LEVERAGING HIGH-RESOLUTION VISION TRANSFORMERS FOR WHOLE-BODY HUMAN POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-body human pose estimation requires accurate localization of dense body, face, hand, and foot landmarks while maintaining efficiency as the number of people increases. Existing approaches face a fundamental trade-off. Top-down methods apply high-resolution Vision Transformers to individual person crops, providing detailed spatial representations but incurring computation for each detected person. In contrast, shared-feature multi-person methods process all instances within a common image representation but struggle to maintain both accuracy and speed for dense whole-body landmarks. We present HR-ViP, a top-down framework that efficiently leverages high-resolution Vision Transformers for fine-grained whole-body localization and multi-person inference. HR-ViP introduces Train-Time Random Token Subsampling (TRTS), an efficient strategy for adapting ViTs to high-resolution inputs, together with feature-map distillation to transfer their spatial representations to smaller student models. Combined with batched person-crop processing, the distilled models enable efficient multi-person whole-body pose estimation while retaining the benefits of high-resolution representations. In the single-person setting on COCO-WholeBody, HR-ViP achieves 72.1 WB AP at a resolution of 1024×768. At 384×288, an 86M-parameter DINOv3-B student matches the 66.8 WB AP of a 300M-parameter DINOv3-L model while increasing throughput from 91 to 136 FPS. In the multi-person setting, HR-ViP achieves 68.1 WB AP at 50 FPS with ten detected people, substantially outperforming the full-image methods in whole-body accuracy while retaining real-time throughput. These results demonstrate that HR-ViP achieves a strong accuracy–throughput trade-off for both fine-grained whole-body localization and multi-person inference.

1 INTRODUCTION

Whole-body human pose estimation Jin et al. (2020) localizes 133 keypoints across the body, face, hands, and feet. Their heterogeneous spatial scales place strong demands on representation quality and spatial resolution.

Whole-body pose estimation faces two distinct challenges. The first is **fine-grained landmark localization**. Accurately localizing small and densely distributed face, hand, and foot keypoints requires spatially detailed representations. High-resolution Vision Transformers (ViTs) are well suited to this setting, but their memory and computational costs grow rapidly with the number of image tokens. Reducing the input resolution improves efficiency but removes spatial details required for accurate localization. Consequently, compact CNN-based architectures often provide favorable runtime, whereas large high-resolution ViTs are primarily regarded as accuracy-oriented solutions.

The second challenge is **efficient multi-person inference**. Top-down methods first detect people and then estimate their poses from normalized person crops. This formulation preserves instance boundaries and supports high-resolution localization, but its computational cost grows with the number of detected people. Each additional person requires crop processing and pose inference, while heatmap refinement and coordinate decoding introduce further person-count-dependent latency. In contrast, bottom-up, single-shot, and query-based methods reuse shared image-level features across people and therefore exhibit weaker dependence on person count. However, extending these full-image approaches from sparse body joints to dense whole-body keypoints requires recovering fine spatial details from the shared representation. Processing the full image at sufficiently high resolution

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

can substantially increase computation, whereas using a lower resolution can reduce localization accuracy for small face, hand, and foot keypoints.

Existing approaches therefore tend to favor either fine-grained localization or efficient multi-person processing. High-resolution top-down models achieve strong whole-body accuracy but incur substantial training costs and person-count-dependent inference latency. Shared-feature multi-person methods reduce repeated computation across people but generally operate in a lower whole-body accuracy range. This motivates the following question: can the localization benefits of high-resolution top-down estimation be retained while reducing both its training cost and multi-person inference latency?

To address this question, we propose **HR-ViP**, a top-down framework that efficiently leverages high-resolution pretrained ViTs for whole-body pose estimation. HR-ViP first adapts a high-capacity ViT to high-resolution inputs using Train-Time Random Token Subsampling (TRTS). TRTS processes a random subset of patch tokens during training while using the complete token grid at inference, reducing high-resolution fine-tuning cost without modifying the inference architecture. The resulting spatial representations are then transferred from the large teacher to a smaller student through feature-map distillation, improving inference efficiency without additional inference cost. We further investigate ResizeConv and selective top-k refinement, which provide a range of accuracy–throughput trade-offs for single-person estimation. Building on this efficient student, HR-ViP extends the same framework to multi-person estimation by batching instance-dependent operations throughout the top-down pipeline. Detected person crops are generated through batched sampling and jointly processed by the pose estimator, while heatmap refinement, coordinate decoding, and transformation to the original image coordinates are likewise implemented as batched tensor operations.

Figure 1 summarizes the resulting accuracy–throughput trade-offs. In single-person estimation, HR-ViP achieves 72.1 WB AP at 1024×768. At 384×288, an 86M-parameter DINOv3-B student matches the 66.8 WB AP of a 300M-parameter DINOv3-L teacher while increasing throughput from 91 to 136 FPS. In multi-person estimation, the distilled student reaches 68.14 WB AP at 50 FPS with ten detected people. These results show that high-resolution top-down estimation can achieve favorable accuracy–throughput trade-offs for both fine-grained whole-body localization and multi-person inference.

Our contributions are summarized as follows:

- We introduce HR-ViP, a top-down whole-body pose estimation framework that unifies high-resolution fine-grained localization with efficient single-person and multi-person inference.
- We introduce TRTS for memory- and computation-efficient adaptation of high-resolution pretrained ViTs, and task-specific feature distillation to transfer their spatial representations to compact students without additional inference-time cost.
- We develop efficient decoder variants that provide complementary accuracy–throughput operating points, together with a batched multi-person pipeline that jointly performs crop processing, pose inference, heatmap refinement, and coordinate decoding.

2 Related Work

High-resolution representations for human pose estimation. Human pose estimation has evolved from convolutional architectures, such as Hourglass Newell et al. (2016) and HRNet Sun et al. (2019), to Transformer-based models, including TokenPose Li et al. (2021b) and ViTPose Xu et al. (2022b). Although ViTs provide strong representations for fine-grained localization, their cost grows rapidly with input resolution. Recent methods improve efficiency through sparse processing, token pruning, and token clustering An et al. (2024); Wang et al. (2024); Zeng et al. (2024), while complementary approaches improve keypoint decoding Zhang et al. (2020); Li et al. (2021a); Hu et al. (2024).

Whole-body human pose estimation. Whole-body pose estimation extends body-joint localization to dense body, face, hand, and foot landmarks Jin et al. (2020). ZoomNet and ZoomNAS address the resulting scale variation through hierarchical localization and architecture search Jin et al. (2020); Xu et al. (2022a), while Sapiens explores large-scale ViT pretraining for high-resolution human analysis Khirodkar et al. (2024). Efficient methods have primarily relied on compact CNN-based architectures: RTMPose develops CSPNeXt-based models, whereas DWPose and RTMW improve

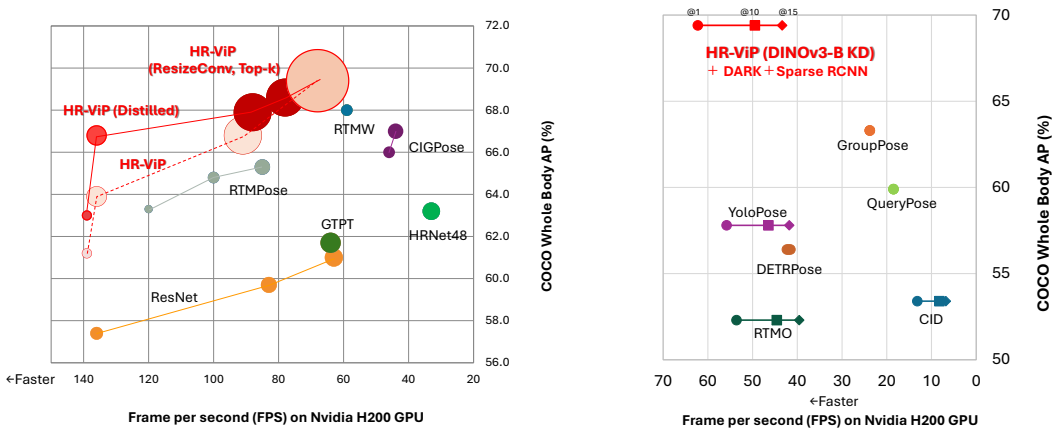

Figure 1: **Accuracy–throughput trade-offs in whole-body pose estimation. (a)** Single-person AP–FPS Pareto frontier on COCO-WholeBody at an input resolution of 384×288 with batch 1. Through feature distillation, efficient decoding, and model scaling, HR-ViP provides a range of accuracy–throughput operating points that outperform previous approaches across the evaluated AP–FPS range. **(b)** Multi-person AP–FPS operating points at fixed person counts. Circles, squares, and diamonds denote FPS@1, FPS@10, and FPS@15, respectively, with markers for the same method forming its person-count trajectory. Full-image methods exhibit weaker dependence on person count, whereas batched HR-ViP maintains substantially higher whole-body accuracy and favorable accuracy–throughput operating points across all three person-count settings.

compact whole-body estimators through knowledge distillation Jiang et al. (2023); Yang et al. (2023b); Jiang et al. (2024). HR-ViP connects these directions by distilling high-resolution ViT features into compact whole-body pose models.

