HumanBench: Towards General Human-centric Perception with Projector Assisted Pretraining

Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, Wanli Ouyang

University of Sydney, SenseTime Research, Zhejiang University, Shanghai AI Laboratory

Outline

- Why Human-Centric Foundation Model ?

- HumanBench: Largest Human-centric Datasets in Academy

- PATH: A Projector Assisted preTraining with Hierarchical weight sharing

- Experimental Results and Future Work

Outline

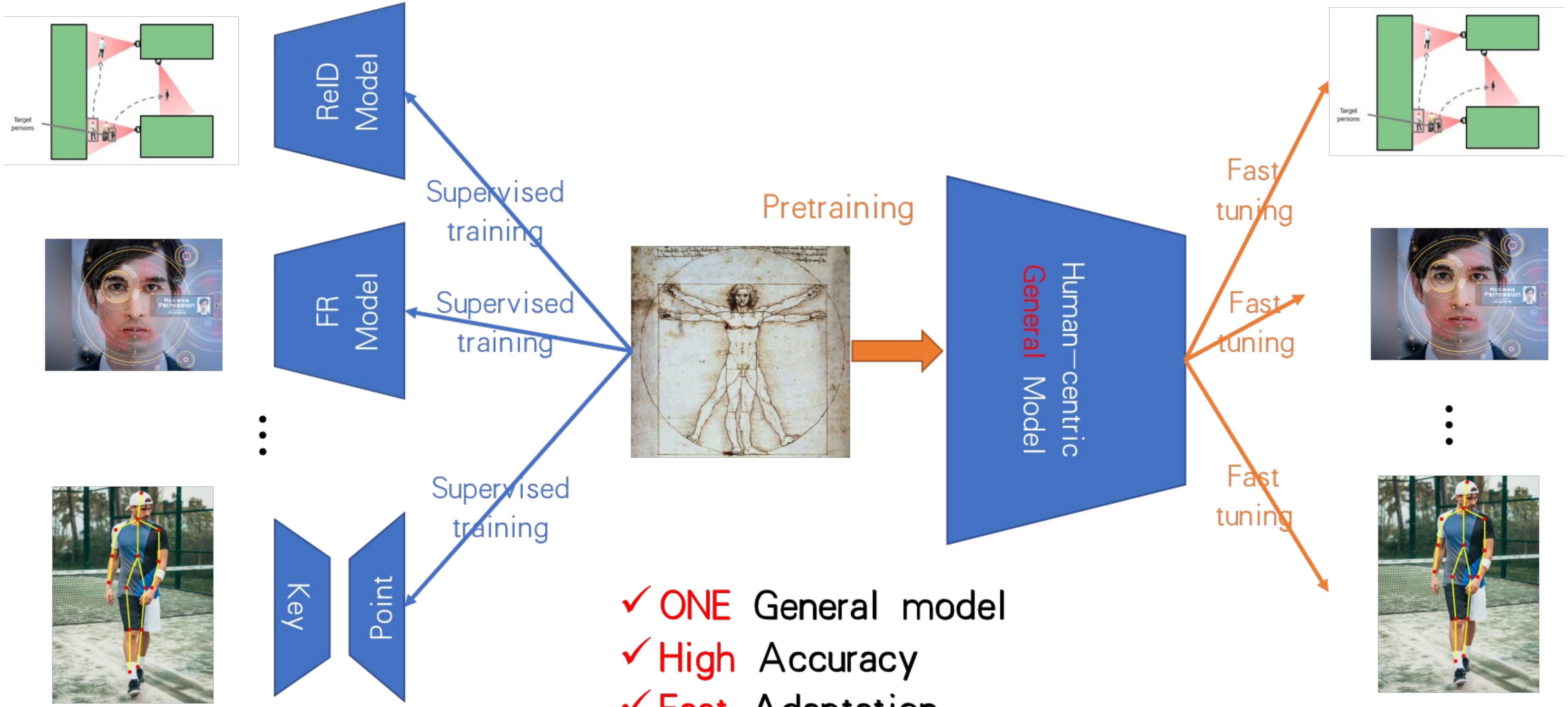
- Why Human-Centric Foundation Model ?

- HumanBench: Largest Human-centric Datasets in Academy

- PATH: A Projector Assisted preTraining with Hierarchical weight sharing

- Experimental Results and Future Work

Diverse Application



✓ONE General model

✓High Accuracy

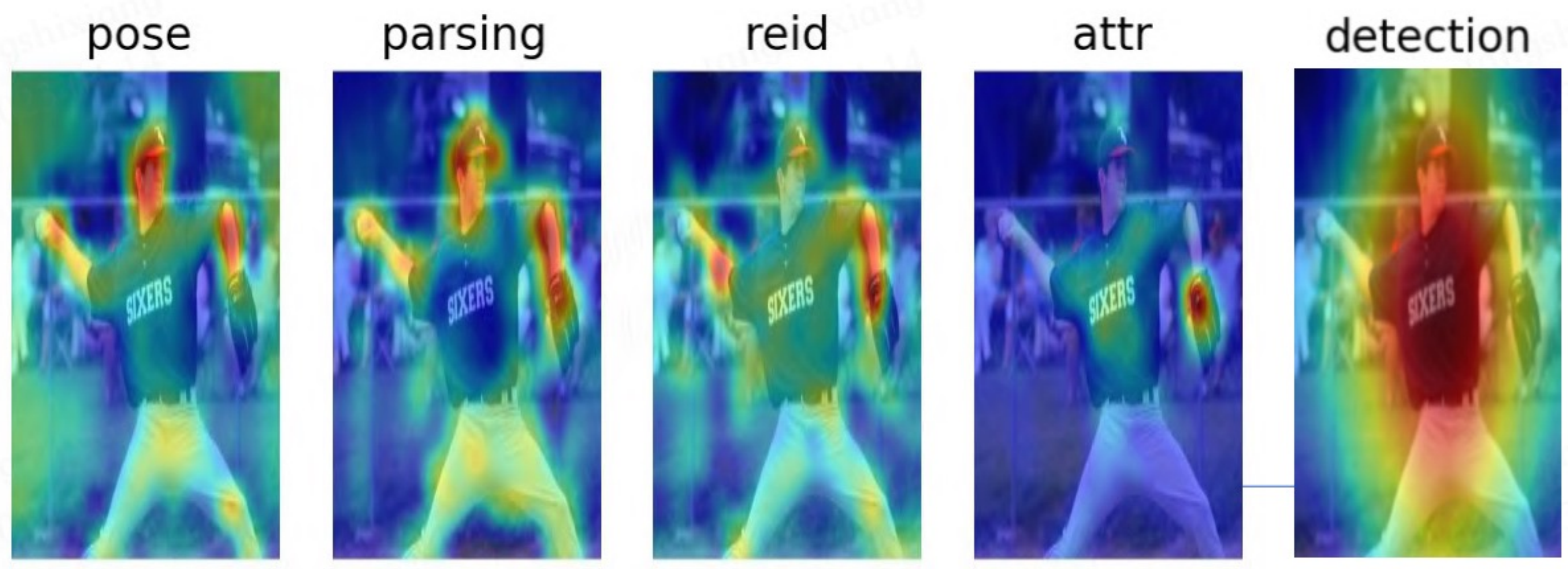
✓Fast Adaptation

✓Less data for downstream tasks

High Correlation among Diverse Human-Centric Tasks

Person Reid, Pedestrian Detection, Attribute: Global Information

Human Parsing, Pose: Local Information



Outline

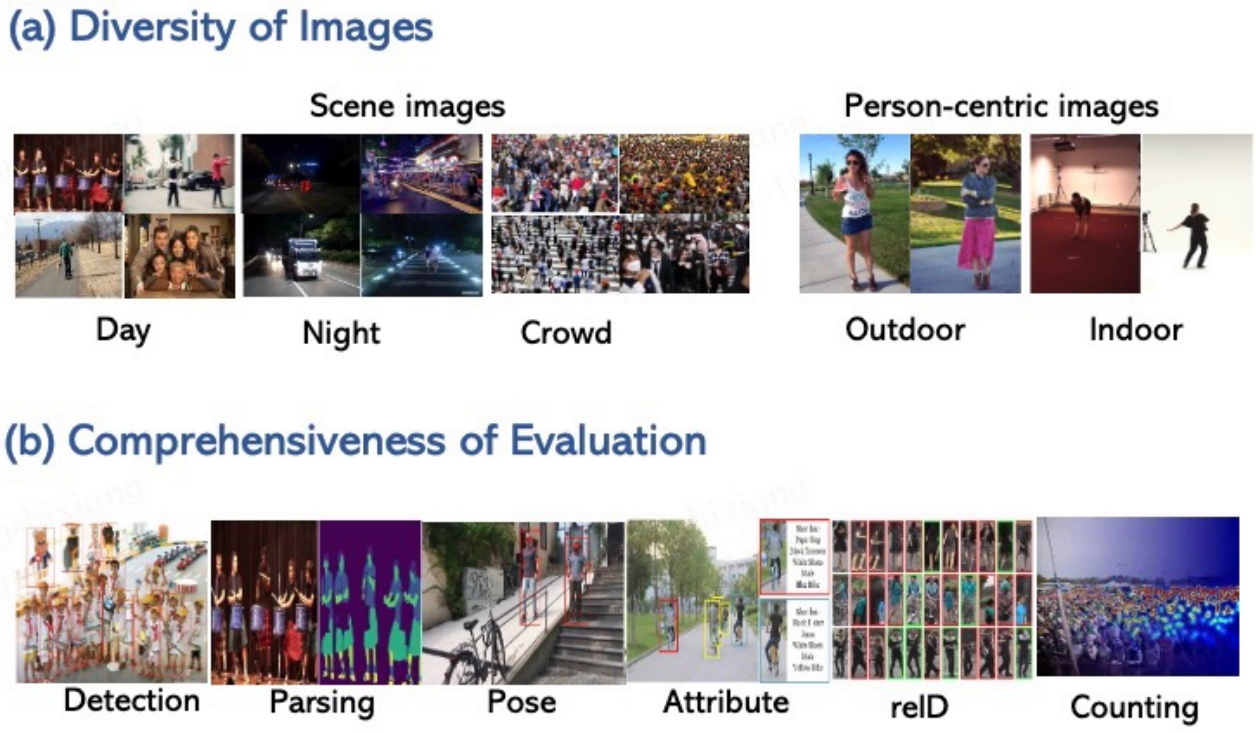
- Why Human-Centric Foundation Model ?
- HumanBench: Largest Human-centric Datasets in Academy
- PATH: A Projector Assisted preTraining with Hierarchical weight sharing
- Experimental Results and Future Work

HumanBench: Largest Human-centric Datasets in Academy

- Diversity of annotations: 6 human-centric tasks

- Diversity of images: scene images, cropped images, indoor images, outdoor images

- Open source: Based on 44 publicly available datasets



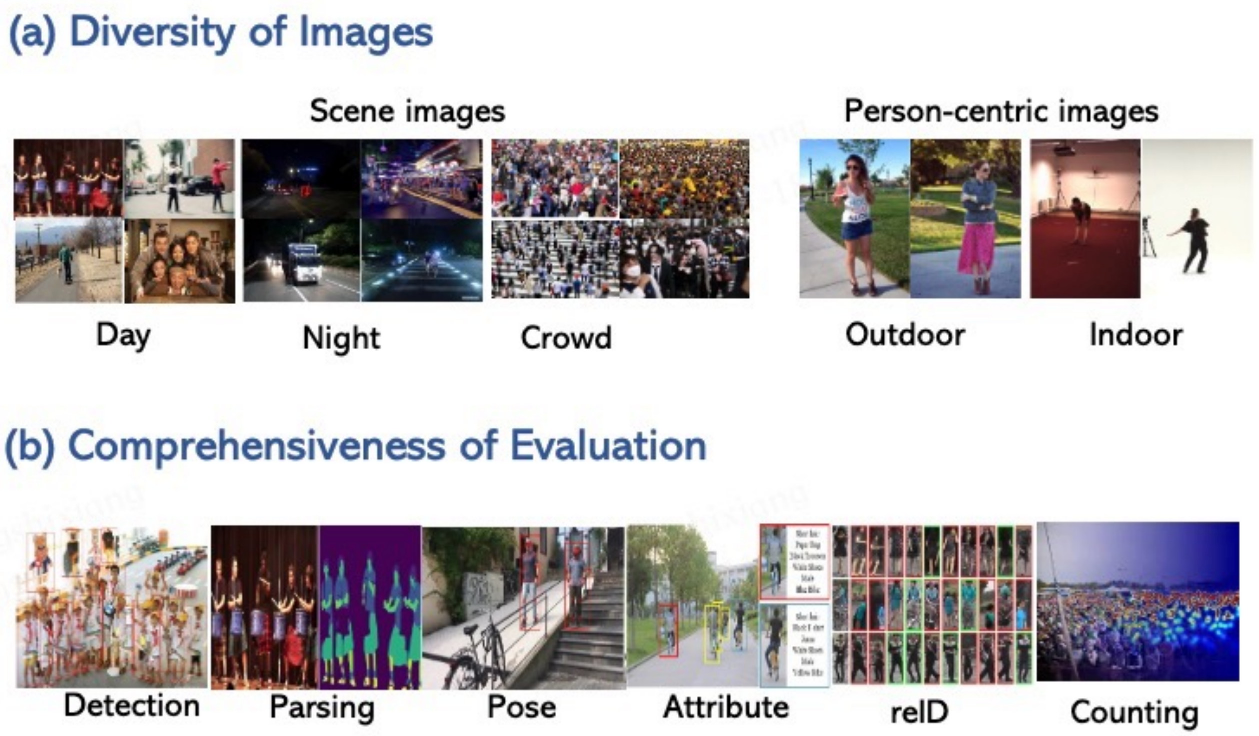
(a) Diversity of Images

(b) Comprehensiveness of Evaluation

HumanBench: Largest Human-centric Datasets in Academy

Pretraining datasets: 11,120,884 images from 37 datasets.

