}(\cdot)$ for the captioning generation in the LLM branch.

$$\mathcal{L}_{LLM} = \mathcal{L}_{wce}(\hat{y}, P_{llm}) + \mathcal{L}_{ce}(\hat{y}_{cap}, P_{cap}) \quad (3)$$

where $\hat{y}$ and $\hat{y}_{cap}$ denote the ground-truth labels and corresponding pedestrian attribute description, respectively. The P_{cap} is the logits generated by the Large Language Model Head.

5 Experiments

Datasets and Evaluation Metric

In this study, we conduct a comprehensive benchmark of 17 pedestrian attribute recognition methods, which represent the most important models in the field of pedestrian attribute recognition. Furthermore, the performance of our methods is compared with existing state-of-the-art (SOTA) PAR methods in our own benchmark and in three publicly available datasets: **PETA** (Deng et al. 2014), **PA100K** (Liu et al. 2017) and **RAPv1** (Li et al. 2016a). Five widely used evaluation metrics are employed for evaluating the performance, including: **mean Accuracy** (mA), **Accuracy** (Acc), **Precision** (Prec), **Recall** and **F1-score** (F1). More details about these evaluation metrics can be found in our supplementary materials.

Comparison on Public PAR Benchmarks

• **Result on MSP60K Dataset.** We collect and analyze public PAR methods from 2015 to 2024 on the MSP60K dataset. As shown in Table 1, methods like HAP (Yuan et al. 2024), and PARformer (Fan et al. 2023), which perform well in the random split, experience significant drops in performance in the cross-domain split. For instance, mA, Acc, and F1 of HAP scores drop by 18.22, 25.53, and 19.20, respectively. Some methods show smaller declines in the cross-domain split, with PromptPAR (Wang et al. 2024a) achieving state-of-the-art results, though still with notable decreases. We also tested MiniGPT-4 (Zhu et al. 2024) in a zero-shot setup on our dataset, with significant drops observed in the cross-domain split. After optimizations, LLM-PAR achieved 80.13, 78.71, 84.39, 90.52, and 86.94 in the random split, and 66.29, 58.11, 65.28, 81.21, and 72.05 in

Methods	Publish	Random Split					Cross-domain Split				
		mA	Acc	Prec	Recall	F1	mA	Acc	Prec	Recall	F1
#01 DeepMAR (Li, Chen, and Huang 2015)	ACPR15	70.46	72.83	84.71	81.46	83.06	54.84	44.97	63.38	58.81	61.01
#02 Strong Baseline (Jia et al. 2021)	-	74.09	73.74	84.06	83.51	83.31	55.91	46.25	63.28	61.34	61.64
#03 RethinkingPAR (Jia et al. 2021)	arXiv20	74.01	74.20	84.17	83.94	84.06	55.98	46.52	62.85	62.09	62.47
#04 SSCNet (Jia, Chen, and Huang 2021)	ICCV21	69.71	69.31	79.22	82.47	80.82	52.84	40.88	56.26	58.64	57.43
#05 VTB (Cheng et al. 2022)	TCSVT22	76.09	75.36	83.56	86.46	84.56	58.59	49.81	65.11	66.11	65.00
#06 Label2Label (Li et al. 2022b)	ECCV22	73.61	72.66	81.79	84.32	82.56	56.38	45.81	59.67	64.20	61.19
#07 DFDT (Zheng et al. 2023a)	EAAI22	74.19	76.35	**85.03**	86.35	85.69	57.85	49.97	65.34	66.18	65.76
#08 Zhou et al. (Zhou et al. 2023)	IJCAI23	73.07	68.76	78.38	82.10	80.20	54.26	41.91	56.23	60.11	58.11
#09 PARFormer (Fan et al. 2023)	TCSVT23	76.14	76.67	84.77	86.93	85.44	57.96	50.63	62.28	71.04	65.82
#10 SequencePAR (Jin et al. 2023)	arXiv23	71.88	71.99	83.24	82.29	82.29	57.88	50.27	65.81	65.79	65.37
#11 VTB-PLIP (Zuo et al. 2023)	arXiv23	73.90	73.16	82.01	84.82	82.93	56.30	46.77	61.20	64.47	62.18
#12 Rethink-PLIP (Zuo et al. 2023)	arXiv23	69.44	68.90	79.82	81.15	80.48	57.18	46.98	63.57	62.16	62.86
#13 PromptPAR (Wang et al. 2024a)	TCSVT24	78.81	76.53	84.40	87.15	85.35	63.24	53.62	**66.15**	71.84	68.32
#14 SSPNet (Shen et al. 2024)	PR24	74.03	74.10	84.01	84.02	84.02	56.15	46.75	62.44	63.07	62.75
#15 HAP (Yuan et al. 2024)	NIPS24	76.92	76.12	84.78	86.14	85.45	58.70	50.59	65.60	66.91	66.25
#16 MambaPAR (Wang et al. 2024c)	arXiv24	73.85	73.64	83.19	84.29	83.28	56.75	47.34	61.92	64.98	62.80
#17 MaHDFT (Wang et al. 2024b)	arXiv24	74.08	74.40	82.82	86.41	83.93	58.67	50.65	62.39	71.13	65.85
Zero-shot	-	56.93	52.97	72.26	64.69	67.46	52.19	39.26	60.12	52.09	55.15
Ours	-	**80.13**	**78.71**	84.39	**90.52**	**86.94**	**66.29**	**58.11**	65.68	**81.21**	**72.05**

