SUPPLEMENTARY MATERIALS FOR "DYNAMIC PATTERN ALIGNMENT LEARNING FOR PRETRAINING LIGHTWEIGHT HUMAN-CENTRIC VISION MODELS"

Anonymous authors
Paper under double-blind review

A THE USE OF LARGE LANGUAGE MODELS

In the preparation of this paper, we employed Large Language Models (LLMs) solely as a writing assistance tool for limited text polishing and language refinement. LLMs were not involved in any aspects of research ideation, conceptual development, technical analysis, algorithm design, experimental execution, or result analyses. All scientific contributions, methodological innovations, and intellectual content remain entirely our own.

B DISCUSSION

Limitation. The performance of the student modle is influenced by the teacher model used for pretraining. Additionally, due to limitations in computational resource, we have only tested our method on image datasets and downstream tasks. However, our approach is also applicable to human-centric video understanding tasks, which will be explored in our future work.

Broader Impact. As demonstrated in Section 4, our method outperforms existing pretraining methods across various downstream tasks, highlighting the potential of DPAL as a novel distillation-based pretraining paradigm. Moreover, DPAL serves as an efficient knowledge distillation technique that enables the development of compact variants of large HVMs, making them suitable for deployment on resource-constrained edge devices. Additionally, DPAL eliminates the necessity of accessing the teacher model's original pretraining datasets by utilizing a relatively small open-source dataset of approximately 1 million images for pretraining. This pretraining paradigm significantly reduces training costs and enhances the accessibility of DPAL for the research community, thereby broadening its potential applications. Furthermore, the codebase developed in this work is publicly released to promote reproducibility and further advancements in research.

C MORE IMPLEMENTATION DETAILS

C.1 MODEL ARCHITECTURE

Backbone. We conducted experiments on various student backbones and teacher backbones, with the corresponding settings presented in Table 1.

Dynamic Pattern Decoder. The dynamic pattern decoder comprises a self-attention module, a router module, three experts and a dynamic expert generator. The router is responsible for assigning experts to different visual tokens. The experts specialize in handing specific patterns. The dynamic expert generator produces weights for experts conditioned on the visual tokens and pattern queries.

Dynamic Expert Generator. The dynamic expert generator consists of self-attention, cross-attention, and FFN modules. We design three learnable expert tokens $T_e = [T_e^1, T_e^2, T_e^3]$, which pass through self-attention, cross-attention, and FFN modules to update the parameters of the three experts. In the cross-attention module, representations from backbone are used as the keys and values, while the expert tokens serve as the queries. The cross-attention module ensures that the parameters of the experts are updated based on the corresponding representations and pattern queries, enabling each expert to selectively focus on the most relevant pattern.

Table 1: Configuration of neural architectures. Both Vision Transformer (ViT-X) and Swin Transformer (Swin-X) are used for investigation.

Arch	Patch size	Embed dim	Heads	Blocks
ViT-Ti	16	192	6	12
ViT-S	16	384	6	12
ViT-B	16	768	12	24

Arch	Patch size	Window size	Embed dim	Heads	Blocks
Swin-Ti	4	7	96	(3,6,12,24)	(2,2, 6,2)
Swin-S	4	7	96	(3,6,12,24)	(2,2,18,2)
Swin-B	4	7	128	(4,8,16,32)	(2,2,18,2)

Router. We designed a router that dynamically assigns experts to distanct visual patterns, thereby decoupling the alignment learning of three visual patterns. We designed a learnable router token T_r, using the representations extracted by the backbone as keys and values. The routing token dynamically adjusts the weights of different experts W_e based on different patterns, enabling the model to effectively capture diverse patterns and enhance its performance in complex visual tasks.

C.2 Pretraining Details

All lightweight HVMs are pretrained using 8 A6000 48G GPUs. We employ the AdamW optimizer(Loshchilov & Hutter, 2017) with an effective batch size of 2048 (i.e., 256 per GPU). As shown in Table 2, each model is pretrained from scratch for 100 epochs. The learning rate is 2.5e-4 and is decayed via Cosine Annealing scheduler(Loshchilov & Hutter, 2016). The single-person image size is 256×128, while the multi-person image is 256×256.