- Diversity of annotations: 6 human-centric tasks
- Diversity of images: scene images, cropped images, indoor images, outdoor images
- Open source: Based on 44 publicly available datasets



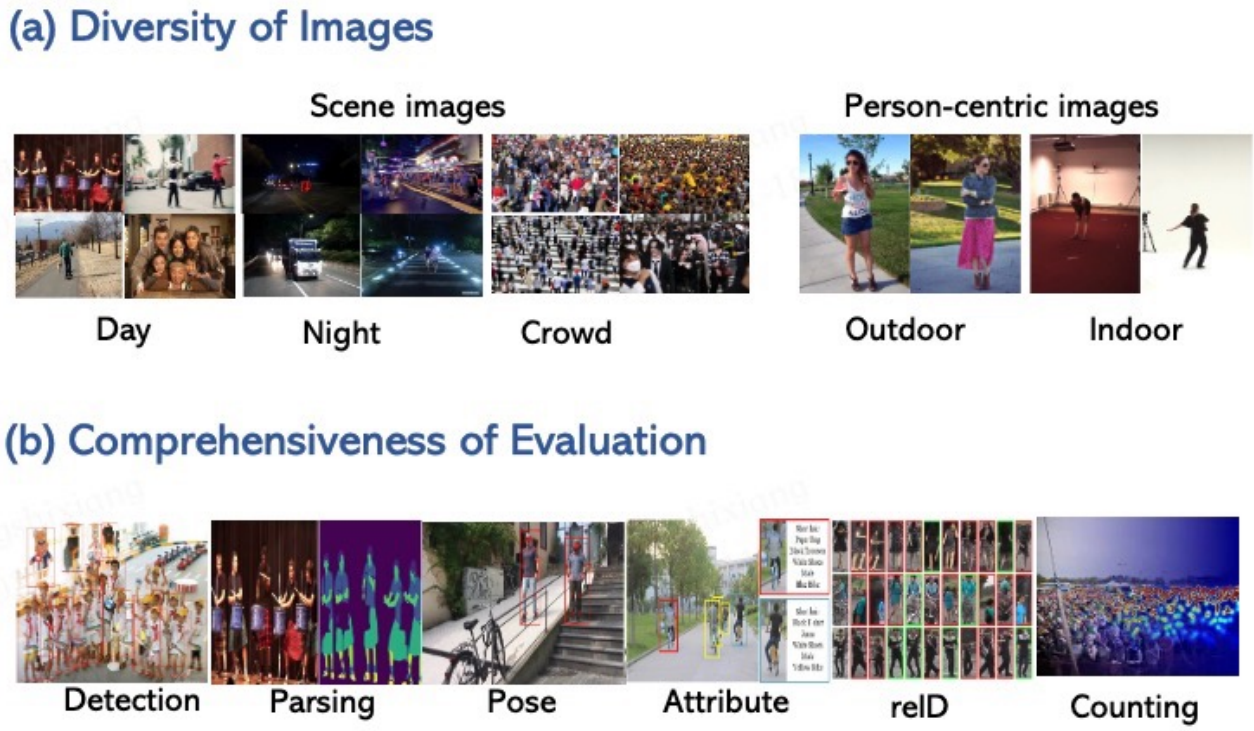
(a) Diversity of Images

(b) Comprehensiveness of Evaluation

HumanBench: Largest Human-centric Datasets in Academy

Pretraining datasets: 11,120,884 images from 37 datasets.

- Diversity of annotations: 6 human-centric tasks
- Diversity of images: scene images, cropped images, indoor images, outdoor images
- Open source: Based on 44 publicly available datasets
- Comprehensiveness of 3 evaluation protocols



HumanBench: Largest Human-centric Datasets in Academy

Evaluation Protocols: 3 protocols.

Efficiency

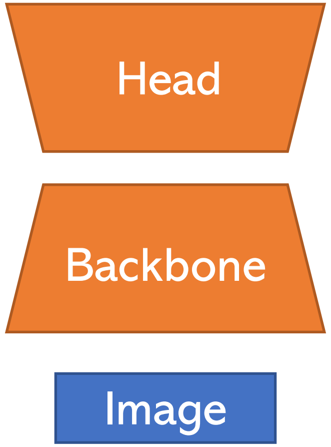
Frozen

Finetuned

HumanBench: Largest Human-centric Datasets in Academy

Evaluation Protocols: 3 protocols.

Efficiency



Full finetune

Full Finetuning: Finetune all parameters using all in the downstream tasks.

Frozen

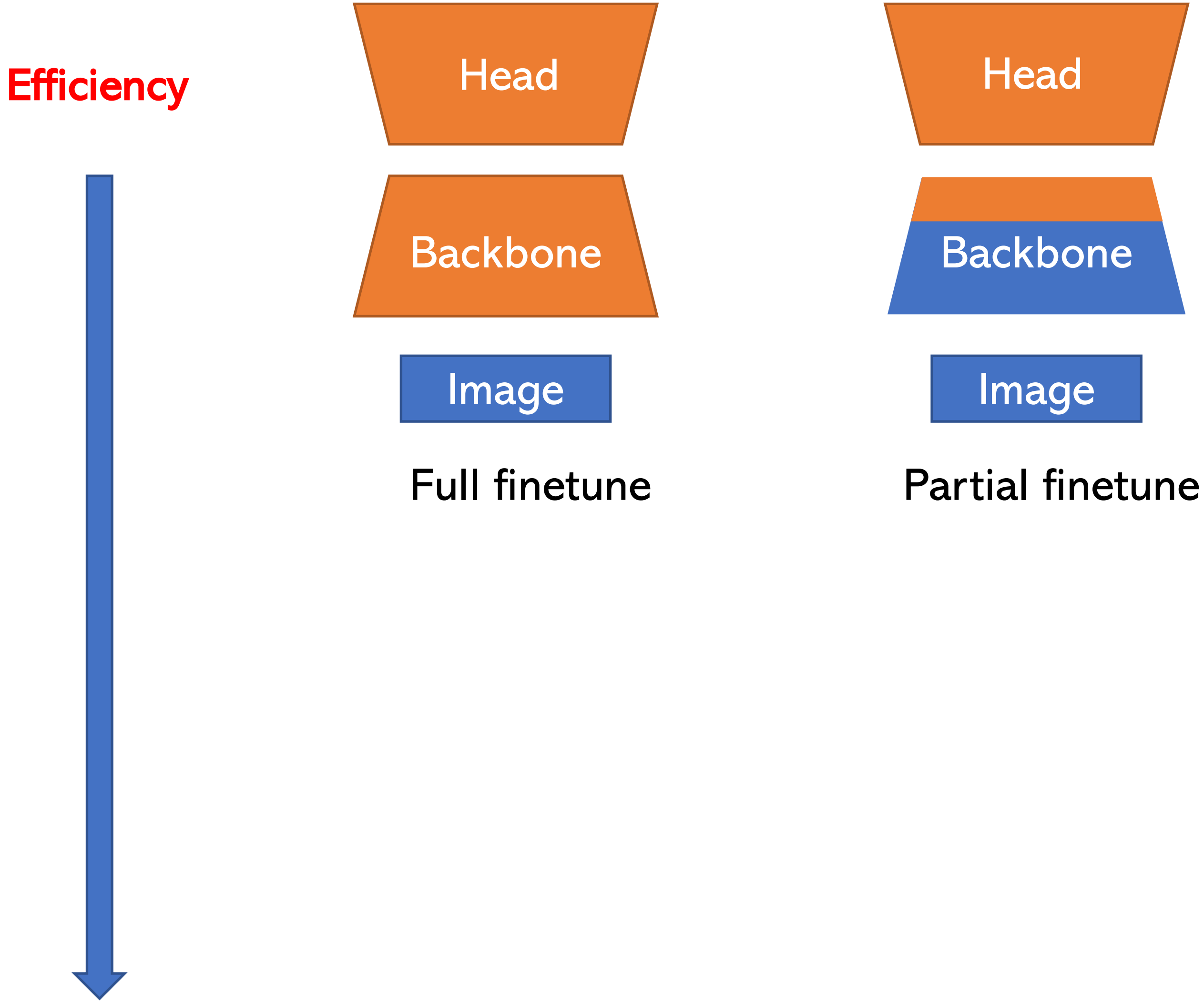
Finetuned

HumanBench: Largest Human-centric Datasets in Academy

Evaluation Protocols: 3 protocols.

Full Finetuning: Finetune all parameters using all in the downstream tasks.

Partial Finetuning: Finetune parameters in the last two layers using all in the downstream tasks.



Full finetune

Partial finetune

Frozen

Finetuned

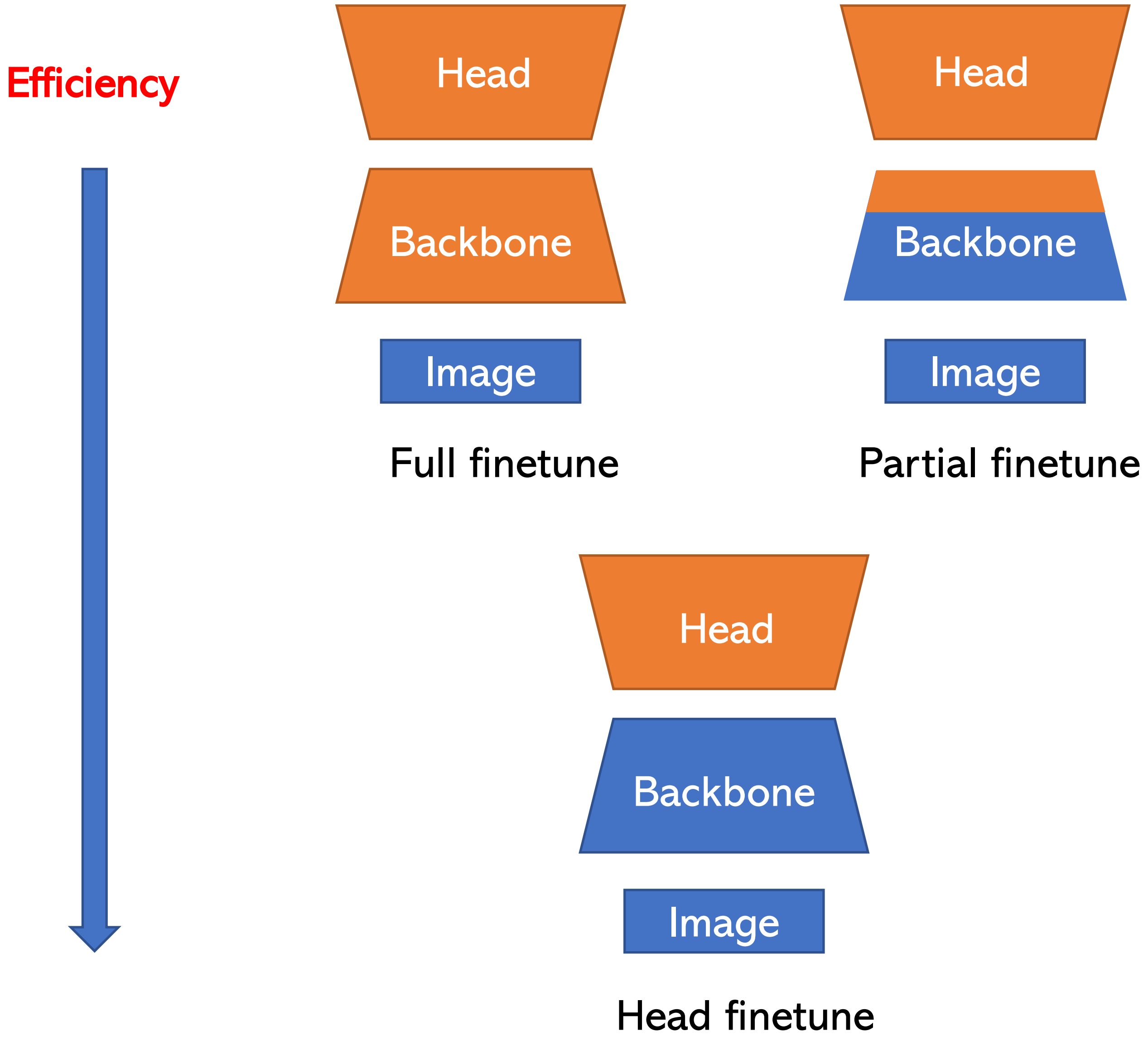
HumanBench: Largest Human-centric Datasets in Academy

Evaluation Protocols: 3 protocols.

Full Finetuning: Finetune all parameters using all in the downstream tasks.

Partial Finetuning: Finetune parameters in the last two layers using all in the downstream tasks.

Head Finetuning: Similar to linear evaluation, only parameters in the task head are finetuned.

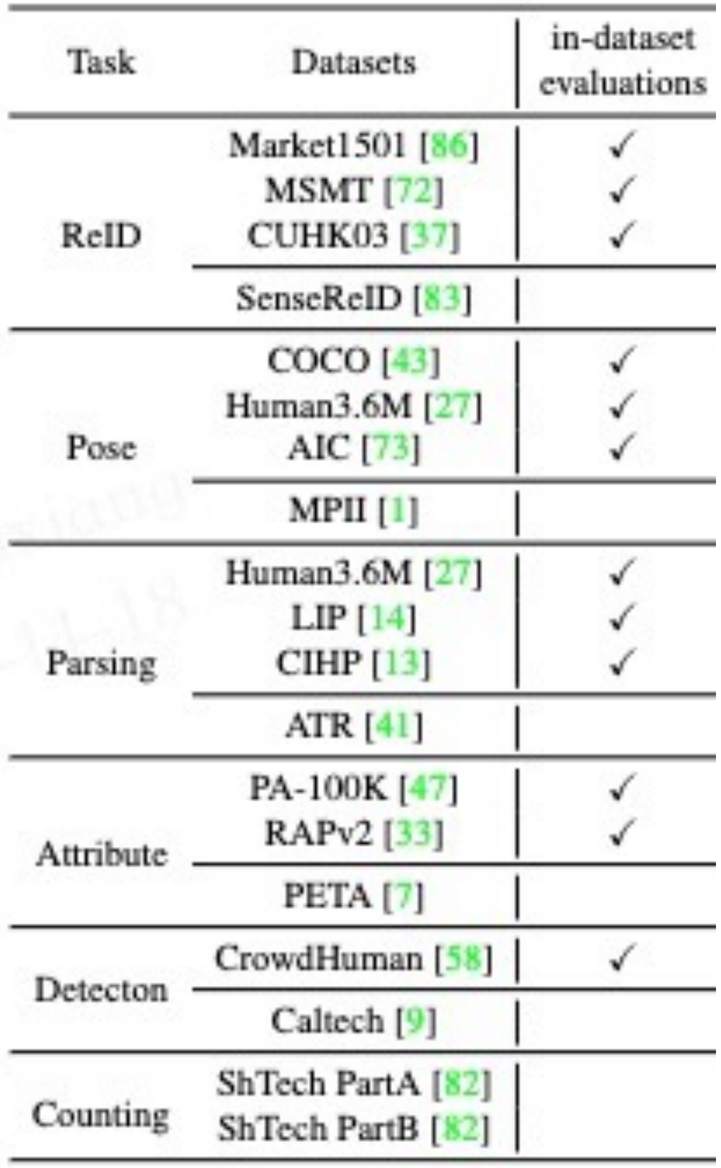


Frozen

Finetuned

HumanBench: Largest Human-centric Datasets in Academy

- In-dataset evaluations: pretraining subset in the pretraining datasets.