Multi-person human pose estimation. Top-down methods estimate poses from detected person crops and provide strong instance-level localization, but their cost grows with person count He et al. (2017); Xiao et al. (2018). Bottom-up methods predict and associate keypoints globally Cao et al. (2017); Newell et al. (2017); Cheng et al. (2020); Geng et al. (2021), while single-shot and query-based methods predict multiple poses from shared full-image features Maji et al. (2022); Lu et al. (2024); Shi et al. (2022); Xiao et al. (2022); Yang et al. (2023a); Liu et al. (2023a); Janampa & Pattichis (2026). Although these full-image formulations scale favorably with person count, most target body pose rather than dense whole-body landmarks. HR-ViP retains top-down high-resolution localization while reducing person-dependent overhead through batched instance processing.

3 Method

3.1 Task and Overview

Given an image $I \in \mathbb{R}^{H \times W \times 3}$ containing N people, the goal is to estimate $K = 133$ whole-body keypoints for each person. For person n, the coordinate of keypoint k is denoted by $\mathbf{y}_{n,k} \in \mathbb{R}^2$, with visibility indicator $v_{n,k} \in \{0, 1\}$. HR-ViP adopts a top-down formulation in which a person detector first predicts bounding boxes, and a pose estimator then predicts body, face, hand, and foot heatmaps from the corresponding normalized person crops.

As illustrated in Figure 2(a), HR-ViP addresses two complementary challenges: **fine-grained landmark localization** and **efficient multi-person inference**. The former requires spatially detailed representations for small face, hand, and foot keypoints, whereas the latter requires efficient processing of multiple detected people.

For fine-grained landmark localization, HR-ViP leverages a high-resolution pretrained Vision Transformer. Train-Time Random Token Subsampling (TRTS) reduces the number of tokens processed during high-resolution fine-tuning while retaining the complete token grid at inference, as shown in Figure 2(c). Feature-map distillation then transfers the teacher representations to a smaller student, as

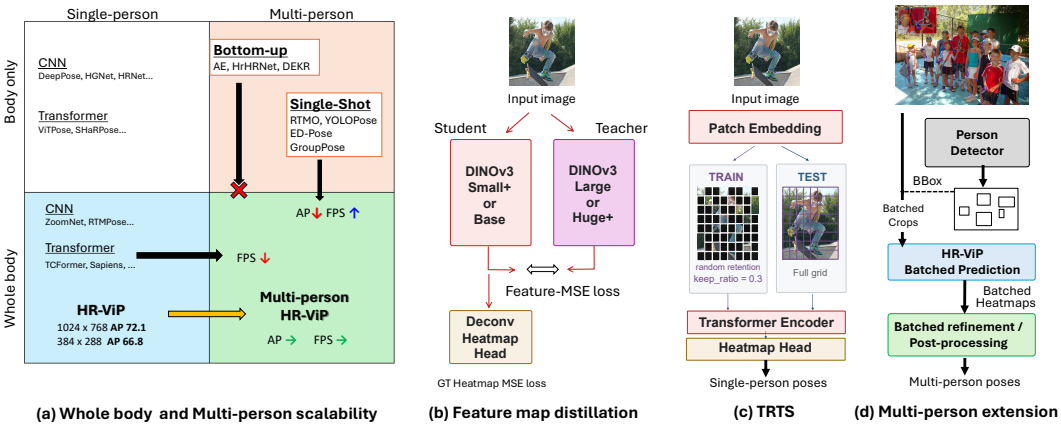

Figure 2: **Overview of HR-ViP. (a)** HR-ViP combines fine-grained whole-body localization with efficient multi-person inference. **(b)** A frozen DINOv3-L/H+ teacher transfers its final spatial feature to a DINOv3-S+/B student. A trainable 1×1 convolution projects the student feature into the teacher channel space before feature-MSE distillation, while the unprojected student feature is passed to the pose head for ground-truth supervision. **(c)** TRTS processes a random subset of patch tokens during training and uses the complete token grid at inference, reducing high-resolution fine-tuning cost without modifying the inference architecture. **(d)** Person cropping, pose inference, heatmap refinement, and coordinate decoding are batched to reduce repeated per-instance execution.

illustrated in Figure 2(b). ResizeConv and selective top-k refinement further provide complementary accuracy–throughput operating points for efficient decoding and fine-grained localization. Together, these components expand the single-person accuracy–throughput frontier.

For efficient multi-person inference, HR-ViP extends the distilled student by batching the instance-dependent operations of the top-down pipeline. Conventional implementations independently perform crop generation, pose inference, heatmap refinement, coordinate decoding, and coordinate transformation for each detected person. HR-ViP instead generates and forwards all person crops as a batch and applies refinement and coordinate decoding jointly across instances, as illustrated in Figure 2(d). This design preserves the fine-grained localization benefits of top-down estimation while improving multi-person throughput through batched instance-dependent processing.

3.2 High-Resolution ViT Baseline for Top-Down Pose Estimation

Accurate localization of face, hand, and foot landmarks requires spatially detailed representations. We therefore use a high-resolution ViT-based top-down model as the foundation of HR-ViP.

High-resolution Vision Transformer backbone. Given a batch of person crops $\mathbf{X} \in \mathbb{R}^{B \times 3 \times H \times W}$, patch embedding produces

$$\mathbf{F}_0 \in \mathbb{R}^{B \times C \times H/P \times W/P}, \tag{1}$$

where P is the patch size and C is the embedding dimension. The resulting patch tokens are processed by D Transformer blocks composed of multi-head self-attention and feed-forward networks. The backbone is initialized from DINOv3 self-supervised pretrained weights Siméoni et al. (2025) and fine-tuned for whole-body pose estimation.

Heatmap head. We use a lightweight deconvolutional head as the base top-down decoder. Given the final spatial feature $\mathbf{F}_{\text{out}} \in \mathbb{R}^{B \times C \times H_d \times W_d}$, the predicted heatmaps are

$$\widehat{\mathbf{H}} = \text{Conv}_{1\times1}\left(\text{Deconv}\left(\text{Deconv}\left(\mathbf{F}_{\text{out}}\right)\right)\right), \tag{2}$$

where $\widehat{\mathbf{H}} \in \mathbb{R}^{B \times K \times H_h \times W_h}$. With two $2\times$ upsampling stages, $H_h = 4H_d$ and $W_h = 4W_d$. We further investigate two efficient decoder variants. ResizeConv fuses coarse semantic features with dense spatial features through interpolation and lightweight convolution, whereas selective top-k refinement applies high-resolution Transformer processing only to informative locations (detailed in Appendix). These variants provide complementary accuracy–throughput operating points.

Heatmap supervision. To reduce coordinate quantization errors for small landmarks, we encode each ground-truth coordinate $\mathbf{y}_{n,k} = (u^*, v^*)$ as a 2D Gaussian heatmap $\mathbf{H}^{\star}_{n,k}$ using unbiased DARK encoding Zhang et al. (2020). The Gaussian is centered at the continuous sub-pixel location rather than at a quantized grid coordinate. The training objective is

$$\mathcal{L}_{\mathrm{GT}} = \frac{1}{\sum_n |V_n|} \sum_n \sum_{k \in V_n} \left\| \hat{\mathbf{H}}_{n,k} - \mathbf{H}^{\star}_{n,k} \right\|_2^2, \tag{3}$$

where $V_n = \{k \mid v_{n,k} = 1\}$ is the set of visible keypoints for person n. At inference, we evaluate both standard heatmap decoding and distribution-aware DARK decoding Zhang et al. (2020), which refines the predicted coordinates to sub-pixel accuracy.

3.3 Efficient High-Resolution Representation Learning

High-resolution ViTs provide strong spatial representations but are expensive to fine-tune and use for inference. HR-ViP addresses these costs in two stages. TRTS enables efficient high-resolution adaptation of a large teacher, and feature-map distillation transfers its representations to a compact student.

Train-Time Random Token Subsampling. Fine-tuning a ViT at high input resolutions, such as 1024×768, is computationally expensive because the number of patch tokens grows with image area and self-attention scales quadratically with the token count. Reducing the input resolution lowers this cost but discards spatial details needed to localize small face, hand, and foot landmarks. Similarly, freezing most of the backbone limits the adaptation of pretrained representations to dense whole-body prediction. We therefore reduce the number of tokens processed only during training, while preserving high-resolution inputs and full-token inference.

