We sincerely thank all reviewers for their comments and for acknowledging that the *clear and strong motivation, straightforward yet powerful key insight, effectiveness on different baselines* (R1), *significantly reduced computational cost while achieving competitive performance* (R3), *the novel Token Reduction Supervision* (R2), *extensive and comprehensive evaluations* (R1, R3). Next, we address the main concerns below:

[R1] How GTR or ITR affects the MVE? For MPVE, using a ResNet-50 backbone, GTR typically causes a larger accuracy drop of 2.6 mm (from 86.7 to 84.1), while ITP is 1.5 mm (from 88.2 to 86.7), . With a lighter EfficientNet-b0 (Eb0) backbone, GTR's drop is 0.8 mm, while ITP improves accuracy by 0.8 mm. This suggests GTR sacrifices accuracy for efficiency, while ITP has less impact and can even improve accuracy on lightweight backbones. We'll include more detailed discussions and visualizations in the revision.

[R1]Novelty of GTR. Existing transformer-based methods treat mesh vertices and skeleton joints as tokens for feature interaction holistically. However, a few skeleton joints can effectively abstract numerous vertices. Based on this body hierarchy, we propose NSR for GTR to hierarchically recover body shape, effectively reducing redundancy. To put it differently, we apply attention matrix decomposition for efficient shape representation in HMR.

[R1]Regressing SMPL parameters as GTR. In Tab 1, although directly regressing SMPL parameters as GTR reduces computation costs and enhances efficiency, it greatly reduces accuracy compared to proposed NSR for GTR.

Table 1. Comparison between different approaches for GTR. We test on FastMETRO with EfficientNet-b0 (Eb0) as a backbone.

Method	GFLOPs↓	Throughput↑	MPJPE↓	PAMPJPE↓
Parameter regression for GTR	?	?	70.2	49.4
NSR for GTR (Ours)	1.6	876.3	**63.2**	**43.9**

[R1] Can NSR be pertrained? No. Since NSR relies on the tokens from the main transformer for training, it needs to be trained together with the main transformer.

B. [R1] Overall memory footprint. Tore effectively saves memory usage: During inference, memory usage when processing one image is saved by ?; 2) During training, the memory costs of a transformer encoder structure and a transformer encoder-decoder structure are saved for **58.3%** and **27.4%** in a standard training scheme with HRNet-W64 as a backbone. More details are in Appendix B.2.

C. [R2] Other related works in token clustering for HMR. The works mentioned in the review are different from ours. 1) TCFormer [4] is a multi-stage method which is much more complicated. The token clustering is progressive with KNN Density Peaking Clustering. A multi-stage Token Aggregation is involved in aggregating features in these stages. There is no consideration of redundancy in the body representation like ours. In contrast, Tore (ours) is specifically proposed for HMR driven by two aspects from both 3D geometry structure (body representation) and 2D image features. Particularly, Tore effectively reduces visual tokens requiring only a single pass. A comparison with TCFormer with the same setting is in Tab 2, where ours surpasses TCFormer on both two datasets. 2) [5] mentioned in the review is **not** for token reduction for an efficient HMR. Instead, it recovers human mesh in UV space to leverage the correspondence between the mesh and the input image. 3) [3] is also **not** for token reduction either, and it even does **not** involve transformers or focus on efficient HMR. It first learns dense correspondence between vertices and the image in UV space with a non-parametric model, then using an IK module to regress SMPL parameters. We will cite all suggested papers. Thanks for improving this submission.

Table 2. TCFormer [4] v.s. Tore (Ours) for HMR on Human3.6M.

Method	Throughput	3DPW		Human3.6M	
		MPJPE	PAMPJPE	MPJPE	PAMPJPE
TCFormer [4]	163.9	80.6	49.3	62.9	42.8
Transformer Encoder+Tore (Ours)	210.1	75.5	46.6	**57.6**	37.1
Transformer Encoder-Decoder+Tore (Ours)	249.2	**72.3**	**44.4**	59.6	**36.4**

E. [R2] More comparisons with other token reduction methods. We compared with PPT [2] that prunes tokens by locating human visual tokens with attention score; see Table 3 where ours is more competitive. A comparison with Not All Tokens are Equal (TCFormer) [4] is given in Tab 2. We cannot compare with TransFusion [1] as it is designed for multi-view 3D human pose estimation (for key points), whereas ours is for HMR from a single-view image. All results and analyses will be included in the revision.