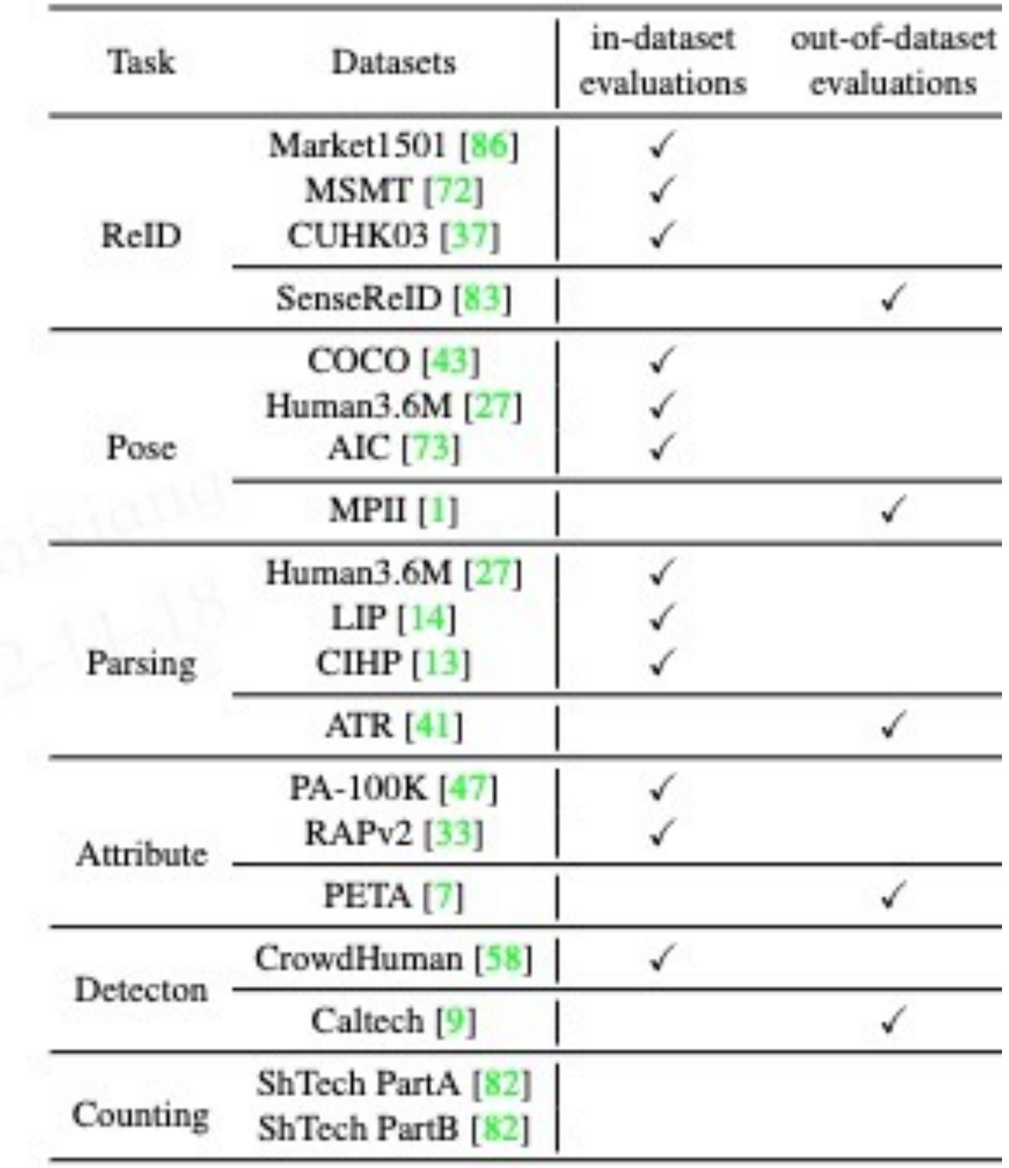


Task	Datasets	in-dataset evaluations
ReID	Market1501 [86]	✓
	MSMT [72]	✓
	CUHK03 [37]	✓
	SenseReID [83]	
Pose	COCO [43]	✓
	Human3.6M [27]	✓
	AIC [73]	✓
	MPII [1]	
Parsing	Human3.6M [27]	✓
	LIP [14]	✓
	CIHP [13]	✓
	ATR [41]	
Attribute	PA-100K [47]	✓
	RAPv2 [33]	✓
	PETA [7]	
Detecton	CrowdHuman [58]	✓
	Caltech [9]	
Counting	ShTech PartA [82]	
	ShTech PartB [82]	

HumanBench: Largest Human-centric Datasets in Academy

- In-dataset evaluations: pretraining subset in the pretraining datasets.

- Out-of-dataset evaluations: pretraining subsets are NOT in the pretraining dataset, but tasks are pretrained.



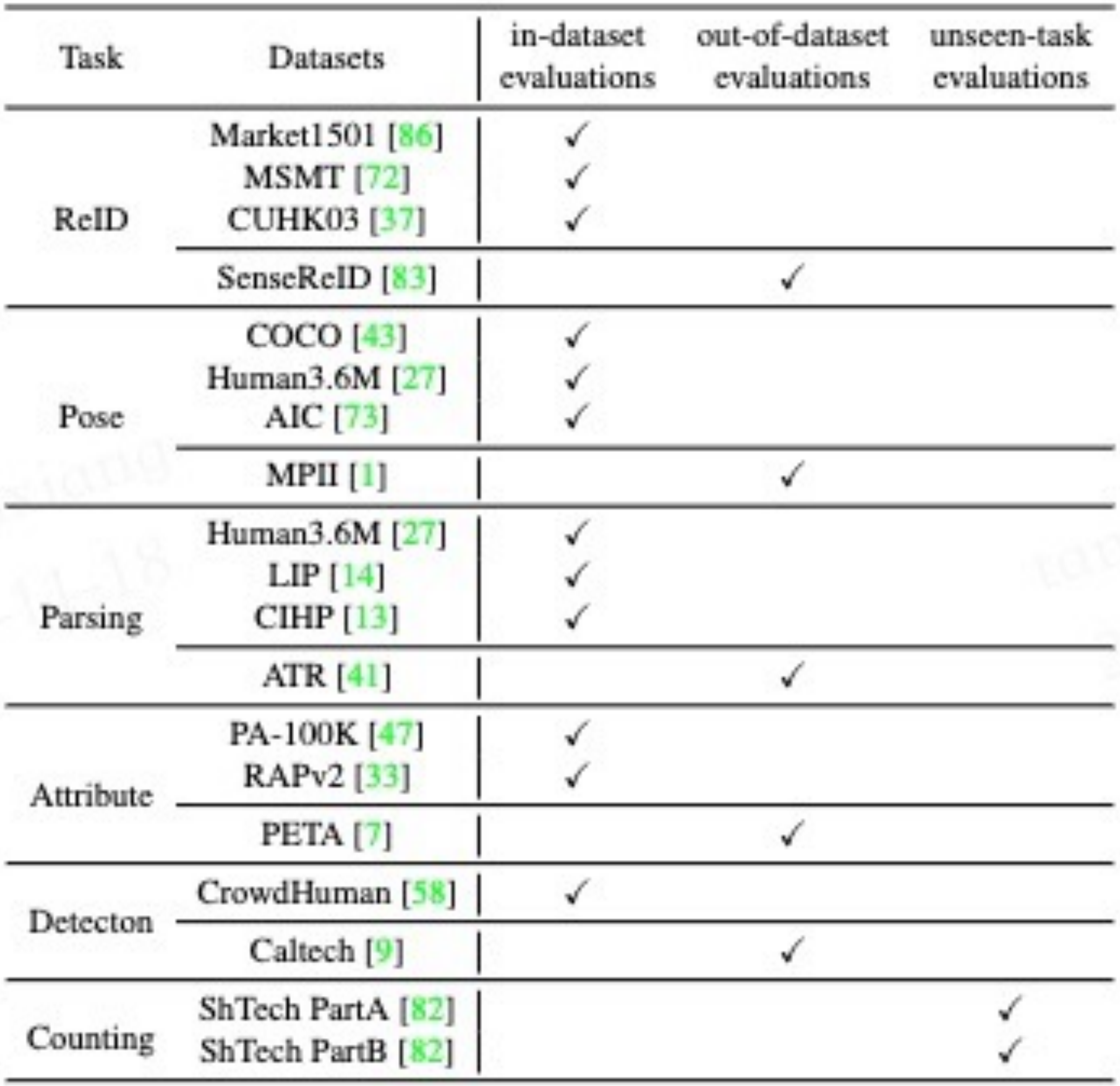
Task	Datasets	in-dataset evaluations	out-of-dataset evaluations
ReID	Market1501 [86]	✓	
	MSMT [72]	✓	
	CUHK03 [37]	✓	
	SenseReID [83]		✓
Pose	COCO [43]	✓	
	Human3.6M [27]	✓	
	AIC [73]	✓	
	MPII [1]		✓
Parsing	Human3.6M [27]	✓	
	LIP [14]	✓	
	CIHP [13]	✓	
	ATR [41]		✓
Attribute	PA-100K [47]	✓	
	RAPv2 [33]	✓	
	PETA [7]		✓
Detecton	CrowdHuman [58]	✓	
	Caltech [9]		✓
Counting	ShTech PartA [82]		
	ShTech PartB [82]		

HumanBench: Largest Human-centric Datasets in Academy

- In-dataset evaluations: pretraining subset in the pretraining datasets.

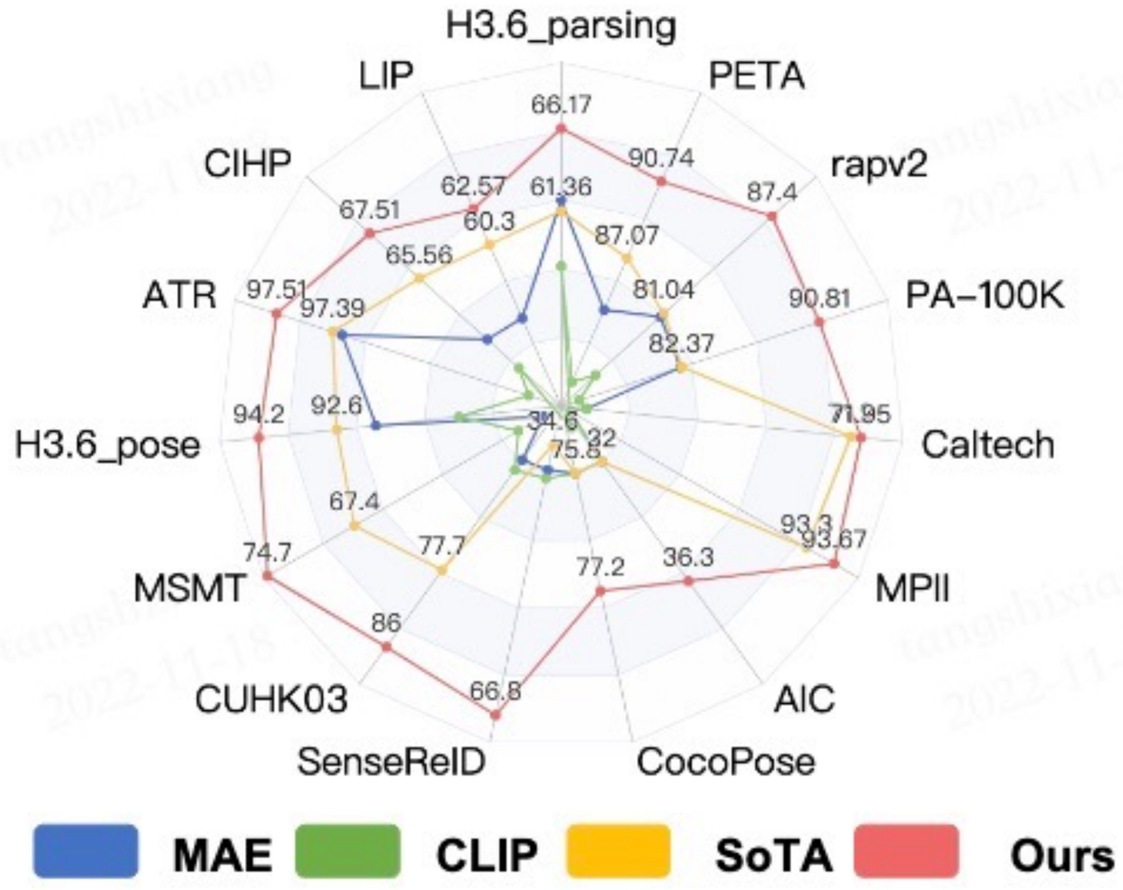
- Out-of-dataset evaluations: pretraining subsets are NOT in the pretraining dataset, but tasks are pretrained.