Inspired by MAE He et al. (2022) and PatchDropout Liu et al. (2023b), we introduce Train-Time Random Token Subsampling (TRTS). As illustrated in Figure 2(c), let $L = H_p W_p$ be the number of patch tokens and $\rho \in (0, 1]$ the retention ratio. At each training iteration, TRTS samples an index set $\mathcal{I}_\rho \subseteq \{1, \ldots, L\}$ with $|\mathcal{I}_\rho| = \lfloor \rho L \rfloor$ and computes

$$\mathbf{Z}_\rho = \mathrm{Transformer}\left(\mathbf{F}_0[\mathcal{I}_\rho]\right), \qquad \mathcal{O}(L^2) \to \mathcal{O}((\rho L)^2). \tag{4}$$

The corresponding positional entries are selected using the same indices. After the Transformer, the retained tokens are scattered back to their original grid locations, with unselected locations initialized to zero before heatmap decoding. Resampling the token subset across iterations exposes the model to different spatial regions without processing the complete sequence at every training step.

At inference, TRTS is disabled and the complete token grid is processed. TRTS therefore reduces the memory and computation of high-resolution fine-tuning without modifying the inference architecture. RoPE handling and additional implementation details are provided in Appendix.

Task-specific feature-map distillation. TRTS enables efficient adaptation of a high-resolution teacher, but the large teacher remains costly at inference. We therefore transfer the task-adapted spatial representation of the frozen teacher to a compact student, as illustrated in Figure 2(b). The student backbone and pose head are optimized using ground-truth pose supervision, while the backbone additionally receives feature-distillation supervision. Let $\mathbf{F}_\ell^{(t)}$ and $\mathbf{F}_\ell^{(s)}$ denote the final spatial features of the teacher and student, respectively. Because their channel dimensions differ, a lightweight 1×1 projection ψ maps the student feature into the teacher feature space. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{GT}} + \lambda_{\mathrm{feat}} \mathcal{L}_{\mathrm{feat}}, \qquad \mathcal{L}_{\mathrm{feat}} = \frac{1}{B C_t |\Omega|} \left\| \psi\left(\mathbf{F}_\ell^{(s)}\right) - \mathrm{sg}\left(\mathbf{F}_\ell^{(t)}\right) \right\|_2^2, \tag{5}$$

where C_t is the teacher feature dimension, $|\Omega| = H_f W_f$ is the spatial feature grid, and sg denotes stop-gradient. The original, unprojected student feature is independently passed to the pose head and optimized by $\mathcal{L}_{\mathrm{GT}}$. Consequently, the student backbone receives gradients from both objectives, whereas the pose head is trained only with ground-truth pose supervision. The teacher and channel projection are removed after training, so distillation introduces no additional inference-time computation.

3.4 BATCHED MULTI-PERSON EXTENSION

A top-down multi-person pipeline first detects person boxes, filters the detections, constructs normalized crops, and applies the pose estimator. The resulting heatmaps are then refined and decoded, after which the predicted coordinates are transformed back to the original image. Standard top-down implementations are often organized around individual person samples, causing crop construction, pose inference, and post-processing to be repeatedly invoked as the number of people increases. HR-ViP instead batches these instance-dependent operations, as illustrated in Figure 2(d).

Given detected boxes $\mathcal{B} = \{\mathbf{b}_n\}_{n=1}^{\hat{N}}$, we construct all normalized person crops as a tensor and apply the pose estimator in one batched forward pass:

$$\mathbf{X}_{1:\hat{N}} = \mathrm{BatchCrop}(\mathbf{I}, \mathcal{B}), \qquad \hat{\mathbf{H}}_{1:\hat{N}} = f_\theta(\mathbf{X}_{1:\hat{N}}). \tag{6}$$

where $\hat{N}$ denotes the number of retained person detections. Gaussian filtering, DARK refinement Zhang et al. (2020), confidence extraction, and coordinate decoding are likewise applied over the combined instance and keypoint dimensions. The decoded crop-space coordinates are mapped to the original image using the inverse transformation associated with each box:

$$\hat{\mathcal{Y}} = \left\{ \mathcal{T}_{\mathbf{b}_n}^{-1}\left(\hat{\mathbf{y}}_{n,k}^{\mathrm{crop}}\right) \;\middle|\; n = 1, \ldots, \hat{N},\ k = 1, \ldots, K \right\}. \tag{7}$$

This design preserves the fine-grained localization benefits of top-down estimation while reducing repeated per-instance execution through batched crop processing, pose inference, refinement, and decoding. Detailed integration and runtime breakdowns are provided in Appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate HR-ViP on COCO WholeBody v1.0 Jin et al. (2020), reporting whole-body and region-specific AP. Single-person evaluation follows the standard MMPose protocol using fixed detector boxes with 56.4 AP, whereas multi-person evaluation measures the complete image-level pipeline. FPS is measured on an NVIDIA H200 with batch size 1 and mixed precision inference, without ONNX or TensorRT acceleration. Multi-person runtime includes detection, cropping, pose inference, refinement, and coordinate decoding; $\mathrm{FPS@}N$ denotes throughput for images containing exactly N detected people. Models are initialized from DINOv3 Siméoni et al. (2025), trained using only COCO-WholeBody for 40 epochs. Full implementation details are provided in the Appendix.

4.2 SINGLE-PERSON WHOLE-BODY POSE

We compare HR-ViP with representative CNN-based, CSPNeXt-based, and Transformer-based whole-body pose estimators. AP results for prior methods are taken from the original papers or the MMPose benchmark. For a controlled runtime comparison, we measure the corresponding configurations on an NVIDIA H200 GPU. No test-time augmentation is used for HR-ViP. Table 1 reports WB AP, region-specific AP, and throughput. At 384×288, the DINOv3-L baseline achieves 66.8 WB AP at 91 FPS, improving over RTMPose-L by 2.0 AP. The DINOv3-H+ model further improves accuracy to 69.4 WB AP at 68 FPS. Increasing the DINOv3-L input resolution yields 69.9 WB AP at 83 FPS for 640×480 and 70.3 WB AP at 33 FPS for 1024×768. Applying DARK refinement at the latter resolution reaches the highest accuracy of 72.1 WB AP at 25 FPS. Notably, feature-map distillation enables the smaller DINOv3-B student to match the 66.8 WB AP of DINOv3-L while increasing throughput to 136 FPS. This result motivates using the distilled DINOv3-B student as the compact pose estimator in our multi-person pipeline.

4.3 MULTI-PERSON WHOLE-BODY POSE

Table 2 compares HR-ViP with representative bottom-up, single-shot, and query-based methods Wang & Zhang (2022); Maji et al. (2022); Lu et al. (2024); Liu et al. (2023a); Xiao et al. (2022); Janampa & Pattichis (2026). For fair comparison with single shot and query-based methods, all detector-based HR-ViP results in Table 2 use detector confidence for instance ranking, which degrades WB AP

Table 1: **Single-person whole-body pose estimation on COCO-WholeBody v1.0.** We report whole-body AP (WB), region-specific AP, and single-crop throughput. HR-ViP runtime is measured on an NVIDIA H200 with batch size 1, mixed precision, and no ONNX or TensorRT acceleration. Prior runtime values are reported for reference. A dash denotes an unavailable result.

Model	Input	WB	Body	Foot	Face	Hand	FPS
CNN and efficient whole-body models							
ResNet50 Xiao et al. (2018)	384×288	57.4	66.6	63.4	73.1	53.6	<u>136</u>
HRNet-W48 Sun et al. (2019)	384×288	63.2	72.2	69.6	77.6	58.7	33
HRNet-W48+DARK Zhang et al. (2020)	384×288	66.1	74.2	70.7	84.1	60.2	27
ZoomNAS Xu et al. (2022a)	384×288	65.4	74.0	61.7	88.9	62.5	–
SAR Wang & Zhang (2024)	384×288	66.6	71.2	69.3	**90.3**	61.3	–
Efficient CSPNeXt-based models							
RTMPose-M Jiang et al. (2023)	384×288	63.3	70.0	67.6	87.8	56.3	120
RTMPose-L Jiang et al. (2023)	384×288	64.8	71.2	69.3	88.2	57.9	100
RTMPose-X Jiang et al. (2023)	384×288	65.3	71.4	69.2	88.9	59.0	85
DWPose-L Yang et al. (2023b)	384×288	66.5	72.2	70.4	88.7	62.1	100
RTMW-L Jiang et al. (2024)	384×288	68.1	73.8	74.5	89.4	64.6	59
CIGPose-L Li et al. (2026)	384×288	66.3	73.0	72.0	88.3	59.8	46
Transformer-based models							
ViTPose+-H Xu et al. (2022b)	256×192	61.2	75.9	77.9	63.3	54.7	–
GTPT-B Wang et al. (2024)	256×192	61.7	72.0	73.0	84.2	47.9	64[†]
TCFormer Zeng et al. (2024)	384×288	64.2	71.8	74.4	79.0	61.4	–
Sapiens-0.3B Khirodkar et al. (2024)	1024×768	62.0	66.4	67.3	87.1	58.1	–
HR-ViP: capacity scaling and distillation							
HR-ViP-S + KD	384×288	63.0	72.8	74.7	76.0	55.8	**139**
HR-ViP-B	384×288	63.9	72.2	74.3	78.5	58.5	<u>136</u>
HR-ViP-B + KD	384×288	66.8	75.1	78.2	78.5	60.0	<u>136</u>
HR-ViP-L	384×288	66.8	75.9	78.4	78.9	60.9	91
HR-ViP-H+	384×288	69.4	<u>76.8</u>	<u>80.4</u>	82.8	64.9	68
HR-ViP-L: input-resolution scaling							
HR-ViP-L	640×480	69.9	76.7	79.6	86.7	65.3	83
HR-ViP-L	1024×768	<u>70.3</u>	76.4	79.5	87.5	<u>66.5</u>	33
HR-ViP-L + DARK	1024×768	**72.1**	**76.9**	**80.5**	<u>89.7</u>	**68.6**	25

[†]Runtime reported under the original evaluation setting and included for reference.

slightly from Table 1. Ground-truth box evaluation uses pose-aware ranking because all boxes have identical confidence.