Table 2: Configurations of pretraining.

Configuration	Value
Batch size	2048
Optimizer	AdamW
Learning rate	2.5e–4
Learning rate decay	Consine scheduler
Weight decay	0.05
Warmup epochs	10
Epochs	100
Image size	256×128

C.3 Finetuning Details

We utilize representative methods from downstream tasks as baselines, subsequently replacing their backbones with our pretrained backbones for finetuning. The list of codebases used for evaluation is presented in Table 3.

Table 3: Implementation codebases and configurations of fine-tuning on 12 datasets.

Task	Dataset	Codebases	Image size	Learning rate	Epoch
I2I ReID	Market1501 (Zheng et al., 2015) MSMT17 (Wei et al., 2018)	SOLIDER (Chen et al., 2023)	256×128	2e-4	120
T2I ReID	CUHK-PEDES(Li et al., 2017) ICFG-PEDES(Ding et al., 2021)	IRRA (Jiang & Ye, 2023)	384×128	1e-4	60
Attribute recognition	PA100(Liu et al., 2017) PETAzs(Deng et al., 2014)	SOLIDER (Chen et al., 2023)	256×128	1e-4	25
Pose estimation	COCO keypoint(Lin et al., 2014)	ViTPose (Xu et al., 2022)	256×192	5e-4	210
Landmark detection	Whole-body COCO(Jin et al., 2020)	ViTPose (Xu et al., 2022)	256×192	5e-4	210
Human parsing	LIP (Liang et al., 2018)	SOLIDER(Chen et al., 2023)	576×384	7e-4	150
Pedestrian detection	CrowdHuman(Shao et al., 2018)	CrowdDet (Chu et al., 2020)	1400×800	2e-4	30
Multiple human parsing	CIHP(Gong et al., 2018)	Cpi-parser (Wang et al., 2024)	1333×800	2e-2	25
Part-level attribute parsing	Fashionpedia(Jia et al., 2020)	KE-RCNN (Wang et al., 2023)	1024×1024	1e-4	32

Table 4: The computational cost in pretraining stage, involving training epoch, training time (Hours) and Memory per GPU (GB).

Setting	ViTKD	MaskedKD	ScaleKD	TinyMiM	Proteus	DPAL
Epochs	300	300	200	300	300	100
Time	41	26	60	15	30	22
Memory	10	22	26	24	16	26
Downstream Tasks						
I2I Person ReID	90.5	79.7	81.6	91.6	93.6	95.2
Human Parsing	52.0	50.9	55.6	53.0	54.3	55.9
Pedestrian Detection	86.6	84.2	87.4	86.4	87.6	88.7

Table 5: Investigating the effect of DPAL on vision-language spatial reasoning task.

Evaluation on OmniSpatial

Method	Vision Encoder	Avg.	Dynamic Reasoning	Spatial Interaction	Complex Logic	Perspective Taking
LLaVA-1.5-7B	CLIP-ViT-L (304M)	34.97	**54.46/31.23**	35.29/**36.19**/33.94	**29.01**/24.08	**55.60**/34.66/35.14
	DPAL-ViT-T (5M)	**35.62**	45.95/26.30	**56.47**/35.24/**42.73**	18.56/**25.16**	53.92/**36.70/46.99**

Evaluation on Robo2VLM

Method	Vision Encoder	Avg.	Spatial Reasoning RS/OS/SR/SU/MV	Goal Reasoning TS-G/TS-S/TS-GL	Interaction Reasoning AU/IP/TU
LLaVA-1.5-7B	CLIP-ViT-L (304M)	21.58	35.32/23.87/16.08/**17.78/17.50**	**31.82/23.79**/19.03	**20.30/21.74**/22.37
	DPAL-ViT-T (5M)	**24.58**	**53.66/27.16/19.29**/2.52/17.31	29.55/19.88/**46.88**	17.24/21.18/**31.82**