Table 1: Comparison with public methods on our datasets. The first and second highest scores are represented by **bold font** and underline, respectively. Zero-shot refers to the use of MiniGPT4 for zero-shot inference to generate all dataset descriptions. It then utilizes BERT to extract text features, followed by training a fully connected layer for classification.

Methods	Publish	PETA					PA100K					RAPv1				
		mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1	mA	Acc	Prec	Rec	F1
SSCNet (Jia, Chen, and Huang 2021)	ICCV21	86.52	78.95	86.02	87.12	86.99	81.87	78.89	85.98	89.10	86.87	82.77	68.37	75.05	87.49	80.43
CAS (Yang et al. 2021)	IJCV21	86.40	79.93	87.03	87.33	87.18	82.86	79.64	86.81	87.79	85.18	84.18	68.59	77.56	83.81	80.56
IAA (Wu et al. 2022)	PR22	85.27	78.04	86.08	85.80	85.64	81.94	80.31	88.36	88.01	87.80	81.72	68.47	79.56	82.06	80.37
DRFormer (Tang and Huang 2022)	NC22	89.96	81.30	85.68	91.08	88.30	82.47	80.27	87.60	88.49	88.04	81.81	70.60	80.12	82.77	81.42
VAC (Guo, Fan, and Wang 2022)	IJCV22	-	-	-	-	-	82.19	80.66	88.72	88.10	88.41	81.30	70.12	81.56	81.51	81.54
DAFL (Jia et al. 2022)	AAAI22	87.07	78.88	85.78	87.03	86.40	83.54	80.13	87.01	89.19	88.09	83.72	68.18	77.41	83.39	80.29
VTB (Cheng et al. 2022)	TCSVT22	85.31	79.60	86.76	87.17	86.71	83.72	80.89	87.88	89.30	88.21	82.67	69.44	78.28	84.39	80.84
PromptPAR (Wang et al. 2024a)	TCSVT24	88.76	82.84	**89.04**	89.74	89.18	87.47	83.78	**89.27**	91.70	90.15	85.45	71.61	79.64	86.05	82.38
PARformer (Fan et al. 2023)	TCSVT23	89.32	82.86	88.06	91.98	89.06	84.46	81.13	88.09	91.67	88.52	84.43	69.94	79.63	**88.19**	81.35
OAGCN (Lu et al. 2023)	TMM23	89.91	82.95	88.26	89.10	88.68	83.74	80.38	84.55	90.42	87.39	**87.83**	69.32	78.32	87.29	82.56
SSPNet (Shen et al. 2024)	PR24	88.73	82.80	88.48	90.55	89.50	83.58	80.63	87.79	89.32	88.55	83.24	70.21	**80.14**	82.90	81.50
SOFA (Wu et al. 2024)	AAAI24	87.10	81.10	87.80	88.40	87.80	83.40	81.10	88.40	89.00	88.30	83.40	70.00	80.00	83.00	81.20
FRDL (Zhou et al. 2024)	ICML24	88.59	-	-	-	89.03	89.44	-	-	-	88.05	87.72	-	-	-	79.16
Zero-shot	-	61.32	50.75	68.57	64.00	65.52	65.26	56.99	79.21	65.20	70.75	65.46	50.90	64.48	65.20	66.06
Ours	-	**92.25**	**84.59**	88.41	**92.94**	**90.39**	**91.09**	**84.12**	87.73	**94.09**	**90.41**	87.80	**71.86**	78.36	**88.20**	**82.64**