- Unseen tasks: tasks are NOT in the pretraining datasets.



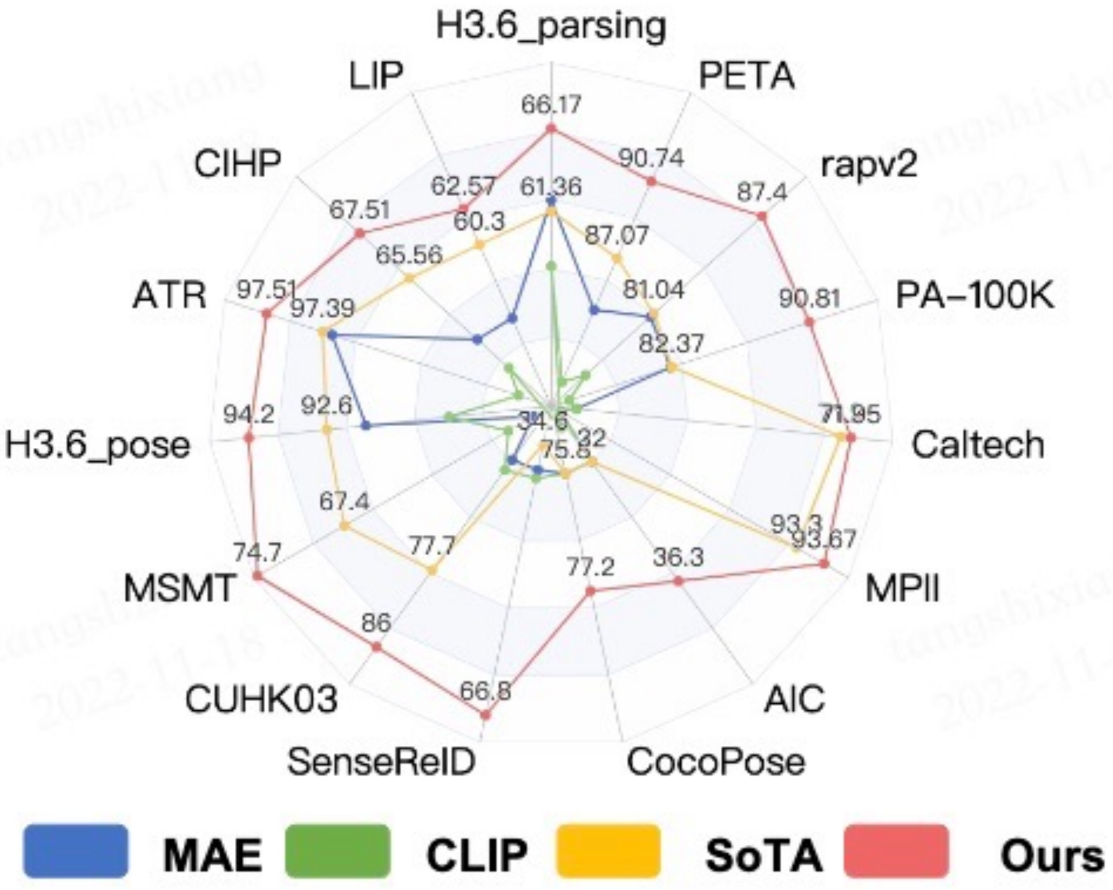
Task	Datasets	in-dataset evaluations	out-of-dataset evaluations	unseen-task evaluations
ReID	Market1501 [86]	✓		
	MSMT [72]	✓		
	CUHK03 [37]	✓		
	SenseReID [83]		✓	
Pose	COCO [43]	✓		
	Human3.6M [27]	✓		
	AIC [73]	✓		
	MPII [1]		✓	
Parsing	Human3.6M [27]	✓		
	LIP [14]	✓		
	CIHP [13]	✓		
	ATR [41]		✓	
Attribute	PA-100K [47]	✓		
	RAPv2 [33]	✓		
	PETA [7]		✓	
Detecton	CrowdHuman [58]	✓		
	Caltech [9]		✓	
Counting	ShTech PartA [82]			✓
	ShTech PartB [82]			✓

HumanBench: Largest Human-centric Datasets in Academy



HumanBench: Largest Human-centric Datasets in Academy

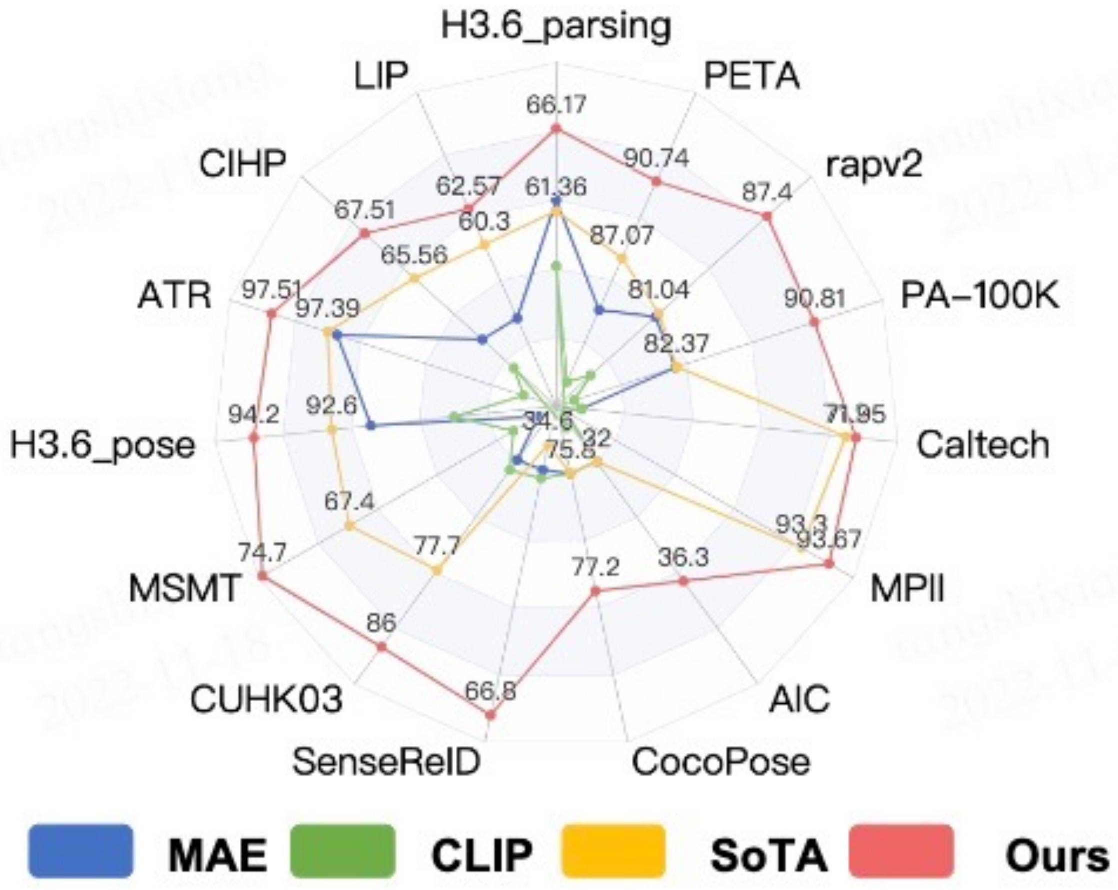
- **Effective than ImageNet and CLIP**
  - MAE>CLIP: Visual-Language datasets are NOT helpful.
  - Ours>MAE: HumanBench are better than ImageNet



HumanBench: Largest Human-centric Datasets in Academy

- **Effective than ImageNet and CLIP**
  - MAE>CLIP: Visual-Language datasets are NOT helpful.
  - Ours>MAE: HumanBench are better than ImageNet

- **Push the limits of states-of-the-art methods on human-centric tasks**
  - Better results than States-of-the-art methods on 17 datasets.



Outline

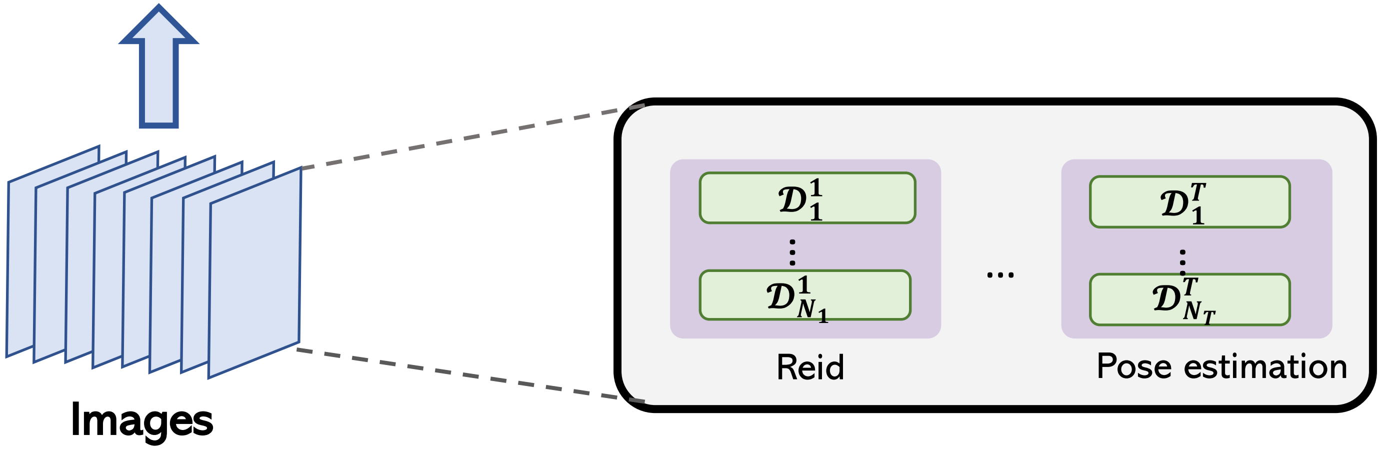
- Why Human-Centric Foundation Model ?