D Ablation Study

D.1 Training Efficiency

From the results listed in Table 4, we observe that the proposed DPAL achieves superior downstream performance while maintaining comparable pretraining costs. As shown in Table 1, all methods require 15 40 hours, while the proposed DPAL requires 22 hours. Second, from the perspective of downstream fine-tuning, we choose a fair and widely-used setting, where only the pre-trained backbone is retained in downstream evaluation, and the alignment module is discarded. In this way, the training costs in downstream tasks are the same for all pre-training methods. Based on this, the DPAL does not bring significant computational burden.

D.2 Ablation Study on Scale of the Dataset

To explore the optimal scale of the dataset, we construct five subsets of varying scales (0.2M, 0.5M, 1M, 2M, and 4M samples) from the LUPerson dataset for pretraining. As shown in Figure 1, we observe that the performance on the 0.2M and 0.5M subsets is significantly worse than on the 1M subset. Moreover, increasing the dataset size does not lead to further performance improvement. Therefore, LUP1M, as the subset of LUPerson, is sufficient to support distillation-based pretraining.

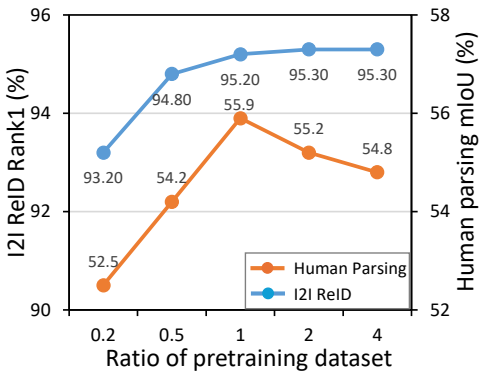


Figure 1: The impact of different scale of datasets on I2I Reid and human parsing tasks.

D.3 Vision-Language Tasks

In this section, we further investigate the effectiveness of proposed method on two vision-language tasks: 1)

Table 6: Investigating the effect of DPAL on vision-language robot control task.

Simulation Task	ACT (ResNet-18-11M) Easy/Hard	RDT (SigLip-400M) Easy/Hard	PI0 (SigLip-400M) Easy/Hard	H-RDT (SigLip-400M) Easy/Hard	Ours (ViT-Ti-5M) Easy/Hard
Grab Roller	66.0/6.0	74.0/43.0	**96.0**/80.0	95.0/52.0	83.0/**57.0**
Place_object_basket	0.0/0.0	42.0/14.0	62.0/10.0	**62.0/19.0**	32.0/4.0

spatial reasoning; and 2) embodied robot control. Specifically, we choose the InternViT-400M (Chen et al., 2025c) as the teacher model, and use proposed DPAL to distill it into ViT-Ti by leveraging the ImageNet-1M (Deng et al., 2009) as a medium. When testing pretrained model on spatial reasoning task, we replace the vision part of LLavA-1.5-vicuna-7B (Liu et al., 2023) with the pretrained ViT-Ti, and test it using OmniSpatial (Jia et al., 2025) and Robo2VLM (Chen et al., 2025a) benchmark. As for the embodied robot control task, we replace the vision part of the H-RDT (Bi et al., 2025) and test it using RobotWin2.0 (Chen et al., 2025b) simulation benchmark. The results reported in Table 5 and Table 6 shows that adopting pretrained lightweight ViT (5M) as the vision encoder achieves competitive performance, which is comparable to that of large vision encoder (400M).

D.4 Ablation Study on Variants of Pattern Decoder

Pattern decoder functions as an adapter to align the outputs of lightweight HVMs to that of large HVMs. This section study the variants of pattern decoder: 1) MAE-Style (He et al., 2022) decoder, which contains two transformer blocks; 2) Standard MoE, where experts in MoE block is the fixed MLP; and 3) proposed D-PaDe, where the experts are dynamically generated via pattern queries with input image. Comparison results reported in Table 7 show that D-PaDe is the best choice for distillation-based pretraining by far.