The distilled DINOv3-B student with YOLOX-L achieves 66.8 WB AP at 58.0 FPS@1 and 34.8 FPS@10. Applying DARK refinement improves accuracy to 68.5 WB AP while retaining 56.2 FPS@1 and 32.6 FPS@10. Combining the same distilled student with Sparse R-CNN yields 68.1 WB AP and the highest HR-ViP throughput in the table, reaching 63.6 FPS@1 and 49.8 FPS@10. These results show that the compact student provides a strong multi-person operating point across different detector configurations.

4.4 Ablation Studies

Efficient training with TRTS. Figure 4 analyzes TRTS for DINOv3-L at 1024×768 with a batch size of 8. Reducing the retention ratio to $\rho = 0.3$ substantially decreases peak GPU memory and training time while largely preserving WB AP. We separately evaluate scalability with DINOv3-H+. At $\rho = 0.3$, TRTS reduces peak memory from 99 to 42 GB and training time from 38 to 19 hours. Under the same batch-size setting, full-token DINOv3-H+ fine-tuning exceeds the memory capacity of an NVIDIA H100, whereas TRTS makes high-resolution backbone adaptation feasible without adding parameters or modifying the inference-time architecture. We further compare token-selection strategies at the same retention ratio. Deterministic regular-grid sampling achieves 70.1 WB AP,

Table 2: **Multi-person whole-body pose estimation on COCO-WholeBody v1.0.** Det. AP denotes COCO bounding-box AP for the person category. FPS@N denotes end-to-end throughput on images containing exactly N detected people. HR-ViP runtime includes person detection, crop generation, pose inference, refinement, and coordinate decoding. Runtime values for prior full-image methods may use different hardware and are reported for reference. A dash denotes an unavailable or inapplicable value.

Pose estimator	Person detector	Input	Det. AP	WB AP	FPS@1	FPS@10	FPS@15
Full-image bottom-up, single-stage, and query-based methods							
CID	–	512^2	–	53.4	13.2	8.2	6.8
YOLOPose-L	–	640^2	–	57.8	55.8	46.5	41.8
RTMO-L	–	640^2	–	52.3	53.6	44.6	39.6
GroupPose (Swin-L)	–	≤ 1333	–	63.3	23.6	23.6	23.6
QueryPose (HRNet-W48)	–	≤ 1333	–	47.6	14.2	14.2	–
QueryPose (DINOv3-L)	–	1280^2	–	59.9	18.5	18.5	18.5
DETRPose (HGNetv2-L)	–	640^2	–	56.4	42.3	42.3	42.3
HR-ViP with the distilled DINOv3-B student							
HR-ViP-B + KD	YOLOX-L	640^2	60.7	66.8	58.0	34.8	27.0
HR-ViP-B + KD + DARK	YOLOX-L	640^2	60.7	**68.5**	56.2	32.6	25.0
HR-ViP-B + KD + DARK	Sparse R-CNN (DINOv3-L)	640^2	58.4	68.1	**63.6**	**49.8**	**43.4**
HR-ViP-L: detector scaling and oracle-box upper bound							
HR-ViP-L	YOLOX-L	640^2	60.7	66.9	43.6	25.7	–
HR-ViP-L	RTMDet-Nano-Person	320^2	40.3	62.1	46.2	25.3	–
HR-ViP-L	DINO-Swin-L	800×1333	**65.9**	66.7	17.9	10.1	–
HR-ViP-L + DARK	Sparse R-CNN (DINOv3-L)	640^2	58.4	68.4	56.4	35.3	–
HR-ViP-L + DARK	Ground-truth boxes	–	–	71.2	–	–	–



YOLOPose **HR-ViP (ours)**

Figure 3: Qualitative comparison on COCO-WholeBody. The visualizations show predicted whole-body skeletons and fine-grained landmarks, including hands. Best viewed digitally.

whereas the default uniform random sampling achieves 72.1 WB AP. This comparison suggests that iteration-wise token diversity is more effective than repeatedly using a fixed spatial pattern. Detailed region-specific results are provided in the Appendix.

In a separate DINOv3-L experiment, decoder-only training achieves 56.3 WB AP, whereas LoRA fine-tuning reaches 69.9 WB AP. Although both provide lighter optimization, this gap highlights the importance of backbone adaptation. Unlike LoRA, TRTS introduces no additional model parameters and directly reduces token-dependent computation, enabling stronger high-resolution teachers whose representations can subsequently be transferred to compact students through feature-map distillation.

Controlled effect of feature distillation. Under the same student architecture and inference setting, feature distillation improves DINOv3-B from 63.9 to 66.8 WB AP while preserving 136 FPS. The largest region-specific gain is observed for the feet (+3.9 AP), followed by the body (+2.9 AP) and

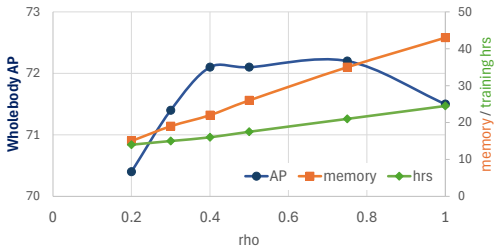

Figure 4: **TRTS efficiency.** WB AP, peak memory, and training time for DINOv3-L at 1024×768 with batch size 8. Full-token training uses $\rho = 1.0$.

Decoder	WB AP	Hand AP	FPS
SimCC	65.8	57.4	**105**
Deconv (biased)	66.1	59.4	91
Deconv (unbiased)	66.8	60.9	91
ResizeConv	67.8	62.8	88
Top-k	**68.1**	**63.9**	78

Table 3: **Decoder ablation.** All variants use DINOv3-L at 384×288 without test-time flip augmentation. Throughput is measured on an NVIDIA H200 with batch size 1. Hand AP averages the left- and right-hand results.

hands ($+1.5$ AP). The distilled 86M-parameter student therefore matches the 66.8 WB AP of the 300M-parameter DINOv3-L baseline while being $1.49\times$ faster.

Distillation objectives. Appendix analyzes the individual distillation objectives using a frozen DINOv3-L teacher and a DINOv3-S student at 384×384 with DARK decoding. Feature-map alignment provides the largest individual improvement, increasing WB AP from 63.1 to 65.0. Among the evaluated objectives, feature MSE provides the strongest and most consistent gain; we therefore use it as the final distillation objective and retain heatmap-level objectives as ablations.

Decoder design. Table 3 compares decoder operating points using DINOv3-L at 384×288. Relative to the unbiased Deconv baseline, SimCC increases throughput from 91 to 105 FPS at a cost of 1.0 WB AP. ResizeConv improves WB AP by 1.0 while retaining 88 FPS. Top-k achieves the best WB AP of 68.1 and Hand AP of 63.9, corresponding to gains of 1.3 and 3.0 AP over the baseline. Thus, SimCC favors speed, ResizeConv provides a balanced operating point, and top-k most effectively improves fine-grained localization.

Batched multi-person processing. The distilled DINOv3-B student with YOLOX-L achieves 66.8 WB AP at 58.0, 34.8, and 27.0 FPS for $N = 1$, 10, and 15, respectively. DARK refinement improves accuracy to 68.5 WB AP while retaining 56.2, 32.6, and 25.0 FPS. Combining the same student with Sparse R-CNN yields 68.1 WB AP and the highest HR-ViP throughput at all three measured person counts, reaching 63.6, 49.8, and 43.4 FPS, respectively. Complete runtime breakdowns and scaling results up to $N = 15$ are provided in Appendix D.4.

Effect of person detector. In Table 2, we further evaluate YOLOX-L Ge et al. (2021), RTMDet-Nano-Person Lyu et al. (2022), DINO-Swin-L Zhang et al. (2023), and Sparse R-CNN Sun et al. (2021) to analyze the effect of detector quality and resolution. For the distilled DINOv3-B student with DARK, Sparse R-CNN achieves 68.1 WB AP and the highest throughput among the evaluated HR-ViP configurations, reaching 63.6 FPS@1 and 49.8 FPS@10. With HR-ViP-L, Sparse R-CNN reaches 68.4 WB AP, while the ground-truth-box upper bound is 71.2 WB AP. These results show that multi-person performance depends on both person detection and pose localization.

