

---

# Unified 2D-3D Discrete Priors for Noise-Robust and Calibration-Free Multiview 3D Human Pose Estimation (Supplementary Material)

---

Geng Chen\* Pengfei Ren\* Xufeng Jian Haifeng Sun<sup>†</sup> Menghao Zhang

Qi Qi Zirui Zhuang Jing Wang Jianxin Liao Jingyu Wang<sup>†</sup>

State Key Laboratory of Networking and Switching Technology,

Beijing University of Posts and Telecommunications

{chengeng, rpf, jianxf, hfsun, zhangmenghao, qiqi8266, zhuangzirui, wangjing, liaojx, wangjingyu}@bupt.edu.cn

In the supplementary material, the following contents are provided:

- Section **A**: Effect of training data volume on model performance, showing consistent improvements with larger datasets.
- Section **B**: Impact of the number of frames and views on performance.
- Section **C**: Visualization of the 2D-3D representation gap and the role of the *UniCodebook* in aligning heterogeneous pose representations to improve performance.
- Section **D**: Validation of our proposed DCSA architecture by comparing it against alternative designs for discrete-continuous feature interaction.
- Section **E**: Analysis of the impact of the *UniCodebook* on temporal modeling under both clean and realistic noisy conditions to evaluate its robustness.
- Section **F**: Investigation of the generalization ability to inter-view interactions by evaluating the effects of view permutation and reduction.
- Section **G**: Visualization of the MPJPE error distributions to compare models with and without the codebook, showing its impact on reducing overall prediction errors.
- Section **H**: Visualization of qualitative comparisons and attention heatmaps to illustrate how the codebook improves predictions.
- Section **I**: Discussion of current limitations of our method and outline potential directions for addressing them in future work.
- Section **J**: Discussion of both positive and negative potential societal impacts of our method.

## A How Do the Training Data Volume Impact Performance?

To investigate the effect of training data volume on the effectiveness of *UniCodebook* in Stage I and its subsequent influence on Stage II, in Table 6, we conduct four independent experiments using 25%, 50%, 75%, and 100% of the valid set of AMASS dataset. In Stage I, increasing the amount of training data leads to a consistent reduction in MPJPE across all four strategies (with a maximum reduction of 0.7mm or 2.6%), while the overall activation rate of the codebook remains stable. This indicates that a larger training set allows the *UniCodebook* to learn a more comprehensive and well-structured discrete representation space. In Stage II, models trained with more data exhibit greater performance improvements, suggesting that a well-trained codebook, enriched with a broader range of pose variations, enhances its effectiveness as a robust prior. This highlights the necessity of utilizing a diverse and extensive dataset to fully exploit the advantages of discrete representations in pose lifting.

---

\*Equal contribution.

<sup>†</sup>Corresponding authors.

Table 6: Ablation on the impact of the AMASS [3] data ratio on the *UniCodebook* at Stage I and its effect on lifting performance improvement at Stage II.

Data Ratio	Stage I				Stage II	
	MPJPE↓				Mean Active Rate(%)↑	MPJPE↓
	2Dto2D	2Dto3D	3Dto2D	3Dto3D		
25%	0.007	33.01	0.007	9.33	83.4	26.74
50%	0.007	30.95	0.006	8.80	<b>87.9</b>	26.55
75%	0.006	29.24	0.006	8.27	82.3	26.51
100%	<b>0.006</b>	<b>28.81</b>	<b>0.005</b>	<b>7.84</b>	83.5	<b>26.34</b>

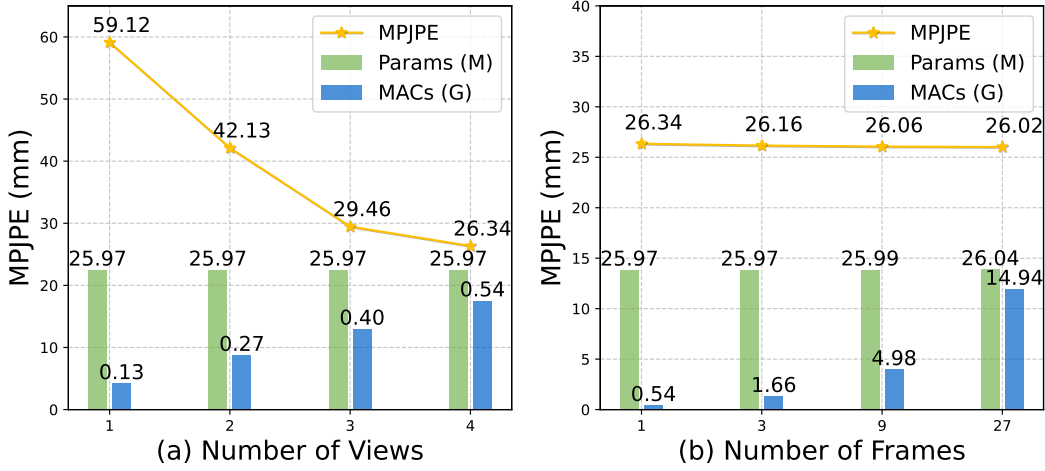


Figure 4: Ablation on the number of views and frames, reporting MPJPE, the number of parameters (M), and the computation cost MACs (G).