Table 2: Comparison with SOTA methods on PETA, PA100K and RAPv1 datasets. The first and second highest scores are represented by **bold font** and underline, respectively.

the cross-domain split, which achieves the best results on nearly all metrics. The experiments on the MSP60K dataset fully validated the effectiveness of our proposed LLM-PAR for attribute recognition.

• **Result on Public Dataset.** As shown in Table 2, we compare our method with several SOTA methods on PETA, PA100K, and RAPv1. Our method outperforms these methods in most metrics. Compared to SSPNet (Shen et al. 2024) with prior guidance on these three datasets, we observe improvements of 3.52/1.79/0.89, 7.51/3.49/1.86, and 4.56/1.65/1.14 in mA, Acc, and F1, respectively. We also get significant improvements compared to OAGCN (Lu et al. 2023), the last best method, with 2.34/1.64/0.89 and 7.35/3.74/3.02 in three metrics on PETA and PA100K, respectively. Though the mA score of 0.03 decreased when we conducted experiments on RAPv1, we still got 2.54/0.08 improvements in Acc and F1 metrics. In contrast to PromptPAR (Wang et al. 2024a) with visual-language modeling, we outperform them by 3.49/1.75/1.21, 3.62/0.34/0.26, and 1.83/0.95/0.74, respectively. Based on the experiments conducted on the four datasets, it is noticed that LLM-PAR delivers impressive results by combining visual classification and LLM modeling within the LLM-augment framework. Furthermore, the AGFA module extracts attribute group-specific features to capture detailed information and integrate them with Q-former into MEQ-Former, thereby enhancing the pedestrian caption details of LLMs.

Component Analysis

We conduct ablation experiments to analyze the contributions of different components in our method, including the visual backbone, AGFA module, and LLM branch. The vi-

#	CLS-Attr	FT Q-Former	LoRA	CLS-LLM	AGFA	CLS-IN	PETA Dataset		
							mA	Acc	F1
1	✓						71.54	58.24	71.96
2	✓	✓					82.89	72.32	81.89
3	✓	✓	✓				90.14	83.25	89.38
4	✓	✓	✓	✓			90.89	83.64	89.60
5	✓	✓	✓	✓	✓		91.78	84.47	90.27
6	✓	✓	✓	✓	✓	✓	92.25	84.59	90.39

Table 3: Component Analysis on the PETA Dataset. The mA, Acc, and F1 results are reported.

sual backbone analysis reveals that the EVA-CLIP (Fang et al. 2023) and Q-Former (Li et al. 2023) alone achieve mA, Acc, and F1 scores of 71.54, 58.24, and 71.96, respectively. Fine-tuning with LoRA (Hu et al. 2022) improves these scores to 90.14, 83.25, and 89.38. When additional LLM branches are incorporated, the scores further improve, reaching 90.89 for mA, 83.64 for Acc, and 89.60 for F1. The efficacy of the AGFA module is confirmed with scores of 91.78, 84.47, and 90.27, highlighting its role in improving feature aggregation and model recognition capabilities. Lastly, the CLS-IN module improves the mA, Acc, and F1 scores by 0.47, 0.12, and 0.12, respectively, indicating its contribution to enhancing the recognition of tail categories and supplementing other categories through shared feature learning.

Ablation Study

In this section, we conduct detailed analysis experiments on the main module of LLM-PAR. This includes the Number of AGFA Layers and the Length of PartQ in the PETA (Deng et al. 2014) dataset.

• **Analysis on the Number of AGFA Layers.** As shown in Table 4, we introduce the AGFA module for extracting pedestrian attribute group features in this study. We analyze the impact of AGFA modules with 1, 3, 6, 9, and 12 layers on recognition performance. Our analysis reveals that increasing the number of AGFA layers improved recognition performance. However, considering computational efficiency, we opt for a 3-layer AGFA module to balance computational burden and performance.