5 Conclusion

We presented HR-ViP, a framework that efficiently leverages high-resolution Vision Transformers for whole-body pose estimation. TRTS enables efficient high-resolution backbone adaptation, while feature-map distillation transfers the learned spatial representations to smaller students without additional inference cost. Efficient decoders provide complementary accuracy–throughput operating points for fine-grained landmark localization. Batched crop processing, pose inference, refinement, and coordinate decoding further reduce repeated per-instance execution in multi-person estimation. HR-ViP thereby achieves favorable accuracy–throughput trade-offs for both fine-grained whole-body localization and multi-person inference. Combining top-down localization accuracy with the person-count independence of shared-feature methods remains an important direction for future work.

GENERATIVE AI DISCLOSURE

Generative AI tools were used for manuscript drafting and editing, translation, figure and table refinement, literature search, reference formatting, experimental feedback, result interpretation, and assistance with software implementation and debugging. All AI-assisted text, references, code, figures, and interpretations were reviewed and verified by the authors. No datasets, experimental measurements, or reported numerical results were generated by generative AI. The authors take full responsibility for the methodology, experiments, claims, and final content of this paper.

REFERENCES

Xiaoqi An, Lin Zhao, Chen Gong, Nannan Wang, Di Wang, and Jian Yang. Sharpose: Sparse high-resolution representation for human pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2), 2024. doi: 10.1609/aaai.v38i2.27826.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5386–5395, 2020.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Shengxiang Hu, Huaijiang Sun, Dong Wei, Xiaoning Sun, and Jin Wang. Continuous heatmap regression for pose estimation via implicit neural representation. In *NeurIPS*, 2024.

Sebastian Janampa and Marios Pattichis. Detrpose: Real-time end-to-end multi-person pose estimation via modified transformer decoder and novel denoising keypoints. In *European Conference on Computer Vision*, 2026.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. URL `https://arxiv.org/abs/2303.07399`.

Tao Jiang, Xinchen Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024.

Bohao Li, Zhicheng Cao, Huixian Li, and Yangming Guo. Cigpose: Causal intervention graph neural network for whole-body pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23484–23494, June 2026.

Jiefeng Li, Siyuan Bian, Ailing Zeng, et al. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021a.

Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 11313–11322, 2021b.

Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, and Jingdong Wang. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15029–15038, 2023a.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. Patchdropout: Economizing vision transformers using patch dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3953–3962, 2023b.

Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1491–1500, 2024. doi: 10.1109/CVPR52733.2024.00148.

Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *CVPRW*, pp. 2637–2646, 2022.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pp. 483–499. Springer, 2016.

Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017.

Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11069–11078, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14454–14463, 2021.

Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11060–11068, 2022.

Dongkai Wang and Shiliang Zhang. Spatial-aware regression for keypoint localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 624–633, June 2024.

Haonan Wang, Jie Liu, Jie Tang, Gangshan Wu, Bo Xu, Yanbing Chou, and Yong Wang. Gtpt: Group-based token pruning transformer for efficient human pose estimation. *ECCV*, 2024.

Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pp. 466–481, 2018.

Yabo Xiao, Kai Su, Xiaojuan Wang, Dongdong Yu, Lei Jin, Mingshu He, and Zehuan Yuan. Querypose: Sparse multi-person pose regression via spatial-aware part-level query. In *NeurIPS*, 2022.

Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: Searching for whole-body human pose estimation in the wild. *TPAMI*, 2022a.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022b.

Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *ICLR*, 2023a.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 4210–4220, 2023b.

Wang Zeng, Sheng Jin, Lumin Xu, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Tcformer: Visual recognition via token clustering transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

A Methodological Comparison with Related Approaches

Table 4: **Methodological comparison with representative pose-estimation approaches.** Whole-body denotes explicit prediction of body, foot, face, and hand keypoints. Multi-person indicates whether the method directly supports inference for multiple people. Full-image methods predict multiple poses from shared image features, whereas top-down methods estimate poses from detected person regions. The table summarizes methodological characteristics rather than directly comparable AP or runtime values.

Method	Formulation	Whole-body	Scope	Representation	Knowledge transfer	Efficiency target
ViTPose Xu et al. (2022b)	Top-down	Partial	Single person	Plain non-hierarchical ViT with a lightweight keypoint decoder	Knowledge transfer from larger to smaller ViTs with heatmap and token KD	Model-capacity scaling and architectural simplicity
GTPT Wang et al. (2024)	Top-down	Yes	Single person	Grouped keypoint tokens, visual-token pruning, and group attention	Not central	Reduced token computation during pose estimation
DWPose Yang et al. (2023b)	Top-down	Yes	Single person	Efficient RTMPose-based whole-body estimator	Two-stage feature, logit, and prediction-head distillation	Compact whole-body inference without test-time overhead
RTMW Jiang et al. (2024)	Top-down	Yes	Single person	Feature pyramid and hierarchical encoding for heterogeneous landmarks	Two-stage distillation with aligned human-keypoint datasets	Real-time 2D and 3D whole-body estimation
Sapiens Khirodkar et al. (2024)	Top-down	Yes	Single person	Large high-resolution human foundation models	Large-scale human-centric self-supervised pretraining	Representation quality and high-fidelity human analysis
QueryPose Xiao et al. (2022)	Full-image query-based	No	Multi person	Instance- and spatially-aware part queries over shared image features	Not central	Sparse end-to-end multi-person prediction
ED-Pose Yang et al. (2023a)	Full-image query-based	No	Multi person	Explicit box detection and keypoint queries in a unified Transformer	Not central	End-to-end person detection and pose estimation

Continued on the next page

Table 4 continued.

Method	Formulation	Whole-body	Scope	Representation	Knowledge transfer	Efficiency target
GroupPose Liu et al. (2023a)	Full-image query-based	No	Multi person	Group-based queries for person instances and body keypoints	Not central	Shared-feature multi-person pose estimation
DETRPose Janampa & Pattichis (2026)	Full-image query-based	No	Multi person	DETR-style person and keypoint prediction from shared image features	Not central	Real-time end-to-end multi-person estimation
HR-ViP	**Batched top-down**	**Yes**	**Both**	High-resolution pretrained ViT with ResizeConv and selective top-k decoding	TRTS-based teacher adaptation and task-specific feature-map distillation	Training memory, compact inference, dense localization, and multi-person latency

Table 4 provides a methodological comparison between HR-ViP and representative approaches related to Vision Transformer scaling, high-resolution processing, knowledge distillation, human-centric pretraining, and multi-person pose estimation. Since these methods differ in their keypoint definitions, training data, evaluation protocols, and runtime environments, we focus on their design characteristics rather than directly comparing reported AP or throughput. Existing methods separately study ViT scaling, inference-time token reduction, whole-body distillation, human-centric pretraining, or shared-feature multi-person estimation. HR-ViP connects these directions by combining training-only token subsampling, task-specific feature distillation, efficient whole-body decoding, and batched top-down inference. It therefore supports both high-accuracy single-person estimation and efficient multi-person processing without modifying the deployed student backbone.

B Additional Method Details

B.1 Network architecture in detail

Figure 5 provides detailed architectures of the baseline HR-ViP model, Train-Time Random Token Subsampling (TRTS), ResizeConv, and selective top-k refinement. These variants share the same pretrained Vision Transformer representation but differ in how they reduce training cost or recover fine-grained spatial information.

Base HR-ViP. The baseline architecture receives a normalized person crop and partitions it into non-overlapping 16×16 patches. For a 384×384 input, patch embedding produces a 24×24 grid containing 576 patch tokens. The tokens are processed by 24 pretrained Vision Transformer blocks, each consisting of layer normalization, multi-head self-attention, a residual connection, and a feed-forward network. The output token sequence is reshaped into a spatial feature map and passed to a lightweight deconvolutional heatmap head. Two successive upsampling stages increase the spatial resolution by a factor of four. A final 1×1 predictor produces 133 heatmaps at 96×96 resolution, corresponding to the body, feet, face, and both hands.

Train-Time Random Token Subsampling. TRTS reduces the computational cost of high-resolution fine-tuning without changing the deployed model. Let $L = H_p W_p$ denote the number of tokens in the complete patch grid. During training, a random subset of $P = \rho L$ patch tokens is retained, where ρ is the token-retention ratio. The selected tokens are processed by the same Transformer encoder used by the baseline model. DINOv3 uses rotary positional embeddings, so the positional entries corresponding to the retained tokens are gathered using the same token indices. This preserves the correspondence between image patches and positional information after subsampling. After the Transformer encoder, the processed tokens are scattered back to their original spatial locations before heatmap prediction. At inference, TRTS is disabled and the complete token grid is processed. Thus, TRTS reduces training-time attention cost without introducing token pruning or structural changes at inference.