Table 7: Ablation study on variants of pattern decoder (%). Aligning ViT-Ti/16 with PATH-B using MAE-style decoder, Standard MoE or D-PaDe.

Setting	I2I ReID	Human Parsing	Detection
w/o decoder	95.2	55.2	87.2
MAE-style	95.1	55.0	87.8
Standard MoE	94.1	55.9	88.5
D-PaDe	95.2	55.9	88.7

D.5 Ablation Study on Variants of Student Model

We evaluate the performance of DPAL on downstream tasks with different model architectures. We use PATH-B as the teacher model and perform distillation for 100 epochs by default. As shown in the Table 8, we employ vision transformer (Dosovitskiy et al., 2020) for the ViT architecture and swin transformer (Liu et al., 2021) for the hybrid architecture. The other settings are the same as in Section 4.1 ans Section 4.2. We observe a consistent improvement in model performance concomitant with the increasing model parameters, as exemplified by I2I ReID task where $Rank1$ increases by 0.6% (+16M), 1.2% (+22M), 1.7% (+44M) compared to ViT-Tiny. However, this improvement is accompanied by a corresponding increase in training costs. Moreover, our method is model-agnostic, demonstrating strong performance on both ViT and hybrid architectures.

D.6 Ablation Study on Model Size of Teacher

We investigate whether employing teacher models with larger size enhances the performance of the student model. Specifically, we employ PATH-B and PATH-L as teacher models to distill ViT-Tiny. The results presented in Table 9 indicate that increasing the size of the teacher model does not yield performance gains across a wide range of downsream tasks. This may be due to the larger gap between

Table 8: Impact of model architecture. We employ PATH-B as teacher model and perform distillation with DPAL on four student architectures.

(a) Single-person discrimitive tasks (%).

Arch	Type	#Param	I2I ReID		T2I ReID		Attribute recognition	
			Market↑	MSMT17↑	CUHK↑	ICFG↑	PA100K↑	PETAzs↑
ViT-Ti/16	ViT	5M	95.2	84.3	64.3	56.0	82.4	74.0
ViT-S/16	ViT	21M	95.8	86.1	65.8	58.5	83.9	74.1
Swin-Ti/4	Hybrid	27M	96.4	86.2	66.9	58.5	83.1	74.9
Swin-S/4	Hybrid	49M	**96.9**	**88.2**	**69.6**	**60.0**	**85.9**	**77.1**

(b) Single-person dense prediction tasks (%).

Arch	Type	#Param	Pose estimation		Landmark detection		Human parsing	
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$	$mIoU\uparrow$	$mAcc\uparrow$
ViT-Ti/16	ViT	5M	72.6	75.8	48.8	61.5	55.9	66.7
ViT-S/16	ViT	21M	73.3	76.3	53.1	65.5	58.1	68.7
Swin-Ti/4	Hybrid	27M	75.1	78.1	53.9	65.7	59.3	69.7
Swin-S/4	Hybrid	49M	**76.3**	**79.4**	**55.6**	**67.2**	**60.7**	**71.5**

(c) Multi-person visual understanding tasks (%).

Arch	Type	#Param	Pedestrian detection		Multiple human parsing		Part-level attribute parsing	
			$AP\uparrow$	$MR\downarrow$	$mIoU\uparrow$	$AP_p\uparrow$	$AP^{box}_{IoU+F_1}\uparrow$	$AP^{segm}_{IoU+F_1}\uparrow$
ViT-Ti/16	ViT	5M	88.7	45.5	51.9	50.3	39.8	37.0
ViT-S/16	ViT	21M	89.2	42.9	**55.9**	**53.4**	42.9	39.3
Swin-Ti/4	Hybrid	27M	**90.2**	**42.3**	55.8	53.3	42.9	39.8
Swin-S/4	Hybrid	49M	89.7	43.1	55.2	52.4	**44.9**	**41.1**

(d) Cross-domain perception tasks (%).