- HumanBench: Largest Human-centric Datasets in Academy

- PATH: A Projector Assisted preTraining with Hierarchical weight sharing

- Experimental Results and Future Work

PATH: A Projector Assisted preTraining with Hierarchical weight sharing

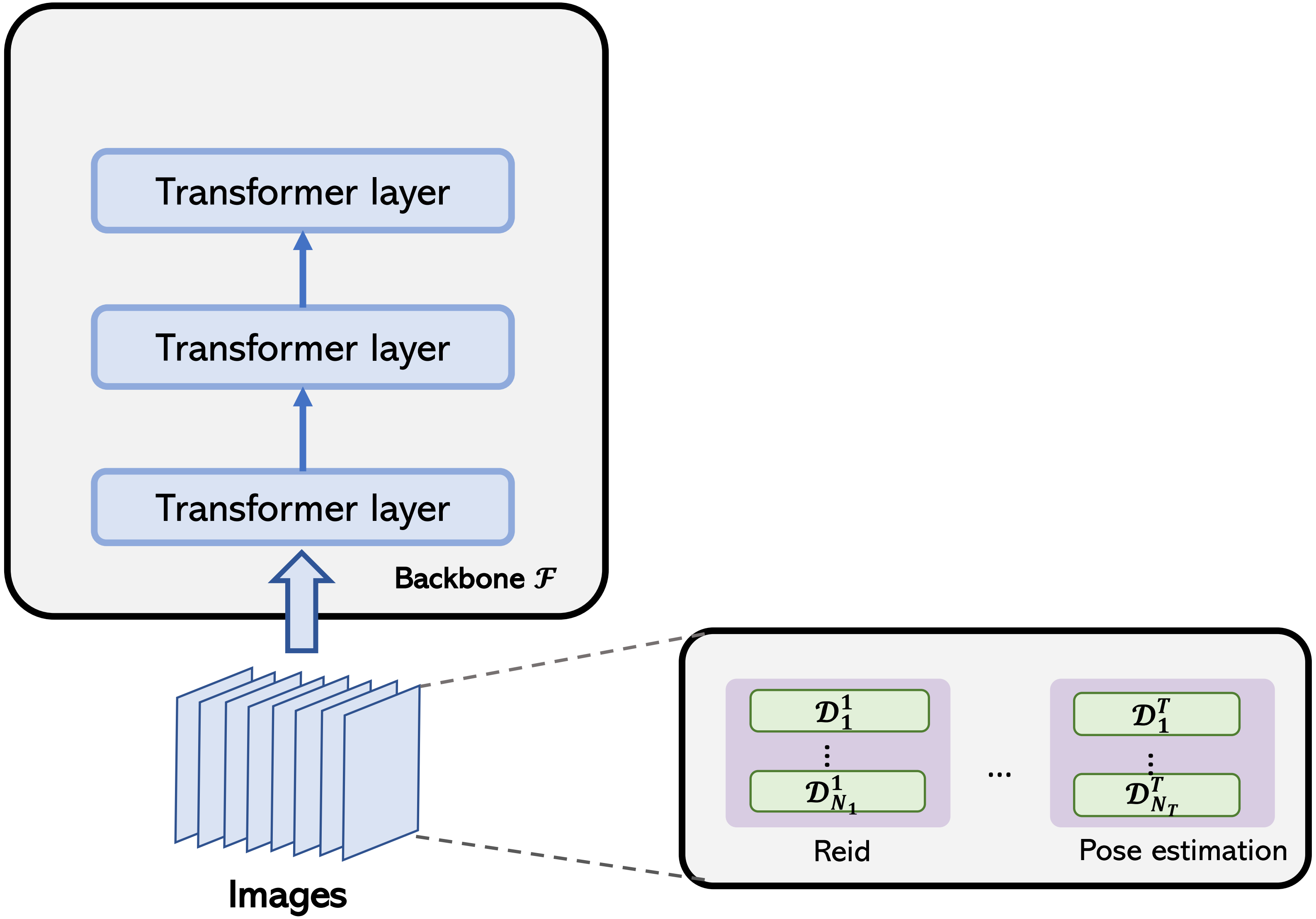


🟦 Shared by all datasets

🟪 Shared by datasets in the task

🟩 Not shared

PATH: A Projector Assisted preTraining with Hierarchical weight sharing

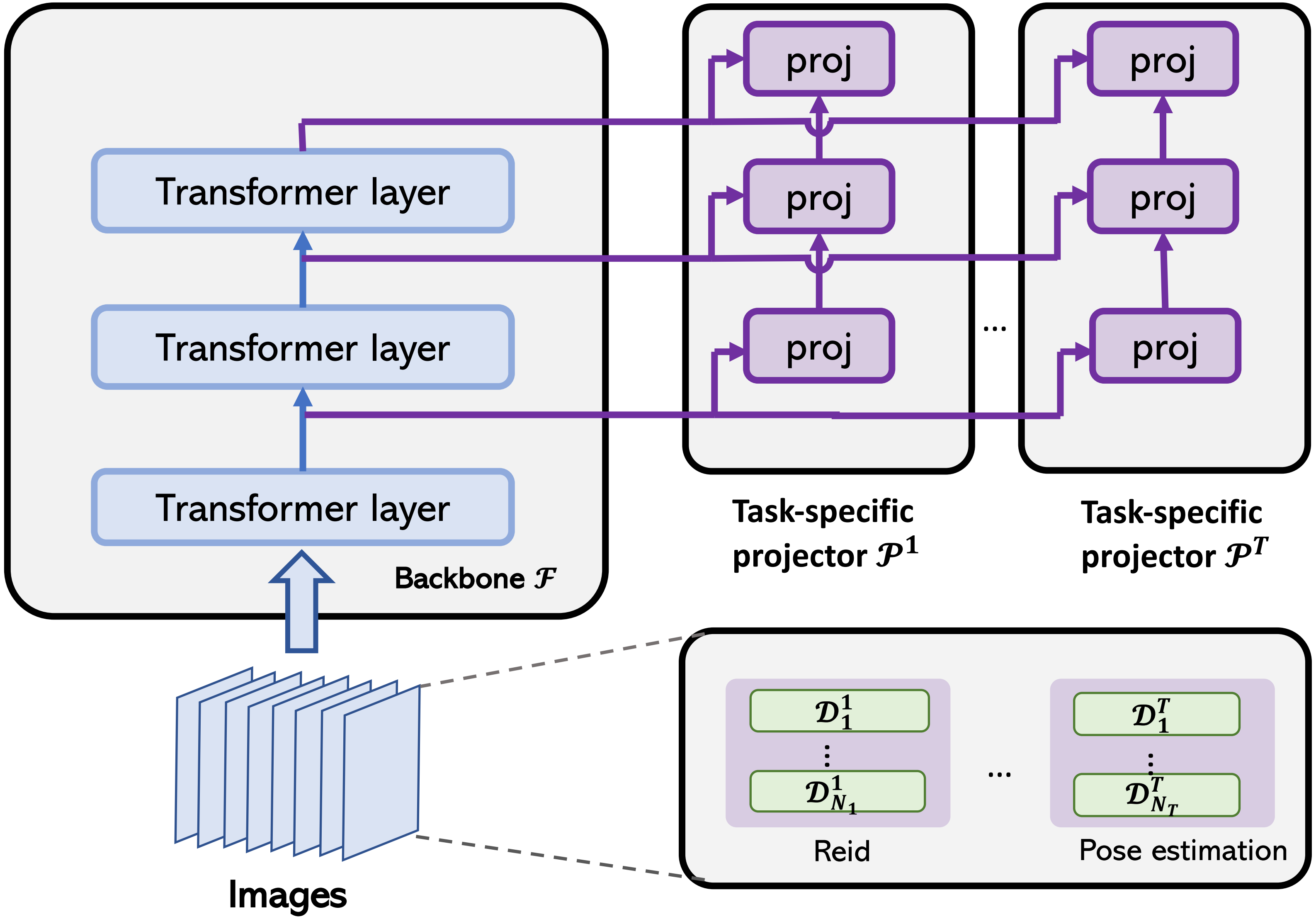


🟦 Shared by all datasets

🟪 Shared by datasets in the task

🟩 Not shared

PATH: A Projector Assisted preTraining with Hierarchical weight sharing

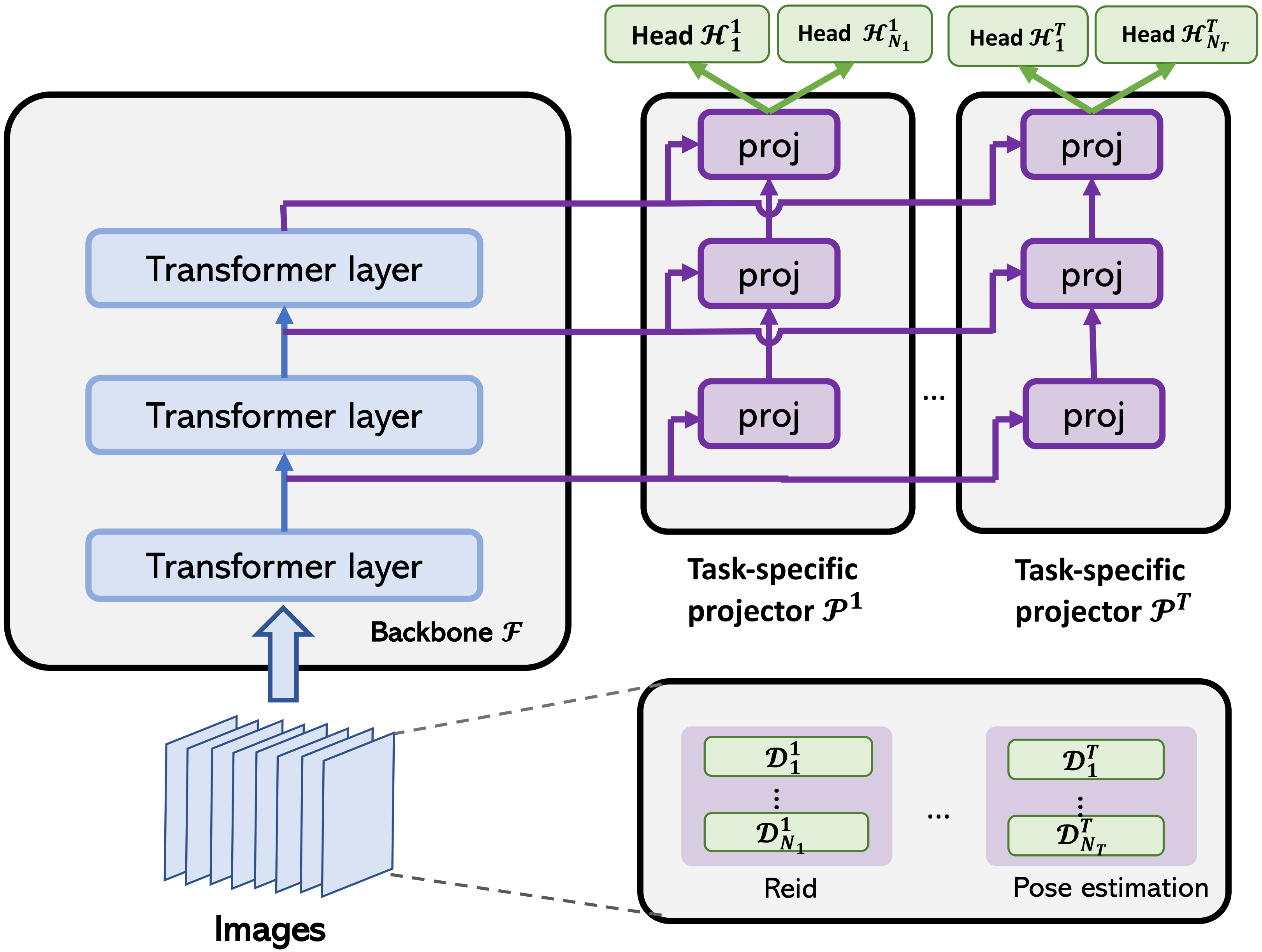


🟦 Shared by all datasets

🟪 Shared by datasets in the task

🟩 Not shared

PATH: A Projector Assisted preTraining with Hierarchical weight sharing

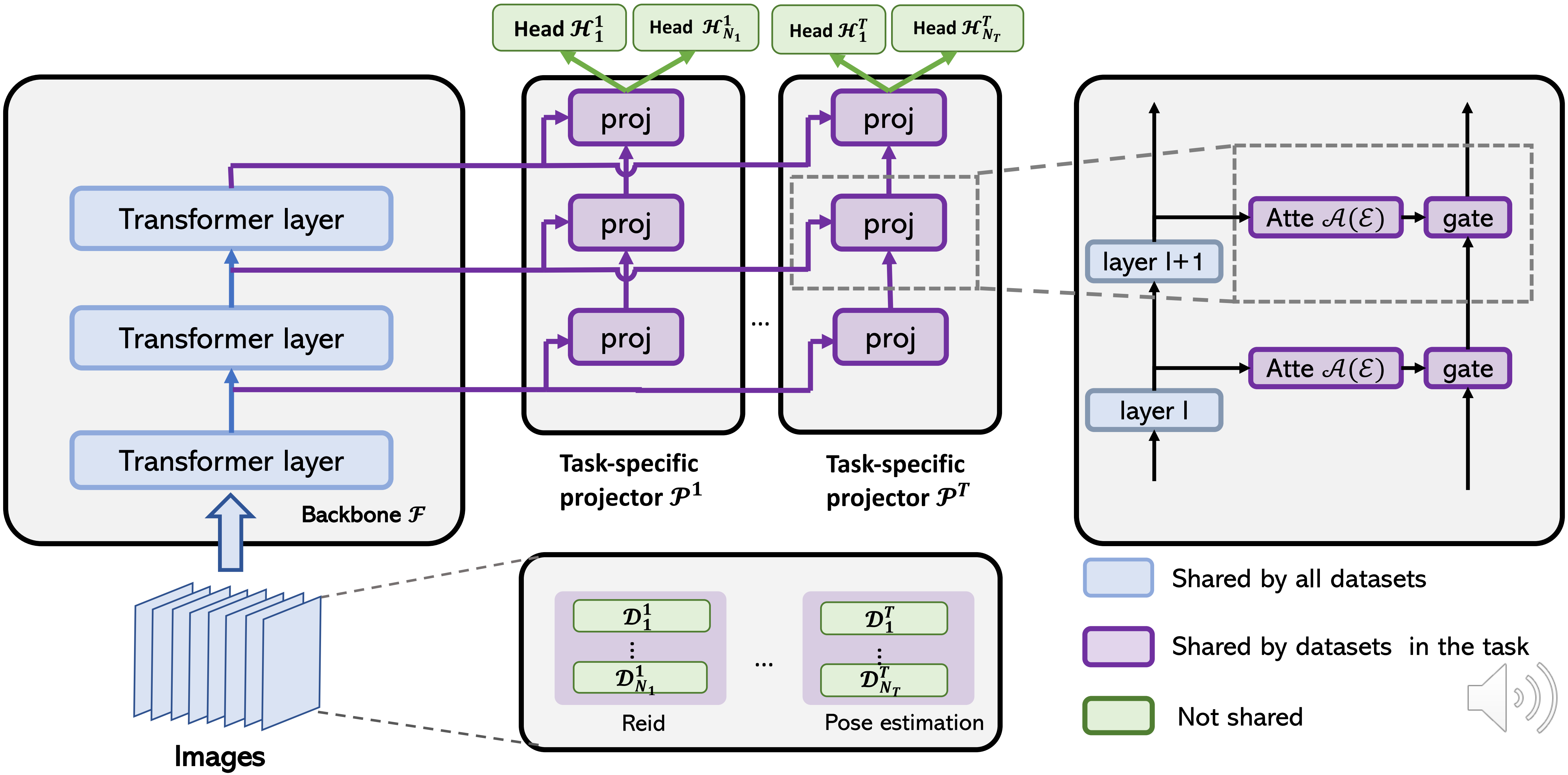


🟦 Shared by all datasets

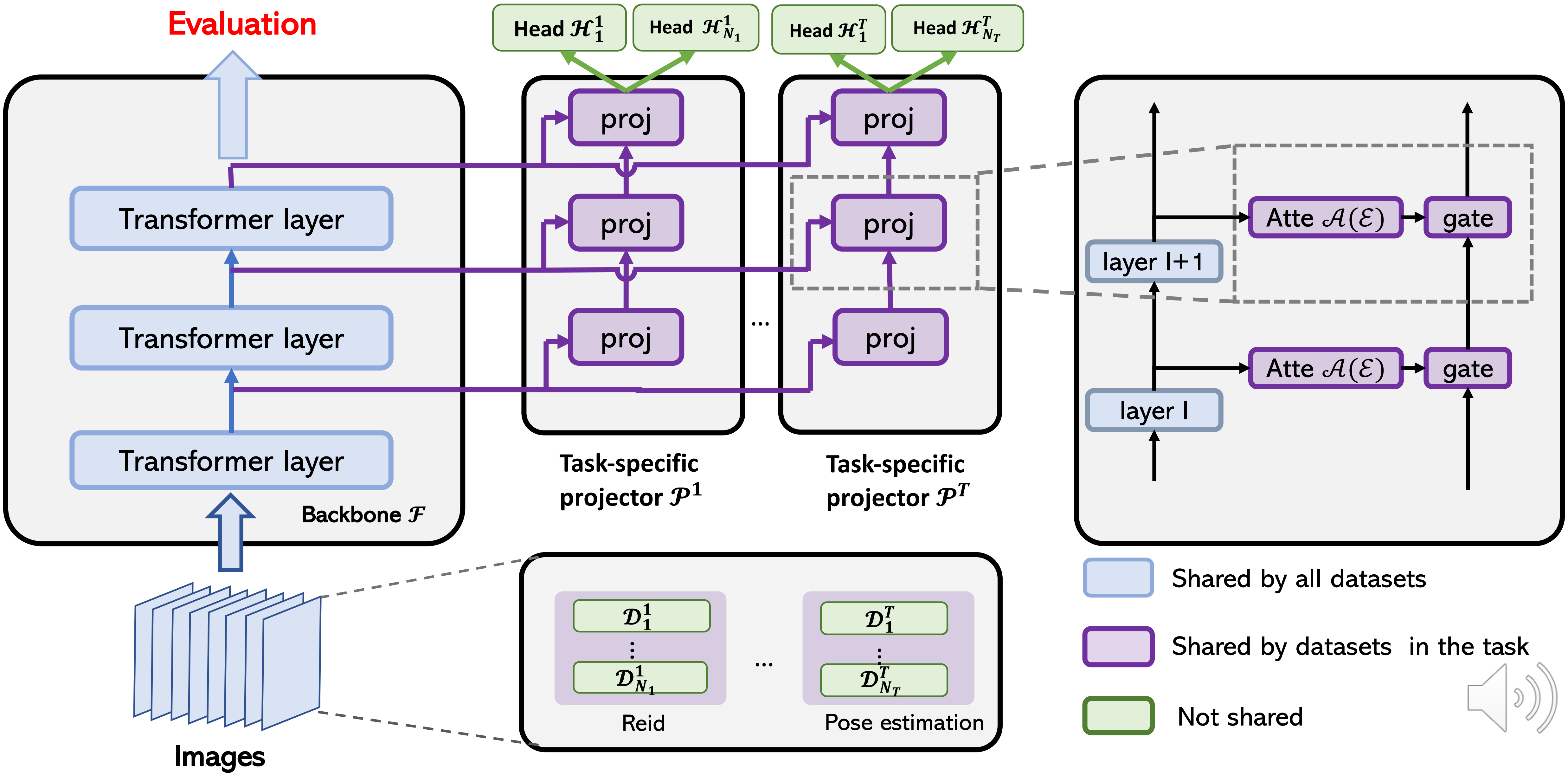
🟪 Shared by datasets in the task

🟩 Not shared

PATH: A Projector Assisted preTraining with Hierarchical weight sharing



PATH: A Projector Assisted preTraining with Hierarchical weight sharing



Outline