ResizeConv decoder. ResizeConv uses separate coarse and dense feature branches. The coarse branch processes a lower-resolution input with the pretrained Transformer and captures semantic information. The dense branch preserves higher-resolution spatial features using patch embedding with substantially lower computational cost than applying the complete Transformer to the dense

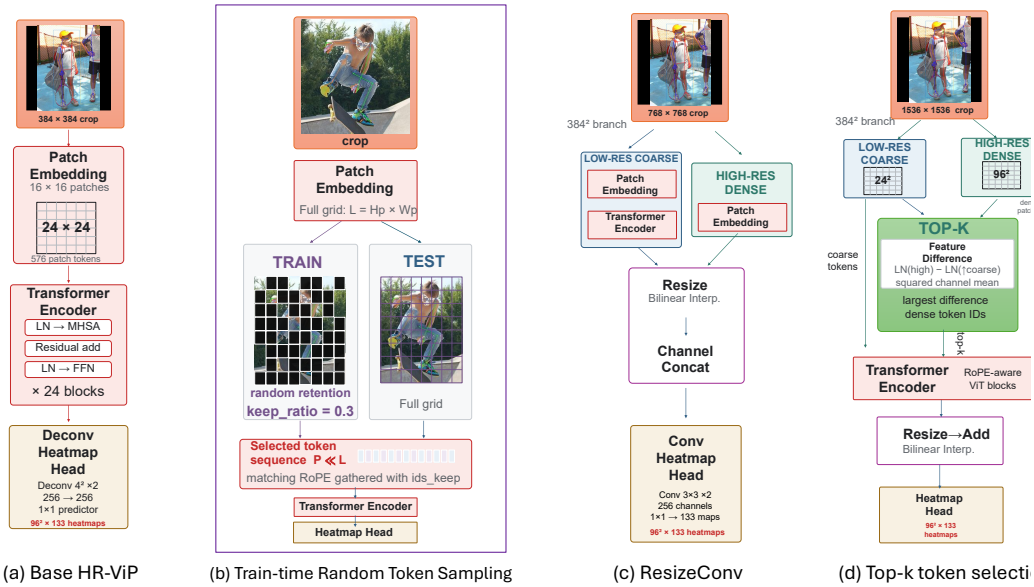

Figure 5: **Detailed architectures of HR-ViP variants. (a)** The base HR-ViP model divides a 384×384 person crop into 16×16 patches, processes the resulting 24×24 token grid using a pretrained Vision Transformer, and predicts 133 whole-body heatmaps with a lightweight deconvolutional head. **(b)** TRTS randomly retains a subset of patch tokens during training and gathers the corresponding rotary positional embeddings using the same retained indices. The complete token grid is restored at inference. **(c)** ResizeConv combines low-resolution semantic features with a dense spatial branch through bilinear resizing, channel concatenation, and lightweight convolutional heatmap decoding. **(d)** Selective top-k refinement ranks dense locations using the normalized discrepancy between high-resolution and upsampled coarse features. Only the selected dense tokens and coarse tokens are processed by the Transformer, after which the refined features are resized and merged for heatmap prediction.

token grid. The coarse features are bilinearly resized to match the spatial resolution of the dense features. The two feature maps are concatenated along the channel dimension and decoded by a lightweight convolutional head. The decoder uses two 3×3 convolutional layers with 256 channels followed by a 1×1 predictor that outputs 133 heatmaps. This design recovers spatial detail through inexpensive resizing and convolutional fusion rather than dense high-resolution self-attention.

Selective top-k refinement. The selective top-k variant processes a low-resolution coarse branch together with a high-resolution dense branch. Let $\mathbf{F}^c$ and $\mathbf{F}^d$ denote the coarse and dense feature maps, respectively. The coarse feature map is first resized to the dense spatial resolution. For each dense location i, an importance score is computed as

$$s_i = \mathrm{mean}_c \left[\mathrm{LN}\left(\mathbf{F}_i^d\right) - \mathrm{LN}(\uparrow \mathbf{F}_i^c)\right]^2, \tag{8}$$

where $\uparrow$ denotes bilinear interpolation and the mean is taken over feature channels. The k locations with the largest discrepancy scores are selected:

$$\mathcal{I}_{\text{top-}k} = \mathrm{TopK}\left(\{s_i\}_{i=1}^{H_d W_d}, k\right). \tag{9}$$

These locations represent regions where the upsampled coarse representation differs most strongly from the dense representation and therefore benefits most from additional high-resolution processing. The selected dense tokens are combined with the complete coarse-token sequence and processed by RoPE-aware Vision Transformer blocks. The refined coarse representation is then resized to the dense grid, and the selected high-resolution features are injected at their original spatial locations. A heatmap head predicts the final 133 whole-body heatmaps. By restricting high-resolution Transformer computation to the selected locations, top-k refinement improves fine-grained localization without processing the complete dense token grid.

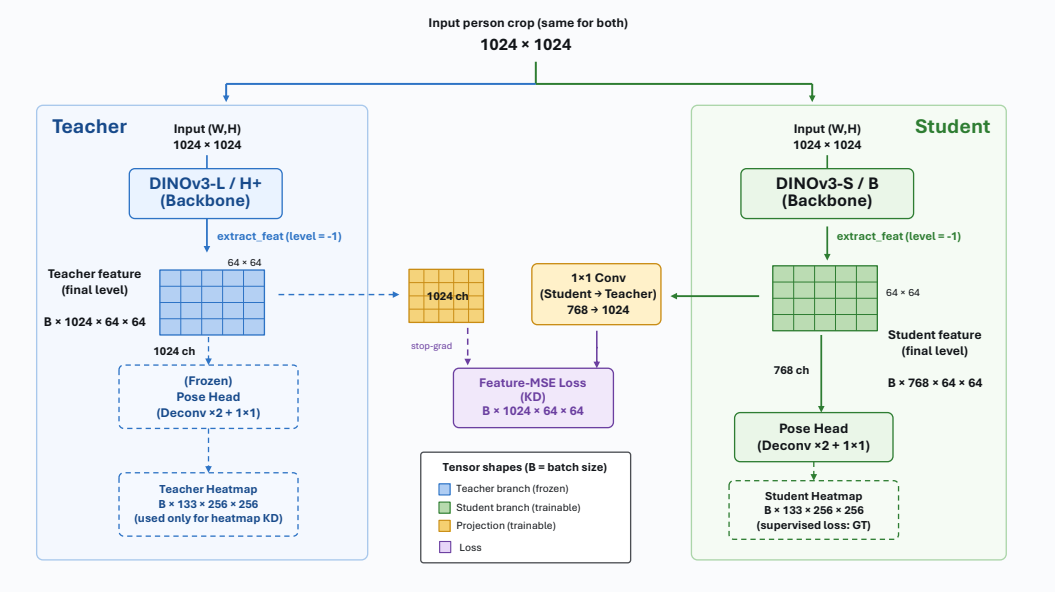

Figure 6: **Task-specific feature-map distillation in HR-ViP.** The frozen DINOv3-L teacher and trainable DINOv3-B student process the same 1024×1024 person crop. Their final backbone features are extracted before the pose head using `extract_feat` with `level=-1`. Both branches produce a 64×64 spatial feature grid. A trainable 1×1 convolution projects the 768-channel student feature into the 1024-channel teacher space, where feature MSE is computed against the stop-gradient teacher feature. Independently, the original student feature is passed to the pose head, which is optimized using ground-truth supervision for 133 whole-body keypoints. At inference, the teacher and channel projection are removed, leaving only the student backbone and pose head. The optional teacher pose head and teacher heatmaps are used only for the heatmap-distillation ablations, whereas the final configuration uses feature MSE alone.

Comparison of architectural roles. The four configurations address different computational bottlenecks. The base model provides a simple high-resolution ViT baseline. TRTS reduces the memory and training-time cost of adapting this baseline while preserving full-token inference. ResizeConv improves spatial recovery using lightweight convolutional fusion. Selective top-k refinement allocates Transformer computation to locations where coarse features provide insufficient spatial detail. Consequently, TRTS primarily improves training efficiency, whereas ResizeConv and top-k provide distinct accuracy and throughput operating points at inference.

B.2 RoPE-Aware Train-Time Random Token Subsampling

For DINOv3 backbones with rotary positional embeddings (RoPE), the retained patch-token indices are also used to gather the corresponding RoPE entries before attention computation. Prefix tokens are excluded from RoPE indexing. The same indices are applied to patch tokens and positional embeddings, preserving their spatial correspondence after subsampling.

After the Transformer encoder, retained tokens are scattered back to their original grid locations, and unselected positions are initialized to zero during training. At inference, HR-ViP processes the complete token grid and its corresponding RoPE entries. The implementation supports both shared and block-specific RoPE representations.