• **Analysis on the Length of PartQ.** As shown in Table 4, we examine the effect of the number of attribute group queries in the AGFA module on performance. Our findings show that using 128 queries obtains the best performance, with performance deteriorating with more than 256 queries and a significant decline observed with 64 queries.

• **Analysis on the Aggregation Strategy of Threes Branches.** To improve the aggregation of results from three branches, we design and evaluate some aggregation strategies, including mean pooling and max pooling, and the performance of each strategy is reported in Table 5. Mean pooling achieves 92.25 and 90.39 in mA and F1 scores, respectively, while max pooling achieves 92.46 and 88.95. We find that mean pooling mitigates the influence of abnormal values on the final result. Additionally, we explore and design the attribute-specific aggregation (ASA), resulting in 91.53 and 90.17.

AGFA	Layers					Querys		
	1	3	6	9	12	64	128	256
mA	91.97	92.25	92.57	92.68	92.77	92.01	92.25	92.20
F1	89.82	90.39	90.61	90.69	90.53	88.28	90.39	90.02

Table 4: Performance comparison for AGFA across different layers and query numbers

Metric	ASA	Mean Pooling	Max Pooling
mA	91.53	92.25	92.46
F1	90.17	90.39	88.95

Table 5: Comparison of different aggregation strategies.

6 Conclusion

This paper addresses the limitations of existing pedestrian attribute recognition (PAR) datasets by introducing MSP60K, a new large-scale, cross-domain dataset with 60,122 images and 57 attribute annotations across eight scenarios. By incorporating synthetic degradation, we further bridge the gap between the dataset and real-world challenging conditions. Our comprehensive evaluation of 17 representative PAR models under both random and cross-domain split protocols establishes a more rigorous benchmark. Moreover, we propose the LLM-PAR framework, which leverages a pre-trained vision Transformer backbone, a multi-embedding query Transformer for partial-aware feature learning, and is enhanced by a Large Language Model for ensemble learning and visual feature augmentation. The experimental results across multiple PAR benchmark datasets demonstrate the effectiveness of our proposed framework. Both the MSP60K dataset and the source code will be released to the public upon acceptance, contributing to future advancements in human-centered research and PAR technology.

In our future work, we plan to further expand the scale of the dataset to conduct more extensive and thorough experimental validations. Moreover, the training and inference of the model still require substantial computational resources. In the future, we will design lightweight models to achieve a better balance between accuracy and performance.

Acknowledgments

This research is jointly supported by the National Natural Science Foundation of China (No. 62376004, 62102205) and the Natural Science Foundation of Anhui Province (No.

2208085J18, 2408085Y032). The authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

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

Deng, Y.; Luo, P.; Loy, C. C.; and Tang, X. 2014. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, 789–792.

Fan, X.; Zhang, Y.; Lu, Y.; and Wang, H. 2023. Parformer: Transformer-based multi-task network for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1): 411–423.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Guo, H.; Fan, X.; and Wang, S. 2022. Visual attention consistency for human attribute recognition. *International Journal of Computer Vision*, 130(4): 1088–1106.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; and Wang, L. 2024. Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17689–17699.

Jia, J.; Chen, X.; and Huang, K. 2021. Spatial and semantic consistency regularizations for pedestrian attribute recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 962–971.

Jia, J.; Gao, N.; He, F.; Chen, X.; and Huang, K. 2022. Learning disentangled attribute representations for robust pedestrian attribute recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1069–1077.

Jia, J.; Huang, H.; Chen, X.; and Huang, K. 2021. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*.

Jin, J.; Wang, X.; Li, C.; Huang, L.; and Tang, J. 2023. Sequencepar: Understanding pedestrian attributes via a sequence generation paradigm. *arXiv preprint arXiv:2312.01640*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, D.; Chen, X.; and Huang, K. 2015. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 111–115. IEEE.

Li, D.; Zhang, Z.; Chen, X.; and Huang, K. 2019. A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios. *IEEE Transactions on Image Processing*, 28(4): 1575–1590.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016a. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, W.; Cao, Z.; Feng, J.; Zhou, J.; and Lu, J. 2022b. Label2label: A language modeling framework for multi-attribute learning. In *European Conference on Computer Vision*, 562–579. Springer.

Li, Y.; Huang, C.; Loy, C. C.; and Tang, X. 2016b. Human Attribute Recognition by Deep Hierarchical Contexts. In *European Conference on Computer Vision*, 684–700. Springer.