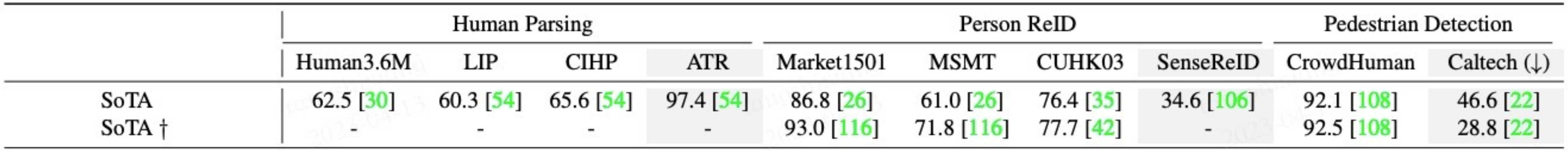
- Why Human-Centric Foundation Model ?

- HumanBench: Largest Human-centric Datasets in Academy

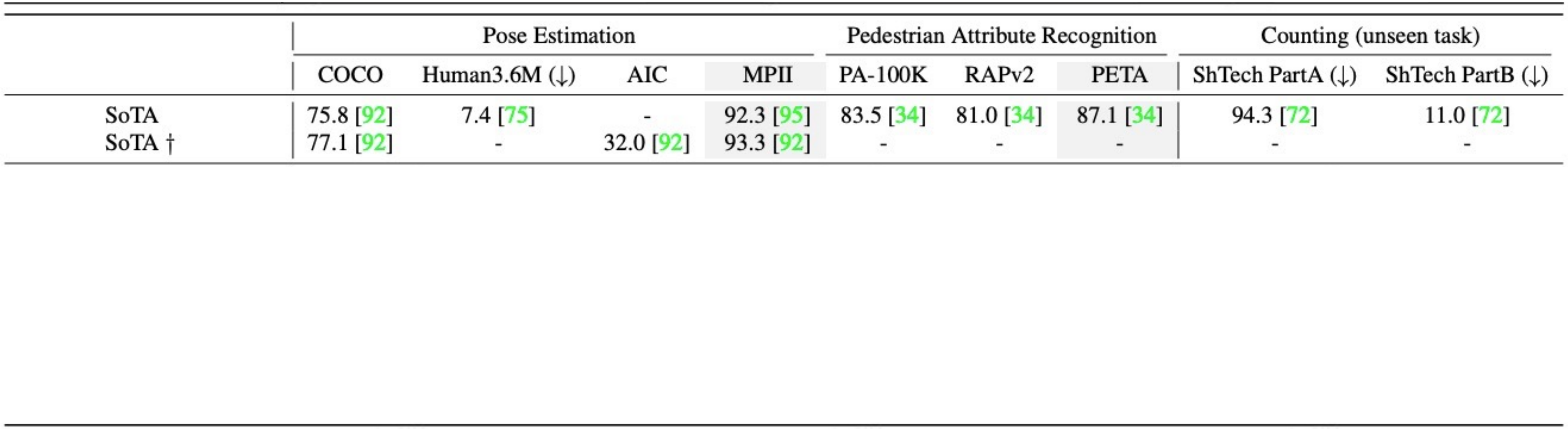
- PATH: A Projector Assisted preTraining with Hierarchical weight sharing

- **Experimental Results and Future Work**

Experimental Results

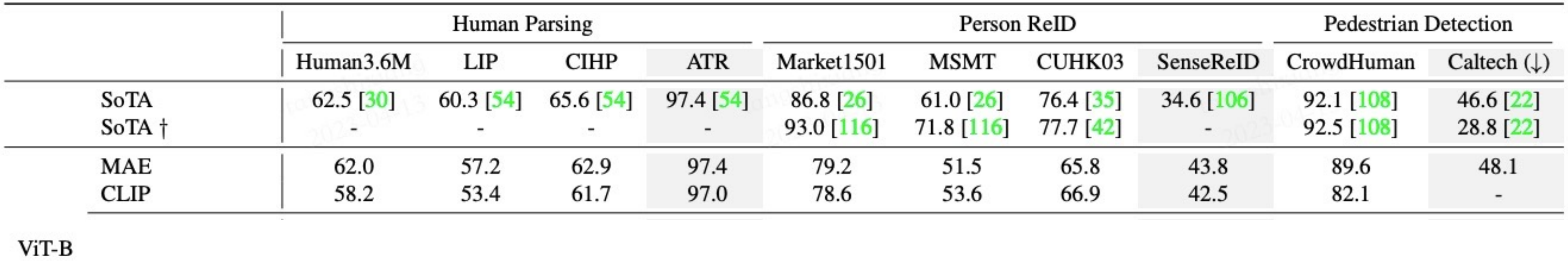


	Human Parsing				Person ReID				Pedestrian Detection	
	Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]



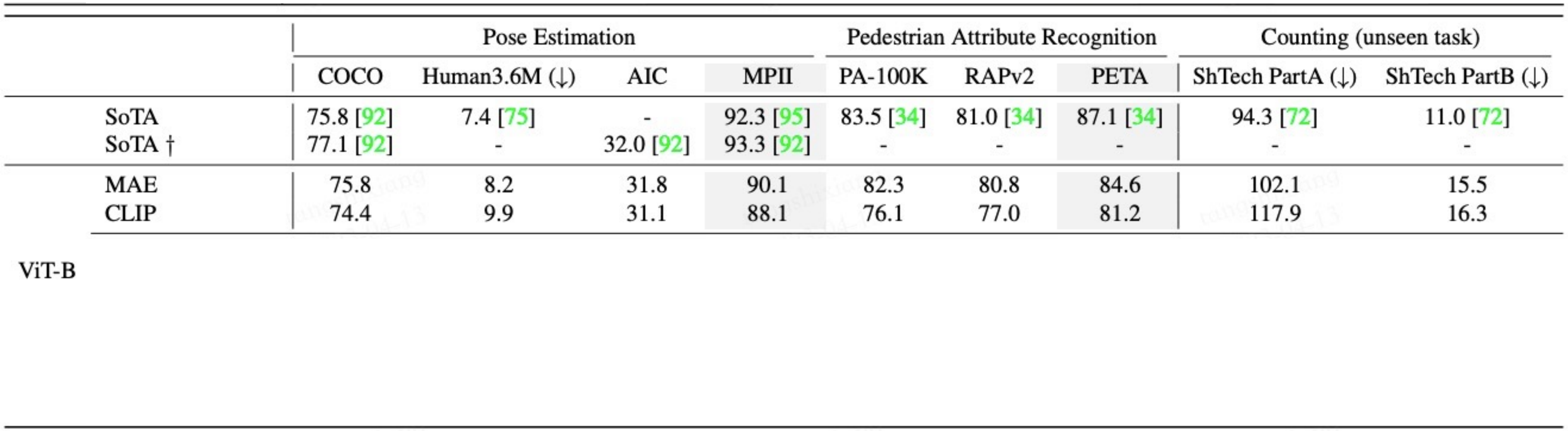
	Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
	COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-

Experimental Results



	Human Parsing				Person ReID				Pedestrian Detection	
	Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-

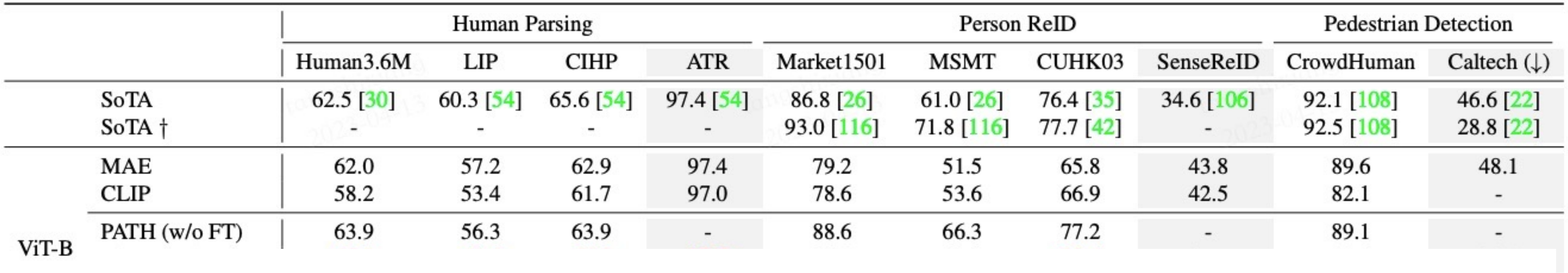
ViT-B



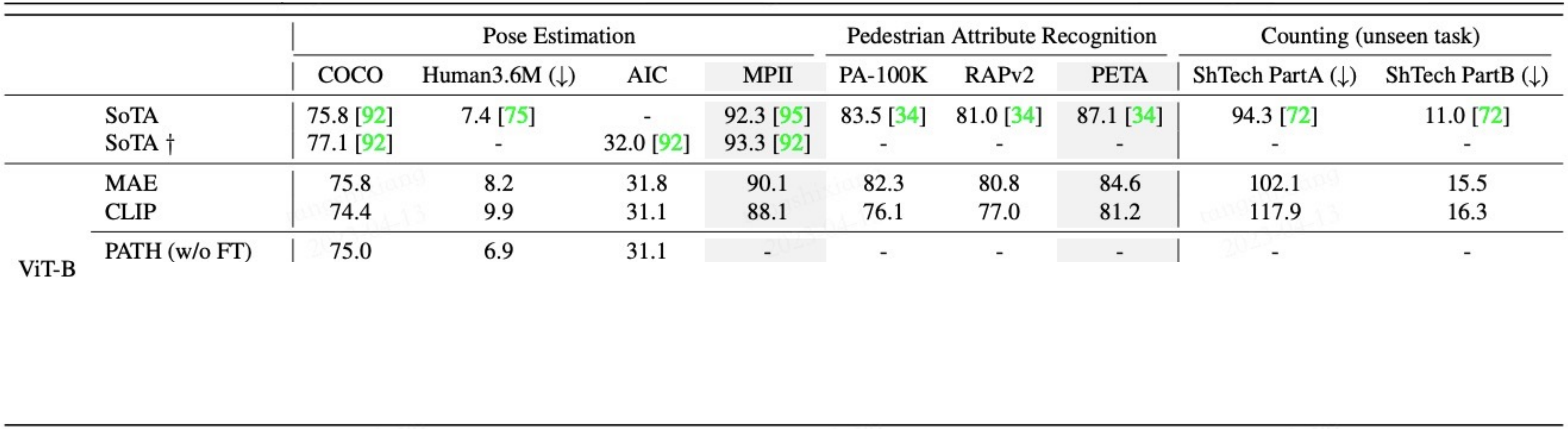
	Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
	COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3

ViT-B

Experimental Results

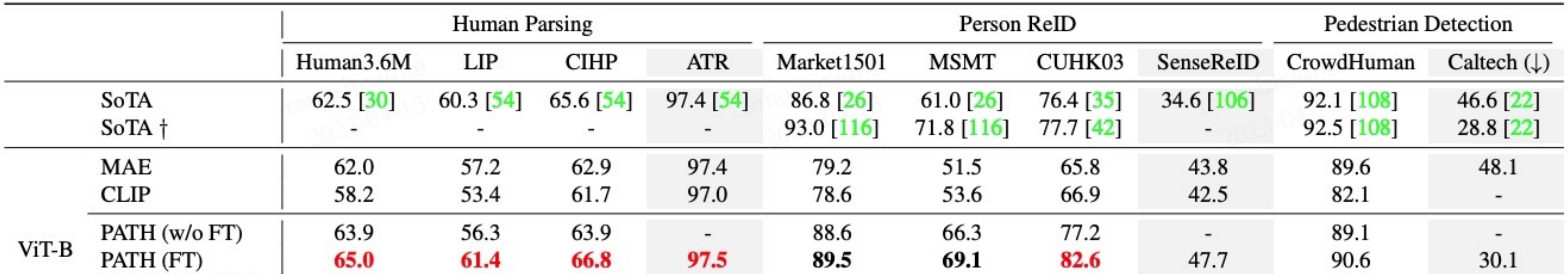


		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
ViT-B	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-