## B How Do the Number of Frames and Views Impact Performance and Computation?

Figure 4 shows that performance improves with more frames and views, highlighting the importance of temporal and multiview information. The performance gain is most pronounced with increasing views, emphasizing the value of complementary perspectives. Meanwhile, the computational cost grows approximately linearly with the number of frames and views.

## C Analysis of 2D-3D Representation Gaps and the Contribution of the *UniCodebook*

As illustrated in Figure 5 (a), 2D and 3D features from separate 2D/3D codebooks exhibit significant distribution discrepancies, manifesting as minimal overlap and distinct clustering patterns. In contrast, Figure 5 (b) reveals the unified feature space achieved through the proposed *UniCodebook*, where both modalities converge into a coherent distribution. This alignment gap in (a) poses critical challenges for 2D-3D lifting tasks. The disjoint distributions force the backbone model to expend additional computational resources on reconciling modality-specific representations (*e.g.*, learning cross-modal feature transformations) rather than focusing entirely on learning robust 2Dto3D mappings. The *UniCodebook* in (b) directly addresses this bottleneck by enforcing cross-modal consistency at the foundational representation level. Through multi-strategy learning, the latent space inherently encodes shared pose semantics across modalities, effectively implicitly decoupling the lifting model from manual feature alignment tasks in Stage II, thereby enhancing both generalization and precision in real-world scenarios. This is demonstrated in Table 4 (a), where MPJPE is reduced by 0.4mm.

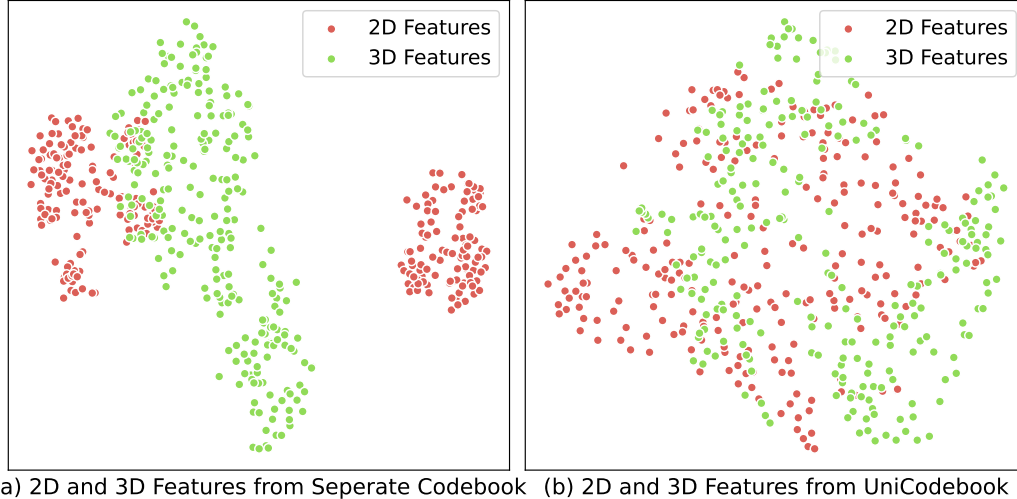


Figure 5: Illustration of the 2D-3D feature gap using t-SNE [5]: (a) features from separate codebooks trained for 2Dto2D and 3Dto3D, respectively, and (b) features from the proposed *UniCodebook*.

Table 7: Ablation study on DCSA design choices.

model	MPJPE↓
Two-path, Weighted Sum	26.39
Single-path	26.51
Two-path, Direct Sum (ours)	<b>26.34</b>

Table 8: Temporal results for models trained on Human 3.6m GT data with T=243.

model	test MPJPE (GT 2d)	test MPJPE (CPN 2d)
MixSTE [6]	21.6	64.6
KTPFormer [4]	<b>19.0</b>	66.6
ours	23.1	<b>52.8(20.7%↓)</b>

## D Which DCSA Design is Most Effective?

We validate our Discrete-Continuous Spatial Attention (DCSA) architecture by comparing it against two alternatives, with results summarized in Table 7. First, we replaced the simple summation of the two branches with a learnable weighted sum. Second, we unified the two-path attention into a single-path attention where continuous queries attend to both continuous and discrete features simultaneously. Both modifications failed to improve performance. The weighted sum offered no benefit, suggesting the model implicitly balances the branches effectively. The single-path attention performed worst, likely because separating the paths allows one branch to specialize on continuous spatial relations and the other on discrete structural priors, thus avoiding feature interference. These results confirm that our proposed two-path design, combined via simple summation, is the most effective.