Arch	Type	#Param	Humanart		Chimpact-Pose		AP-10K	
			$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$
ViT-Ti/16	ViT	5M	69.9	73.4	21.9	25.5	67.0	70.3
ViT-S/16	ViT	21M	72.0	75.5	24.7	28.4	69.0	72.4
Swin-Ti/4	Hybrid	27M	72.9	76.1	25.2	29.4	69.4	73.0
Swin-S/4	Hybrid	49M	**75.1**	**78.3**	**27.8**	**32.1**	**71.5**	**74.7**

the larger teacher models and the student model, which is also mentioned in the TinyMIM(Ren et al., 2023).

E VISUALIZATION RESULTS

We provide additional visualization results in Figure 2. First, we visualize the class token representation space of PATH (Tang et al., 2023), Proteus(Zhang et al., 2025), TinyMIM(Ren et al., 2023) and DPAL for two single-person images to investigate the models' ability to learn global identity patterns. As shown in Figure 2 (a), DPAL and PATH distinctly separate the two instances in the representation space, whereas the other methods do not. Second, we conduct principal component analysis (PCA) visualization to investigate the model's capability in capturing local shape patterns. DPAL successfully captures local body shape patterns comparable to those of PATH, while the others fail to capture the whole structure of a person instance. Third, we perform PCA visualization on multi-person images. Similar to PATH, DPAL is able to distinguish different individuals that are depicted by different colors in the visualization. This demonstrates that DAPL has successfully enabled lightweight model to acquire multi-person interaction patterns.

Table 9: Impact of teacher size. We use ViT-Tiny as the student model and perform DPAL distillation separately with teacher models of two different sizes.

(a) Single-person discrimitive tasks (%).

Teacher	I2I ReID		T2I ReID		Attribute recognition	
	Market↑	MSMT17↑	CUHK↑	ICFG↑	PA100K↑	PETAzs↑
PATH-B	95.2	**84.3**	64.3	**56.0**	82.4	74.0
PATH-L	**95.2**	83.7	**66.0**	55.9	**82.7**	**74.3**

(b) Single-person dense prediction tasks (%).

Teacher	Pose estimation		Landmark detection		Human parsing	
	$AP\uparrow$	$AR\uparrow$	$AP\uparrow$	$AR\uparrow$	$mIoU\uparrow$	$mAcc\uparrow$
PATH-B	72.6	75.8	**48.8**	**61.5**	**55.9**	**66.7**
PATH-L	**72.7**	**78.2**	48.6	61.2	55.7	66.6

REFERENCES

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint arXiv:2507.23523*, 2025.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025a.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025b.