Table 3. Performance statistics of PPT and Tore (Ours). We test on FastMETRO with Eb0 as backbone for HMR on Human3.6M.

Method	GFLOPs↓	Throughput↑	MPJPE↓	PAMPJPE↓
PPT	?	?	68.4	46.2
Tore (Ours)	?	?	**63.2**	**43.9**

Table 4. Influence of different pruning rates. We test on FastMETRO with Eb0 as backbone for HMR on Human3.6M.

Metrics	No Pruning	Tore@0.2	Tore@0.5	Tore@0.75
PAMPJPE ↓	45.8	**43.9**	43.9	44.7
MPJPE ↓	69.2	**63.2**	64.2	65.3
GFLOPs ↓	7.1	1.6	1.4	1.2

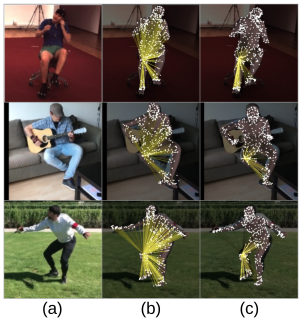
D. [R2] "The throughout and flops look good, while MPVE/MPJPE/PMPJPES doesn't look amazing." This work aims to save computation costs while maintaining the competitive accuracy of HMR. Although token reduction leads to drops in accuracy, the drop is slight. As said by R1 and R3, "efficiency got significantly improved while the accuracy was not affected" and "significantly reduced computational cost while still achieving competitive performance". Tore improves the accuracy on the backbones ResNet and Eb0, and hand mesh recovery while greatly saving the cost; See main paper Tab 1, 2 and 8.

G. [R3] Practical impact of improvement in GFLOPs and Throughput. During inference, FM-EB0+Tore increases the throughput by **68%** to 870.5 images/s compared with the original FM-EB0. The runtime for each image is saved for 68%. During training, FM-EB0+Tore saves the

memory costs by ? from ? to ? compared with FM-EB0. These suggest a superior improvement for real-world deployment.

H. [R3] How is the pruning rate determined? The pruning rate is a hyperparameter. Ablation study on different pruning rates is in Tab 4. By default, we set $= 20\%$, which is a good trade-off in practice.

I. [R3] More qualitative results of ablation studies. We will add more qualitative results of ablations in the revision. Here, we show that the model FM-Eb0-Tore@20% (c) learns joint-vertex interactions in a manner like the blending weights in SMPL that efficiently encodes body shape, whereas the baseline model FM-Eb0 (b) redundantly models local joints and global vertices, making the interaction costly. below can be removed.


(a) (b) (c)

I. [R1,R2,R3]Ablations & Exposition & References We will carefully revise the paper and add all suggested experiments, discussions and references.

J. [R3] Qualitative ablation results. The visualization of the cross attention w./w.o. Tore is shown here. We will supplement more results in the revision.

How to make them into one page... I, J and references...

References

[1] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S.-Y. Lin, and X. Xie. Transfusion: Cross-view fusion with transformer for 3d human pose estimation. *arXiv preprint arXiv:2110.09554*, 2021. 1

[2] H. Ma, Z. Wang, Y. Chen, D. Kong, L. Chen, X. Liu, X. Yan, H. Tang, and X. Xie. Ppt: token-pruned pose transformer for monocular and multi-view human pose estimation. In *ECCV*, 2022. 1

[3] Z. Wang, J. Yang, and C. Fowlkes. The best of both worlds: combining model-based and nonparametric approaches for 3d human body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2318–2327, 2022. 1

[4] W. Zeng, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, and X. Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11101–11111, 2022. 1

[5] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang. 3d human mesh regression with dense correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7054–7063, 2020. 1