B.3 Task-Specific Feature-Map Distillation

Figure 6 illustrates the feature-map distillation used in HR-ViP. The frozen DINOv3-L teacher and trainable DINOv3-B student process the same 1024×1024 person crop. Distillation is applied to the final spatial feature returned by `extract_feat` with `level=-1`:

$$\mathbf{F}^{(t)} \in \mathbb{R}^{B\times1024\times64\times64}, \qquad \mathbf{F}^{(s)} \in \mathbb{R}^{B\times768\times64\times64}. \tag{10}$$

The spatial grids already match, so no interpolation is required. A trainable 1×1 convolution projects the student feature into the teacher channel space:

$$\widetilde{\mathbf{F}}^{(s)} = \psi\Big(\mathbf{F}^{(s)}\Big), \qquad \psi : \mathbb{R}^{768} \to \mathbb{R}^{1024}. \tag{11}$$

The feature-distillation loss is

$$\mathcal{L}_{\text{feat}} = \frac{1}{BC_tH_fW_f} \left\| \widetilde{\mathbf{F}}^{(s)} - \text{sg}\Big(\mathbf{F}^{(t)}\Big) \right\|_2^2, \tag{12}$$

where $C_t = 1024$, $H_f = W_f = 64$, and sg denotes stop-gradient. The original student feature is independently passed to the pose head and optimized using ground-truth heatmap supervision. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{GT}} + \lambda_{\text{feat}}(e)\mathcal{L}_{\text{feat}}, \qquad \lambda_{\text{feat}}(e) = 0.2 \min\left(1, \frac{e}{5}\right). \tag{13}$$

where e denotes the training epoch. Thus, the student backbone receives gradients from both objectives, while the pose head and channel projection are updated by $\mathcal{L}_{\text{GT}}$ and $\mathcal{L}_{\text{feat}}$, respectively. Heatmap MSE and spatial KL distillation are evaluated only in the ablation in Table 7; the final configuration uses feature MSE alone. At inference, only the student backbone and pose head are retained.

C Implementation Details

C.1 Single-Person Implementation

Single-person models are implemented using `MMPose` and validated every five epochs. We use AdamW with an initial learning rate of 10^{-4}, $\beta = (0.9, 0.999)$, weight decay 0.01, and layer-wise learning-rate decay of 0.8 over D Transformer layers. Mixed-precision training uses `bfloat16` through `AmpOptimWrapper`.

The learning rate is linearly warmed up for 500 iterations and decayed by a factor of 0.1 at epochs 25 and 35. Training augmentation includes horizontal flipping, half-body augmentation, bounding-box perturbation, and affine resizing. Ground-truth heatmaps use unbiased encoding. Standard or DARK decoding is applied as specified for each configuration. The Gaussian standard deviation is set to $\sigma = 2$, 4, and 5 for input resolutions of 384, 640, and 1024, respectively.

Single-person evaluation follows the standard MMPose protocol using fixed person-detection boxes generated by a detector with 56.4 AP. Person-level scores combine bounding-box and keypoint confidence scores, using a keypoint threshold of 0.2. Duplicate predictions are removed using OKS-NMS with a threshold of 0.9.

C.2 Multi-Person Implementation and Integration

Processing stages. Given an input image $\mathbf{I}$, the complete multi-person pipeline consists of the following stages:

1. **Person detection.** The detector predicts person bounding boxes, confidence scores, and category labels from the complete image.
2. **Detection post-processing.** Predictions are filtered using the person category and confidence threshold. When enabled, additional non-maximum suppression removes duplicate boxes.
3. **Crop construction.** Each predicted box is enlarged using a padding factor of 1.25 and converted into the center–scale representation used by top-down pose estimation.
4. **Batched affine sampling.** All person crops from the same image are transformed into normalized 384×384 pose inputs. The batched implementation constructs the affine transformations on the GPU and samples all crops using a single tensor operation.
5. **Batched pose inference.** The normalized crops are stacked into a tensor $\mathbf{X} \in \mathbb{R}^{N\times3\times384\times384}$ and processed by the distilled DINOv3-B HR-ViP estimator in one batched forward pass.

6. **Heatmap refinement and coordinate decoding.** The predicted heatmaps $\hat{\mathbf{H}} \in \mathbb{R}^{N\times 133\times h\times w}$ are decoded using standard or DARK refinement, where h and w denote the heatmap height and width. Gaussian modulation, local coordinate refinement, and confidence extraction are implemented over the combined person and keypoint dimensions.
7. **Coordinate restoration.** The decoded coordinates are transformed from the normalized crop system back to the original image coordinate system using the inverse affine transformations.

The integrated runtime includes all stages from detector inference through pose decoding and coordinate restoration. Image-file reading, dataset loading, and visualization are excluded.

HR-ViP is integrated into top-down multi-person pipelines implemented with `MMPose`, `MMDetection`, `Detectron2`, and `Detrex`. We evaluate YOLOX-L, RTMDet-Nano-Person, DINO-Swin-L, and Sparse R-CNN detectors. Unless otherwise specified, the pose estimator is the distilled DINOv3-B HR-ViP model and predicts 133 COCO-WholeBody keypoints from 384×384 normalized person crops. Given an input image, the detector first predicts person boxes and confidence scores. The boxes are filtered, enlarged by a factor of 1.25, and converted into affine transformations. All person crops are sampled on the GPU and stacked into

$$\mathbf{X} \in \mathbb{R}^{N\times3\times384\times384}, \tag{14}$$

which is processed in one batched pose-estimation pass. Heatmap refinement, confidence extraction, coordinate decoding, and coordinate restoration are also applied jointly over the person and keypoint dimensions. When enabled, DARK decoding uses Gaussian modulation with a kernel size of 11. Predictions are transformed back to the original image coordinates and duplicate poses are removed using OKS-NMS with a threshold of 0.9. For the multi-person comparison in Table 2, detector confidence is used for instance ranking. Detector input resolution, detector AP, WB AP, and end-to-end throughput are reported for each configuration in that table.

MMPose integration. The sequential MMPose pipeline constructs and processes a separate pose sample for each detection. Our implementation instead represents the detected instances as a batch throughout preprocessing, pose inference, and decoding. Affine transformations are constructed as tensors, and all normalized person crops are sampled in one operation before being passed to the pose estimator as a tensor of shape $\mathbf{X} \in \mathbb{R}^{N\times3\times H\times W}$.

The predicted heatmaps $\widehat{\mathbf{H}} \in \mathbb{R}^{N\times K\times H_h\times W_h}$ are also decoded jointly on the GPU. Peak coordinates and confidence scores are obtained by flattening the spatial dimensions and applying batched `argmax` operations over all person–keypoint pairs. Standard quarter-pixel refinement or DARK refinement is then applied using vectorized PyTorch operations. In particular, the batched DARK implementation performs Gaussian modulation, derivative estimation, and sub-pixel coordinate updates without iterating over individual people or keypoints. The resulting coordinates are scaled from the runtime heatmap resolution to the pose-input resolution using a dynamically computed scale factor. Keeping heatmap decoding and refinement as batched tensor operations avoids per-instance NumPy decoding and repeated Python-level dispatch, reducing the overhead of crop construction, pose inference, and post-processing.

Sparse R-CNN integration. We train a single-class Sparse R-CNN person detector on the COCO WholeBody training set for six epochs using `Detectron2` and `Detrex`. The detector uses a pretrained DINOv3-L backbone with a Simple Feature Pyramid. The detector backbone is evaluated once per image, and the box branch applies ROIAlign to the detector features for person-box prediction. For pose estimation, the retained person boxes are jointly processed by a separate `ROIAlignV2` operator applied directly to the input image. All person regions are pooled into a batched tensor of 384×384 RGB crops and passed to the distilled DINOv3-B HR-ViP estimator in one forward pass.

The person boxes, ROI-aligned image crops, predicted heatmaps, and decoded keypoints remain grouped as batched instance data throughout the pipeline. Heatmap peak extraction, standard or DARK refinement, confidence extraction, and coordinate restoration are likewise performed jointly across person instances and keypoints. The measured throughput benefits from a single full-image DINOv3-L backbone evaluation, batched `ROIAlignV2` crop extraction, and batched pose decoding and post-processing. The pose computation remains person-dependent because one ROI-aligned RGB crop is processed for each retained detection.

Table 5: **Controlled scaling of backbone capacity and input resolution.** We report WB AP and single-crop throughput with standard and DARK decoding. KD does not change the student architecture or inference cost.

			Standard		+DARK	
Backbone	Params	Input	WB AP	FPS	WB AP	FPS
DINOv3-S+	26M	384×288	60.6	139	63.1	77
DINOv3-S+ + KD	26M	384×288	63.0	139	65.0	77
DINOv3-B	86M	384×288	63.9	136	65.9	72
DINOv3-B + KD	86M	384×288	66.8	136	67.9	72
DINOv3-L	300M	384×288	66.8	91	69.0	57
DINOv3-H+	840M	384×288	69.4	68	71.2	46
DINOv3-L	300M	256×192	61.2	100	64.7	67
DINOv3-L	300M	640×480	69.9	83	71.1	45
DINOv3-L	300M	1024×768	70.3	33	72.1	25

Table 6: **Controlled analyses of pretraining and TRTS. (a)** Effect of the pretrained representation using ViT-L at 1024×768. **(b)** TRTS across DINOv3 backbone capacities at 1024×768; OOM denotes full-token training that exceeds the evaluated H100 memory limit. **(c)** Token-sampling strategies at the same retention ratio using DINOv3-L and full-token inference with DARK decoding.