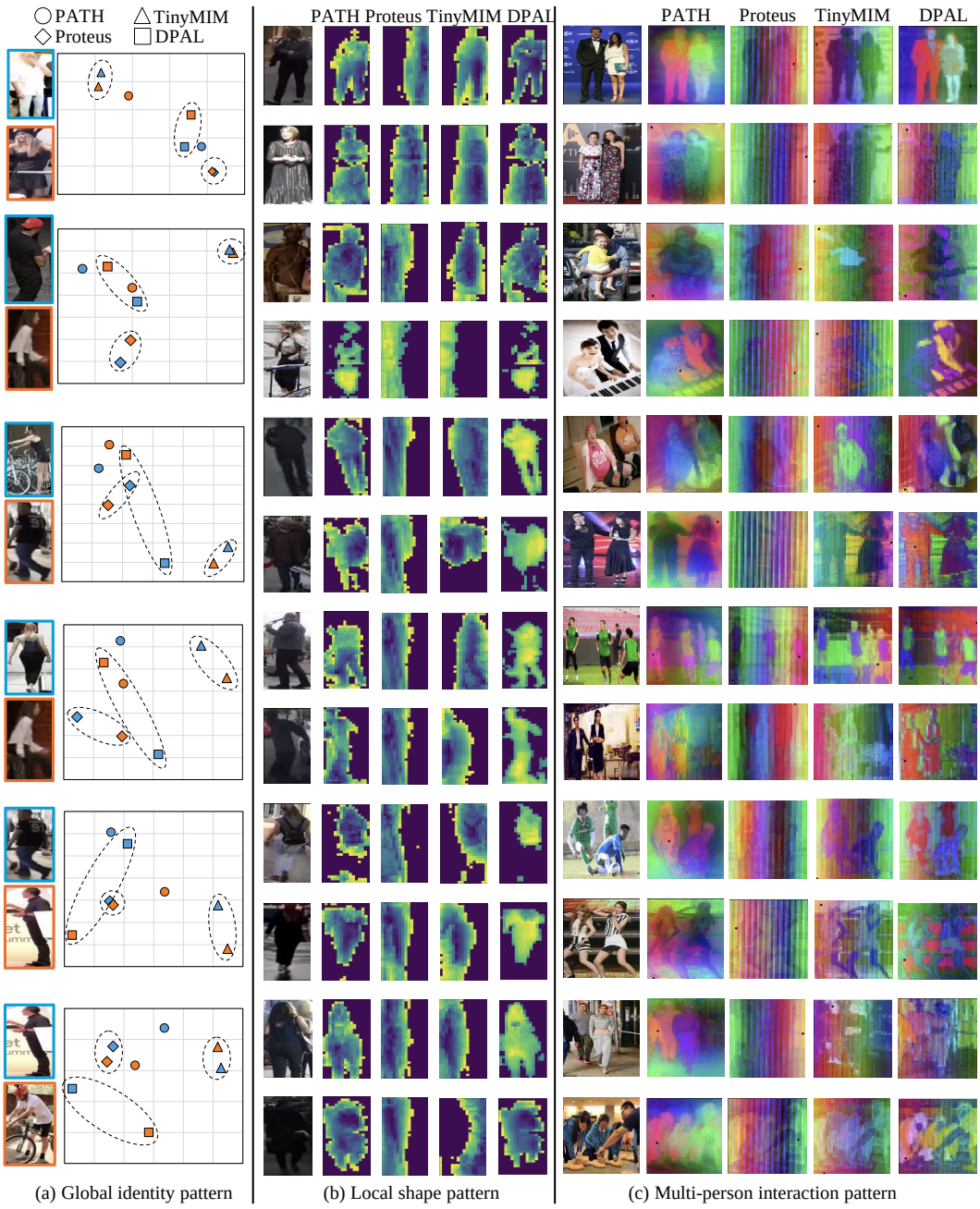
Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15050–15061, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025c.

Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 789–792, 2014.



(a) Global identity pattern (b) Local shape pattern (c) Multi-person interaction pattern

Figure 2: Visualization of learned patterns among four models. From left to right: (a) global identity pattern, (b) local shape pattern, and (c) multi-person interaction pattern.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *European Conference on Computer Vision*, pp. 770–785, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025.

Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In *European Conference on Computer Vision*, pp. 316–332, 2020.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pp. 196–214. Springer, 2020.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. 2017.

Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in Neural Information Processing Systems, 2023.

Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pp. 350–359, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Sucheng Ren, Fangyun Wei, Zheng Zhang, and Han Hu. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3687–3697, 2023.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, et al. Humanbench: Towards general human-centric perception with projector assisted pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21970–21982, 2023.

Xuanhan Wang, Jingkuan Song, Xiaojia Chen, Lechao Cheng, Lianli Gao, and Heng Tao Shen. Ke-rcnn: Unifying knowledge-based reasoning into part-level attribute parsing. *IEEE Transactions on Cybernetics*, 53(11):7263–7274, 2023.

Xuanhan Wang, Xiaojia Chen, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Cpi-parser: Integrating causal properties into multiple human parsing. *IEEE Transactions on Image Processing*, 33: 5771–5782, 2024.

Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 79–88, 2018.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via imagenet-1k. In *The Thirteenth International Conference on Learning Representations*, 2025.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015.