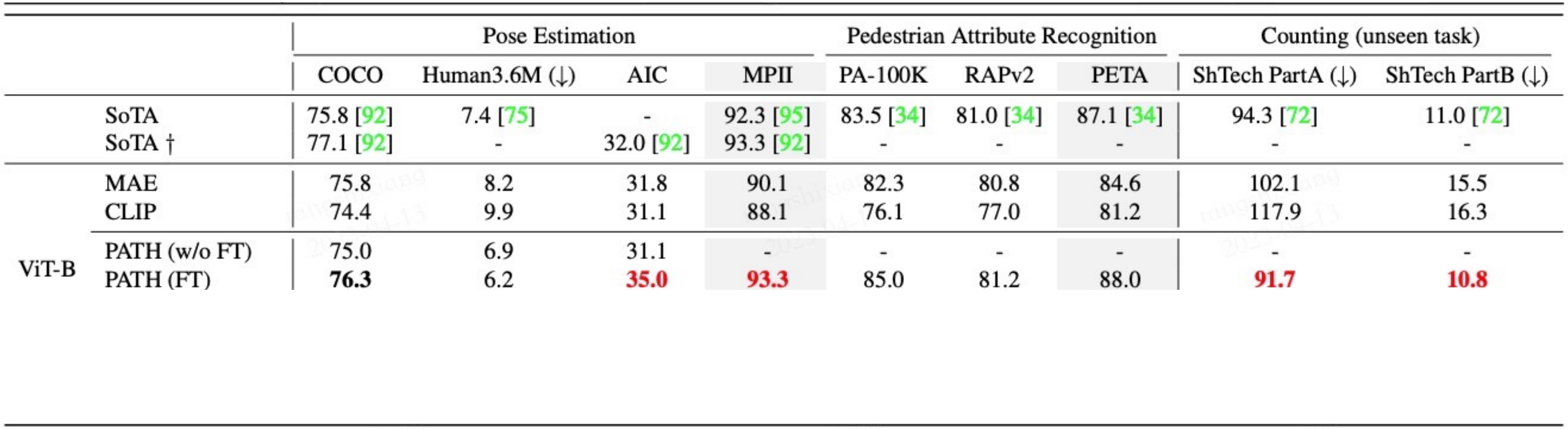


		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-

Experimental Results

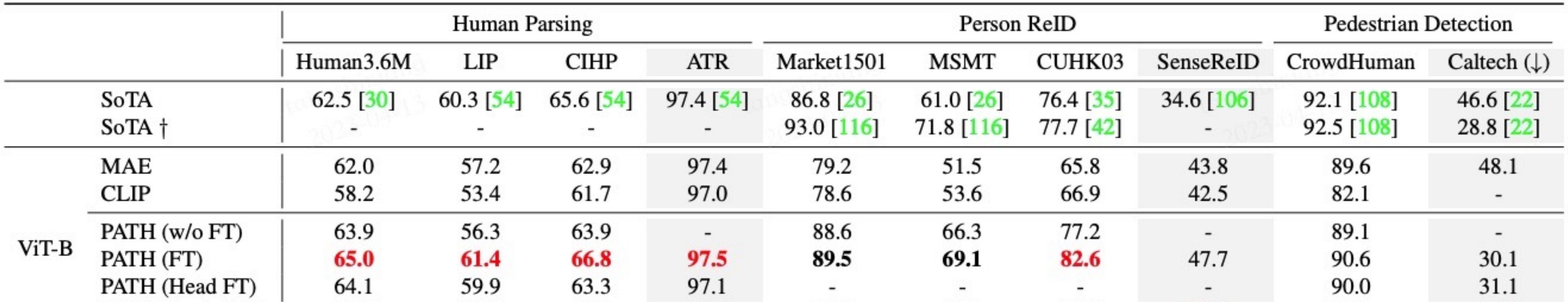


		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
ViT-B	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-
	PATH (FT)	**65.0**	**61.4**	**66.8**	**97.5**	**89.5**	**69.1**	**82.6**	47.7	90.6	30.1

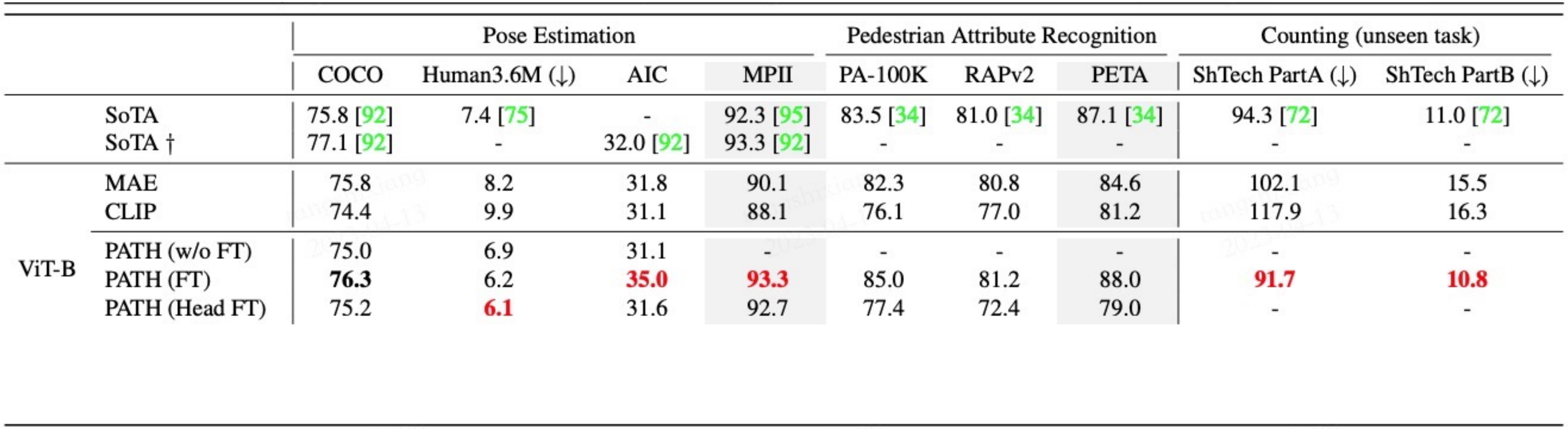


		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
ViT-B	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-
	PATH (FT)	**76.3**	6.2	**35.0**	**93.3**	85.0	81.2	88.0	**91.7**	**10.8**

Experimental Results

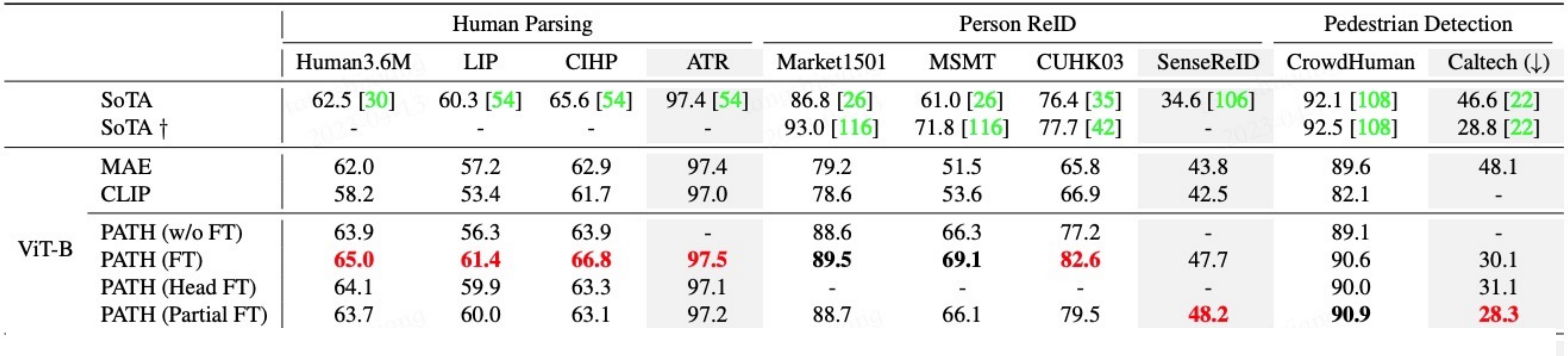


		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
ViT-B	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-
	PATH (FT)	**65.0**	**61.4**	**66.8**	**97.5**	**89.5**	**69.1**	**82.6**	47.7	90.6	30.1
	PATH (Head FT)	64.1	59.9	63.3	97.1	-	-	-	-	90.0	31.1

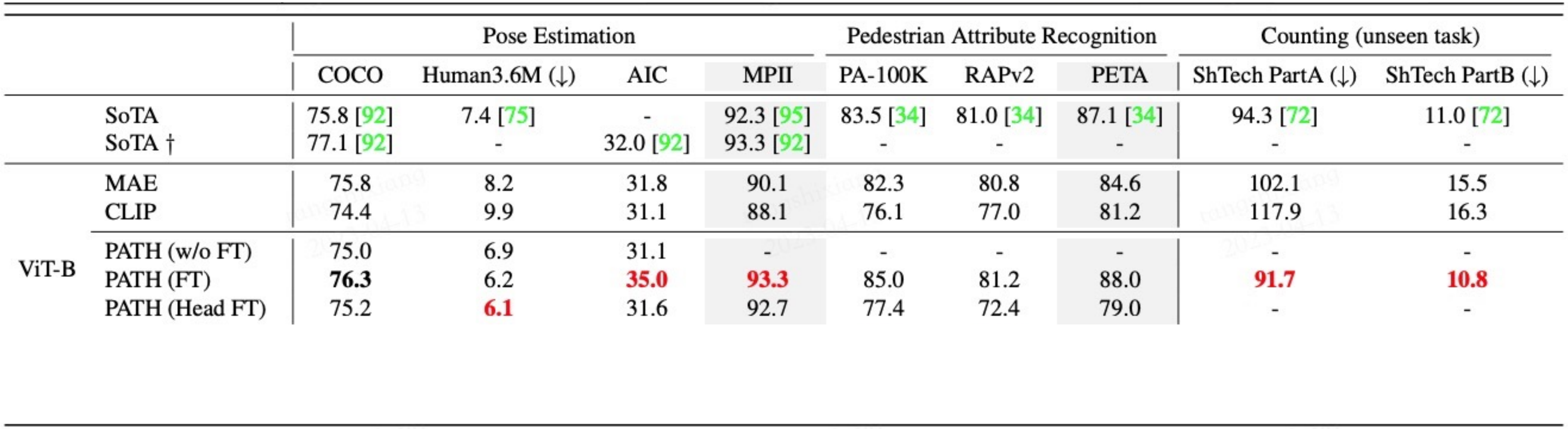


		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-
	PATH (FT)	**76.3**	6.2	**35.0**	**93.3**	85.0	81.2	88.0	**91.7**	**10.8**
	PATH (Head FT)	75.2	**6.1**	31.6	92.7	77.4	72.4	79.0	-	-

Experimental Results

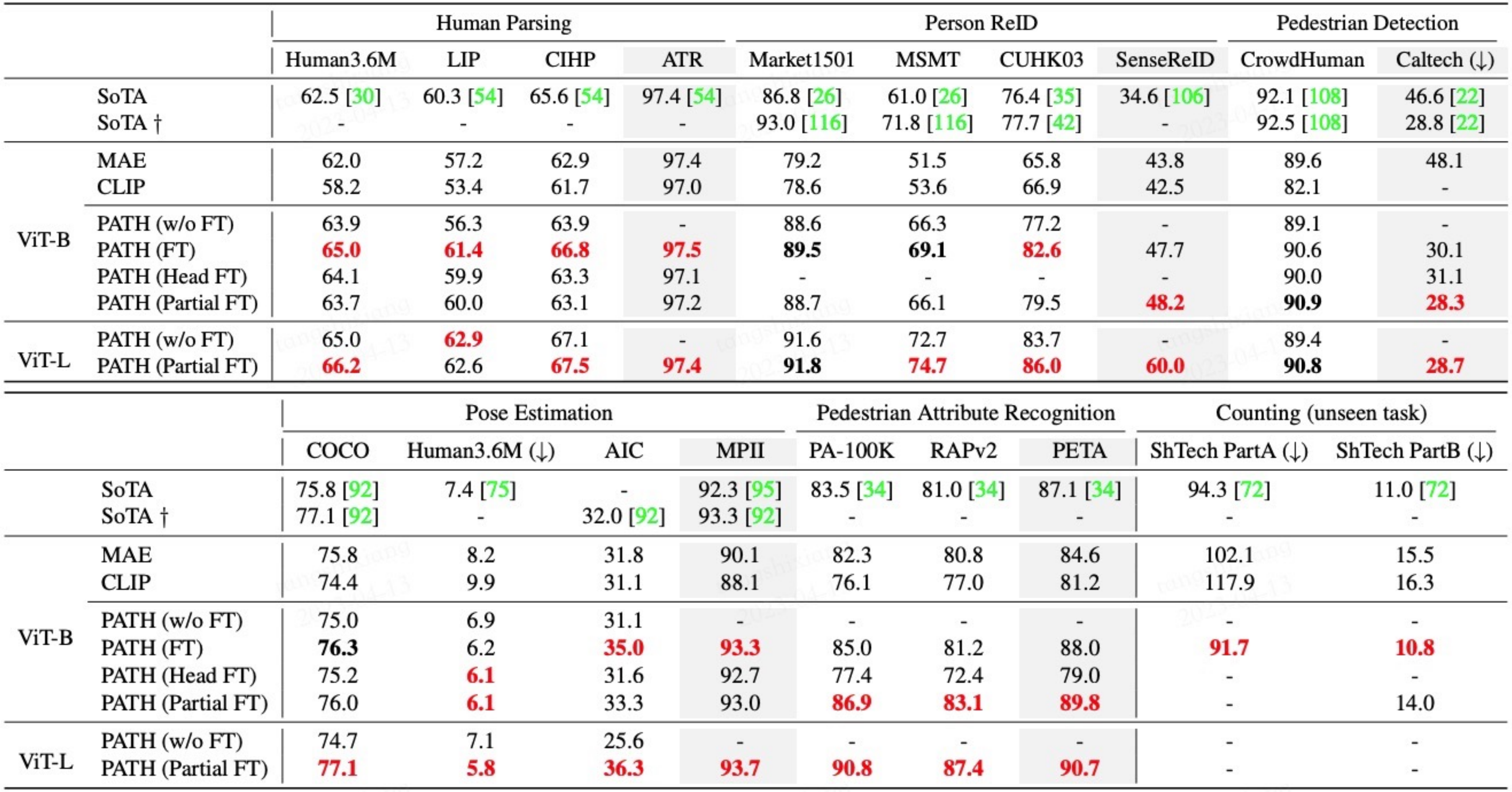


		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
ViT-B	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-
	PATH (FT)	**65.0**	**61.4**	**66.8**	**97.5**	**89.5**	**69.1**	**82.6**	47.7	90.6	30.1
	PATH (Head FT)	64.1	59.9	63.3	97.1	-	-	-	-	90.0	31.1
	PATH (Partial FT)	63.7	60.0	63.1	97.2	88.7	66.1	79.5	**48.2**	**90.9**	**28.3**