Li, Y.; Xiao, Z.; Yang, L.; Meng, D.; Zhou, X.; Fan, H.; and Zhang, L. 2024. AttMOT: Improving Multiple-Object Tracking by Introducing Auxiliary Pedestrian Attributes. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95: 151–161.

Liu, J.; Zhu, X.; Liu, F.; Guo, L.; Zhao, Z.; Sun, M.; Wang, W.; Lu, H.; Zhou, S.; Zhang, J.; et al. 2021. Opt: omni-perception pre-trainer for cross-modal understanding and generation. *arXiv preprint arXiv:2107.00249*.

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 *Proceedings of the IEEE international conference on computer vision*, 350–359.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 22631–22648. PMLR.

Lu, W.-Q.; Hu, H.-M.; Yu, J.; Zhou, Y.; Wang, H.; and Li, B. 2023. Orientation-aware pedestrian attribute recognition based on graph convolution network. *IEEE Transactions on Multimedia*, 26: 28–40.

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 *International Conference on Machine Learning*, 8748–8763. PMLR.

Shen, J.; Guo, T.; Zuo, X.; Fan, H.; and Yang, W. 2024. SSP-Net: Scale and spatial priors guided generalizable and interpretable pedestrian attribute recognition. *Pattern Recognition*, 148: 110194.

Tan, Z.; Yang, Y.; Wan, J.; Guo, G.; and Li, S. Z. 2020. Relation-aware pedestrian attribute recognition with graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12055–12062.

Tang, Z.; and Huang, J. 2022. DRFormer: Learning dual relations using Transformer for pedestrian attribute recognition. *Neurocomputing*, 497: 159–169.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*.

Wang, X.; Jin, J.; Li, C.; Tang, J.; Zhang, C.; and Wang, W. 2024a. Pedestrian Attribute Recognition via CLIP based Prompt Vision-Language Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Wang, X.; Kong, W.; Jin, J.; Wang, S.; Gao, R.; Ma, Q.; Li, C.; and Tang, J. 2024b. An Empirical Study of Mamba-based Pedestrian Attribute Recognition. arXiv:2407.10374.

Wang, X.; Wang, S.; Ding, Y.; Li, Y.; Wu, W.; Rong, Y.; Kong, W.; Huang, J.; Li, S.; Yang, H.; et al. 2024c. State space model for new-generation network alternative to transformers: A survey. *arXiv preprint arXiv:2404.09516*.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121: 108220.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, 3–19.

Wu, J.; Huang, Y.; Gao, M.; Niu, Y.; Yang, M.; Gao, Z.; and Zhao, J. 2024. Selective and Orthogonal Feature Activation for Pedestrian Attribute Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6039–6047.

Wu, J.; Huang, Y.; Gao, Z.; Hong, Y.; Zhao, J.; and Du, X. 2022. Inter-Attribute awareness for pedestrian attribute recognition. *Pattern Recognition*, 131: 108865.

Yang, Y.; Tan, Z.; Tiwari, P.; Pandey, H. M.; Wan, J.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision*, 129: 2731–2744.

Yuan, J.; Zhang, X.; Zhou, H.; Wang, J.; Qiu, Z.; Shao, Z.; Zhang, S.; Long, S.; Kuang, K.; Yao, K.; et al. 2024. Hap: Structure-aware masked image modeling for human-centric perception. *Advances in Neural Information Processing Systems*, 36.

Zheng, A.; Wang, H.; Wang, J.; Huang, H.; He, R.; and Hussain, A. 2023a. Diverse features discovery transformer for pedestrian attribute recognition. *Engineering Applications of Artificial Intelligence*, 119: 105708.

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. 2023b. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhou, Y.; Hu, H.-M.; Xiang, Y.; Zhang, X.; and Wu, H. 2024. Pedestrian Attribute Recognition as Label-balanced Multi-label Learning. In *Forty-first International Conference on Machine Learning*.

Zhou, Y.; Hu, H.-M.; Yu, J.; Xu, Z.; Lu, W.; and Cao, Y. 2023. A solution to co-occurence bias: attributes disentanglement via mutual information minimization for pedestrian attribute recognition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Zuo, J.; Yu, C.; Sang, N.; and Gao, C. 2023. Plip: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*.