(a) Pretraining

Pretraining	WB AP
DINOv2	70.8
MAE	70.9
DINOv3	**71.5**

(b) Backbone capacity

Backbone	Full	TRTS	Δ
DINOv3-B	69.6	69.1	−0.5
DINOv3-L	71.5	71.4	−0.1
DINOv3-H+	OOM	**72.4**	–

(c) Token sampling

Sampling	WB	Body	Foot	Face	Hand
Regular grid	70.1	75.7	78.9	87.9	66.1
Uniform random	**72.1**	**76.9**	**80.5**	**89.7**	**68.6**

D Additional Experimental Analyses

D.1 Model Capacity and Input Resolution

Table 5 analyzes backbone capacity at 384×288 and input-resolution scaling with DINOv3-L. Increasing model capacity and resolution improves accuracy at the cost of throughput. Feature distillation improves compact students without changing their inference cost, while resolution scaling provides its most favorable gains before the highest-resolution setting.

D.2 Pretraining and TRTS Analysis

Table 6 summarizes the controlled pretraining and TRTS analyses. DINOv3 provides the strongest pretrained representation. TRTS largely preserves accuracy across model scales while enabling high-resolution DINOv3-H+ training when full-token fine-tuning exceeds the evaluated memory limit. Uniform random sampling also outperforms a fixed regular grid, supporting iteration-wise token diversity. Because these panels use their respective controlled protocols, the values should not be compared directly with Table 1.

D.3 Knowledge Distillation Analysis

Table 7 summarizes the distillation analyses. Feature MSE provides the primary improvement, whereas adding heatmap MSE or spatial KL provides no further gain. Distillation consistently improves the DINOv3-B student across resolutions, while scaling the teacher beyond DINOv3-L provides only small, resolution-dependent differences. We therefore use DINOv3-L and feature MSE as the default configuration.

Table 7: **Knowledge-distillation analyses.** **(a)** Ablation of distillation objectives using a frozen DINOv3-L teacher, a DINOv3-S student, 384×384 training inputs, and DARK decoding. **(b)** Effect of teacher capacity and student input resolution using a DINOv3-B student and standard decoding. Parentheses indicate gains over the non-distilled student.

(a) Distillation objectives

HM MSE	HM KL	Feature	WB AP
			63.1
✓			63.8
	✓		63.1
		✓	**65.0**
✓		✓	**65.0**
	✓	✓	63.7
✓	✓	✓	**65.0**

(b) Teacher capacity and resolution

Input	No KD	DINOv3-L	DINOv3-H+
384×288	63.90	65.95 (+2.05)	**66.85** (+2.95)
640×480	66.76	**69.30** (+2.54)	69.00 (+2.24)
1024×768	67.98	70.50 (+2.52)	**70.60** (+2.62)

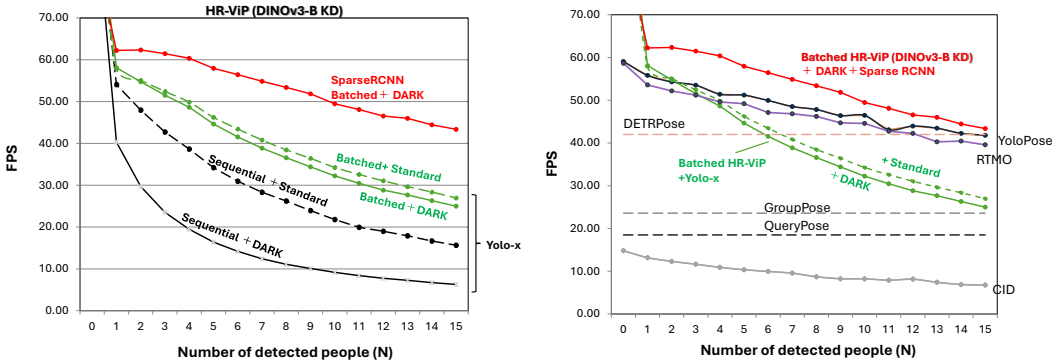

Figure 7: **Runtime scaling with the number of detected people. Left:** Sequential processing exhibits substantial person-count dependent slowdown, particularly with DARK refinement. Batching crop sampling, pose inference, refinement, and decoding substantially improves scaling and reduces the throughput difference between standard and DARK decoding. **Right:** Comparison with full-image multi-person methods and integrated HR-ViP pipelines. The Sparse R-CNN integration batches the remaining instance-dependent operations and maintains over 40 FPS at $N = 15$, while retaining substantially higher whole-body accuracy than the evaluated full-image methods.

D.4 Runtime Scaling with the Number of People

We analyze end-to-end throughput as the number of detected people increases. Runtime includes person detection, crop sampling, pose inference, heatmap refinement, coordinate decoding, and coordinate restoration. Image loading and visualization are excluded. Table 8 isolates the effect of batching under the same YOLOX-L setting. At $N = 10$, batching improves standard-decoding throughput from 21.80 to 34.26 FPS and DARK throughput from 9.17 to 32.25 FPS. The larger improvement with DARK shows that sequential heatmap refinement and decoding introduce substantial per-person overhead. Once these operations are batched, standard and DARK decoding exhibit similar throughput. The left panel of Figure 7 extends the controlled comparison across person counts. Sequential DARK processing declines most rapidly because refinement and decoding are repeatedly invoked for each person. Batching these operations largely removes this additional overhead: the standard and DARK curves become close, while retaining the crop-normalized localization of top-down estimation. The right panel compares HR-ViP with full-image multi-person methods. The integrated Sparse R-CNN pipeline batches crop generation, pose inference, refinement, decoding, and coordinate restoration, maintaining over 40 FPS at $N = 15$. Full-image methods show weaker dependence on person count, whereas HR-ViP provides a higher whole-body accuracy range in the evaluated comparison.

Table 8: **Effect of batching at fixed person counts.** Pose latency includes pose inference, refinement, decoding, and coordinate restoration. Total latency includes person detection and the pose pipeline.

Processing	Post-process	N	Count	Detector (ms)	Pose (ms)	Total (ms)	FPS
Sequential	Standard	1	913	9.31	9.17	18.48	54.07
Batched	Standard	1	913	9.32	7.96	17.28	57.88
Sequential	DARK	1	913	9.31	13.67	22.98	43.52
Batched	DARK	1	913	9.30	7.93	17.23	**58.08**
Sequential	Standard	10	61	9.29	36.78	46.07	21.80
Batched	Standard	10	61	9.27	19.92	29.19	**34.26**
Sequential	DARK	10	61	9.31	99.86	109.17	9.17
Batched	DARK	10	61	9.26	21.75	31.01	**32.25**

Table 9: **High-throughput DINOv3-B operating point.** Throughput is measured on an NVIDIA H200 with batch size 1 and mixed-precision inference. Hand AP is averaged over the left and right hands.

Decoder	WB AP	Body AP	Foot AP	Face AP	Hand AP	FPS
SimCC	62.5	72.3	73.3	76.8	53.4	**153**

D.5 High-Throughput DINOv3-B with SimCC

We evaluate SimCC decoding with the compact DINOv3-B backbone as a speed-oriented HR-ViP operating point. As shown in Table 9, the model achieves 62.5 WB AP at 153 FPS with a 384×288 input on an NVIDIA H200. Note that this DINOv3-B configuration differs from the DINOv3-L SimCC result in Table 3 and provides a more speed-oriented operating point.

SimCC improves throughput by reducing spatial decoding cost, but provides lower accuracy for fine-grained landmarks, particularly the hands. It therefore complements the heatmap-based variants, which provide stronger fine-grained localization, as a speed-oriented endpoint in the HR-ViP accuracy–throughput trade-off.

D.6 Extended Qualitative Results

Figure 8 compares representative full-image methods with HR-ViP under variations in person count, scale, overlap, and visibility. The examples illustrate successful fine-grained localization and failure cases caused by missed detections, inaccurate boxes, and ambiguous small landmarks.

E Limitations

HR-ViP remains dependent on person detection. Missed detections, duplicate detections, and inaccurate boxes cannot be fully recovered by the top-down pose estimator and disproportionately affect small landmarks under occlusion and inter-person overlap. Although batched processing substantially reduces refinement and decoding overhead, HR-ViP still processes one crop per detected person and therefore slows down in highly crowded scenes.

Full-image methods scale more favorably with person count but do not yet match the whole-body accuracy of HR-ViP in our evaluation. Combining the fine-grained localization accuracy of top-down estimation with the crowd-level scalability of full-image approaches remains an open problem. Future work may explore shared full-image representations, crowd-adaptive processing, and hybrid top-down and single-stage formulations.



RTMO **YOLOPose** **HR-ViP (ours)**

Figure 8: **Qualitative comparison on COCO-WholeBody.** Columns show RTMO, YOLOPose, and HR-ViP. Rows cover different person counts, scales, overlap patterns, and visibility conditions.



Figure 9: Additional qualitative results on challenging in-the-wild images. The examples demonstrate robust whole-body pose estimation across diverse scenarios, including single- and multi-person scenes, large scale variations, unusual viewpoints, dynamic actions, human–object interactions, partial occlusion, and densely crowded environments. The visualizations show the predicted person bounding boxes and whole-body keypoints, including body, face, hand, and foot landmarks.