		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-
	PATH (FT)	**76.3**	6.2	**35.0**	**93.3**	85.0	81.2	88.0	**91.7**	**10.8**
	PATH (Head FT)	75.2	**6.1**	31.6	92.7	77.4	72.4	79.0	-	-

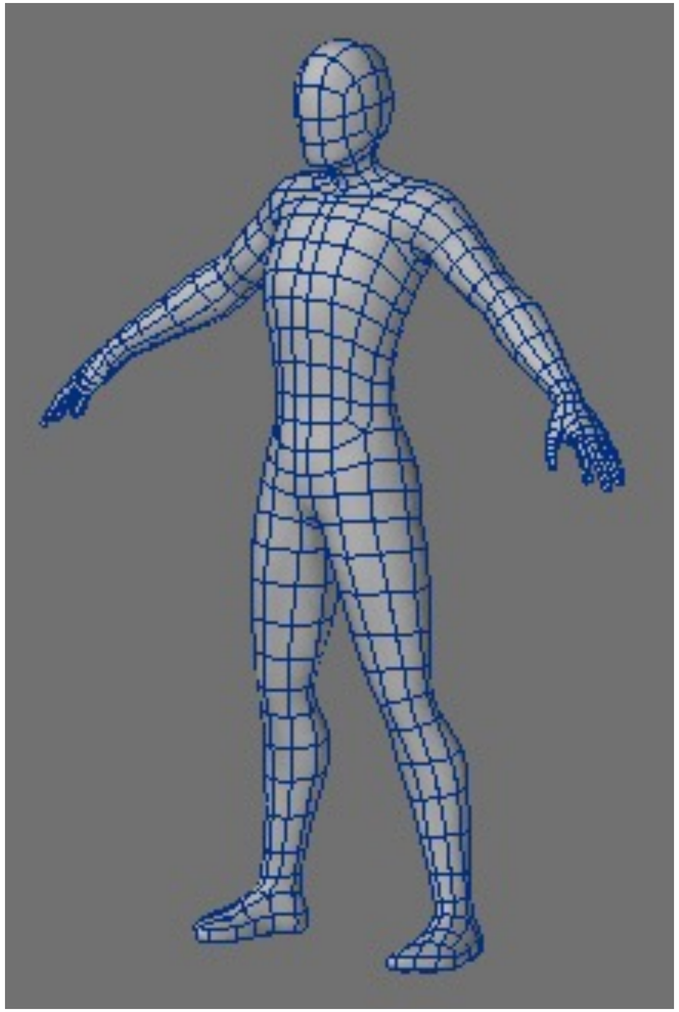
Experimental Results



		Human Parsing				Person ReID				Pedestrian Detection	
		Human3.6M	LIP	CIHP	ATR	Market1501	MSMT	CUHK03	SenseReID	CrowdHuman	Caltech (↓)
	SoTA	62.5 [30]	60.3 [54]	65.6 [54]	97.4 [54]	86.8 [26]	61.0 [26]	76.4 [35]	34.6 [106]	92.1 [108]	46.6 [22]
	SoTA †	-	-	-	-	93.0 [116]	71.8 [116]	77.7 [42]	-	92.5 [108]	28.8 [22]
ViT-B	MAE	62.0	57.2	62.9	97.4	79.2	51.5	65.8	43.8	89.6	48.1
	CLIP	58.2	53.4	61.7	97.0	78.6	53.6	66.9	42.5	82.1	-
	PATH (w/o FT)	63.9	56.3	63.9	-	88.6	66.3	77.2	-	89.1	-
	PATH (FT)	**65.0**	**61.4**	**66.8**	**97.5**	**89.5**	**69.1**	**82.6**	47.7	90.6	30.1
	PATH (Head FT)	64.1	59.9	63.3	97.1	-	-	-	-	90.0	31.1
	PATH (Partial FT)	63.7	60.0	63.1	97.2	88.7	66.1	79.5	**48.2**	**90.9**	**28.3**
ViT-L	PATH (w/o FT)	65.0	**62.9**	67.1	-	91.6	72.7	83.7	-	89.4	-
	PATH (Partial FT)	**66.2**	62.6	**67.5**	**97.4**	**91.8**	**74.7**	**86.0**	**60.0**	**90.8**	**28.7**

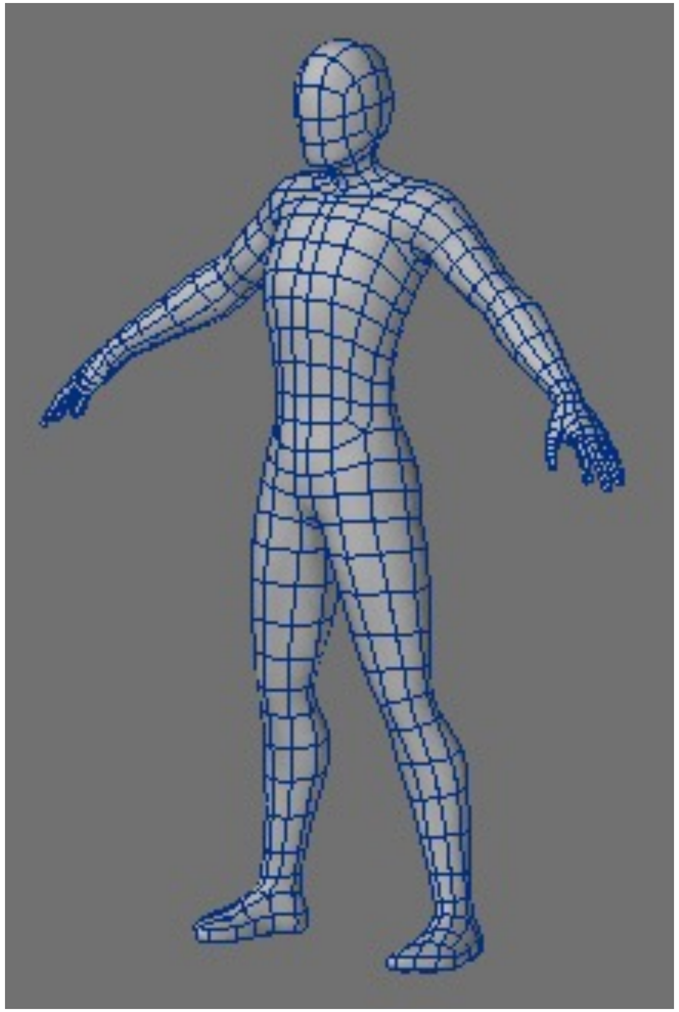
		Pose Estimation				Pedestrian Attribute Recognition			Counting (unseen task)	
		COCO	Human3.6M (↓)	AIC	MPII	PA-100K	RAPv2	PETA	ShTech PartA (↓)	ShTech PartB (↓)
	SoTA	75.8 [92]	7.4 [75]	-	92.3 [95]	83.5 [34]	81.0 [34]	87.1 [34]	94.3 [72]	11.0 [72]
	SoTA †	77.1 [92]	-	32.0 [92]	93.3 [92]	-	-	-	-	-
ViT-B	MAE	75.8	8.2	31.8	90.1	82.3	80.8	84.6	102.1	15.5
	CLIP	74.4	9.9	31.1	88.1	76.1	77.0	81.2	117.9	16.3
	PATH (w/o FT)	75.0	6.9	31.1	-	-	-	-	-	-
	PATH (FT)	**76.3**	6.2	**35.0**	**93.3**	85.0	81.2	88.0	**91.7**	**10.8**
	PATH (Head FT)	75.2	**6.1**	31.6	92.7	77.4	72.4	79.0	-	-
	PATH (Partial FT)	76.0	**6.1**	33.3	93.0	**86.9**	**83.1**	**89.8**	-	14.0
ViT-L	PATH (w/o FT)	74.7	7.1	25.6	-	-	-	-	-	-
	PATH (Partial FT)	**77.1**	**5.8**	**36.3**	**93.7**	**90.8**	**87.4**	**90.7**	-	-

Future Work



Future Work

1. How to learn general human-centric features with self-supervised learning framework?

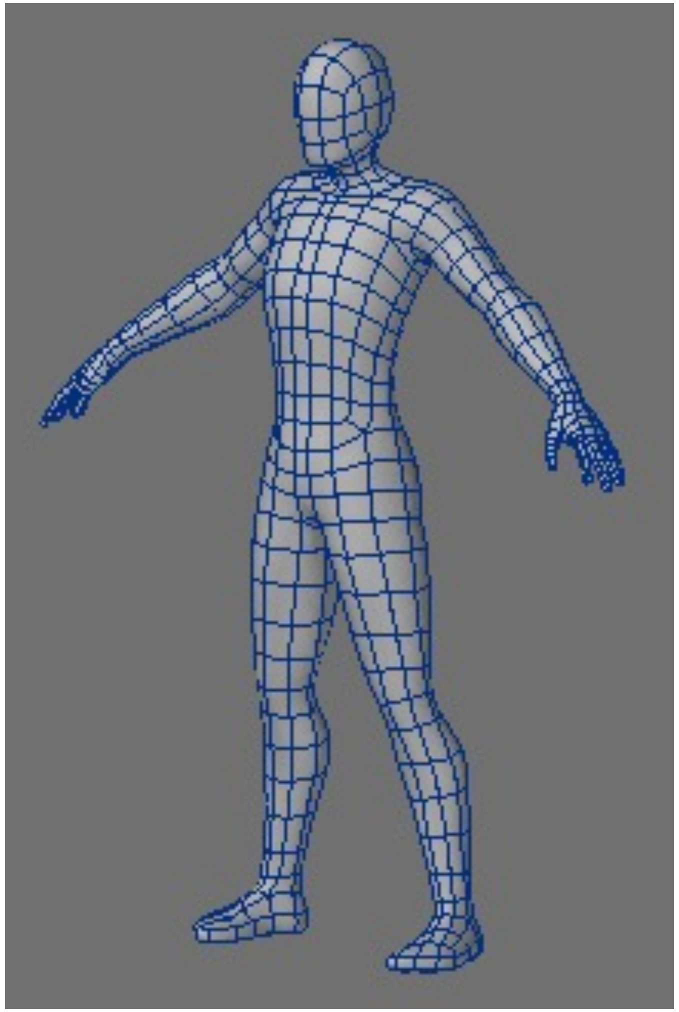


Future Work

1. How to learn general human-centric features with self-supervised learning framework?

2. Multimodal and 2D-3D aligned Human-Centric Model.
    - Text, Video, Point Cloud
    - Body, Hand, Face

3. Foundation-driven 4D Digital Human Generation.



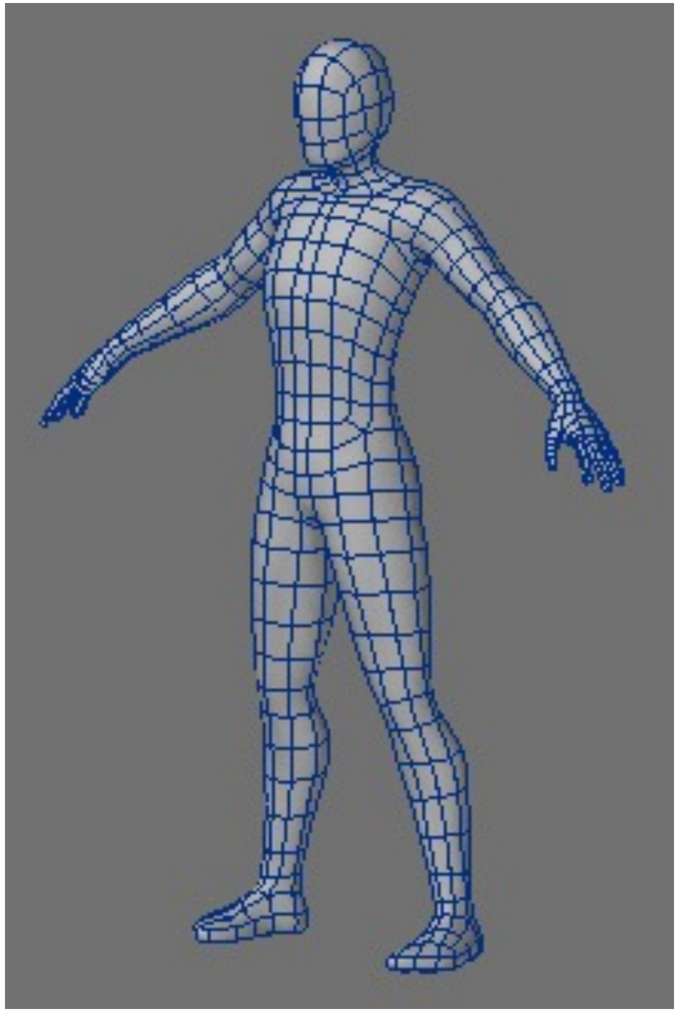
Future Work

1. How to learn general human-centric features with self-supervised learning framework?

2. Multimodal and 2D-3D aligned Human-Centric Model.

- Text, Video, Point Cloud
- Body, Hand, Face

3. Foundation-driven 4D Digital Human Generation.



Thank you!

Code