## E Does the *UniCodebook* Benefit Temporal Modeling?

### E.1 Performance Drop in Standard Temporal Setting

Our model shows clear advantages in multi-view fusion but no gains in temporal modeling. We guess this difference stems from the intrinsic properties of the data.

In the multi-view setting, poses captured from different viewpoints exhibit large appearance and coordinate discrepancies (e.g., an average MPJPE of 161 px between views in Human3.6M). The

Table 9: Results of testing with shuffled and varying view configurations. All models were trained on four views in a fixed order.

model	test shuffle	test views	MPJPE
ours w/ PE	×	4	26.0
ours w/ PE	✓	4	59.7
ours wo/ PE	×	4	26.1
ours wo/ PE	✓	4	26.1
ours wo/ PE	×	3	59.3

*UniCodebook* provides a shared latent representation, aligning heterogeneous 2D observations of the same 3D pose and enabling stable cross-view fusion.

In the temporal setting, however, adjacent 2D poses are already smooth and highly correlated (*e.g.*, an average MPJPE of 1 px between frames in Human3.6M). The temporal cues mainly rely on fine-grained motion continuity, which continuous features can naturally capture. Discretization through the codebook may blur these subtle variations and disrupt temporal smoothness, thereby degrading performance compared with transformer-based models that directly process continuous trajectories.

## E.2 Robustness Under Realistic Conditions (GT→CPN Setting)

Although our method falls short of SOTA under clean GT 2D inputs, it shows superior robustness when evaluated under realistic noise. We adopt a “train-on-GT, test-on-CPN” protocol, where models trained with ideal 2D poses are tested using noisy 2D detector predictions (*e.g.*, from CPN). This better reflects real-world deployment, where 2D detectors are affected by occlusion and motion blur.

As shown in Table 8, our model achieves a 13.8 mm (20.7%) improvement over the strongest baseline (KTPFormer [4]) under noisy inputs, highlighting its noise resilience and stable generalization. Therefore, while our model may underperform slightly in clean temporal benchmarks, it demonstrates clear practical advantages under realistic, imperfect 2D detections.

## F Does the Model Learn Generalizable Inter-View Interactions?

To investigate whether the model learns generalizable inter-view relationships, rather than overfitting to specific camera geometries, we evaluated its performance under two challenging conditions: a shuffled (permuted) view order and a reduced number of views. These tests are designed to probe the model’s robustness to variations in the multiview configuration at test time. The results are summarized in Table 9.

Our analysis yields two key insights. First, the experiments highlight that Positional Encoding (PE) is the primary component that hinders generalization to permuted view orders. As shown in Table 9, the model variant incorporating PE (ours w/ PE) experiences a severe performance collapse when the view order is shuffled, with the MPJPE dramatically increasing from 26.0mm to 59.7mm. This suggests that PE can cause the model to learn view-dependent biases specific to the training data. In contrast, the model without PE demonstrates complete robustness to view permutation, maintaining a stable MPJPE. This is an expected outcome, as the network is composed entirely of attention and feed-forward network (FFN) layers, both of which are inherently permutation-invariant. Without an explicit positional signal from PE, the model treats the input views as an unordered set. It is also worth noting that the inclusion of PE can accelerate model convergence.

Second, while the model without PE is robust to view permutation, its performance still degrades significantly when the number of input views is reduced. As shown, when testing the 4-view trained model on only 3 views, the MPJPE rises to 59.3mm. This limitation is not a consequence of the *UniCodebook* but rather points to the continuous representation-based baseline, which learns to operate on a fixed number of views during training. Designing a framework that can effectively generalize to an arbitrary number of views without performance degradation remains a compelling direction for future research.

## G Which Actions Show the Largest Error Drop? Action-Level Error Distribution Analysis with Codebook Priors

Figure 6 visualize the MPJPE error histograms of models trained *without* and *with* the proposed codebook prior. We observe a clear leftward shift in the overall distribution across all action categories when the codebook is used, indicating that more predictions fall into lower-error regions. This demonstrates that the discrete codebook serves as an effective noise-resilient prior, especially in scenarios where 2D keypoint detections are noisy and even unreliable.

Among all actions, *sitting* and *sitting down* exhibit the most significant relative improvements. These actions involve severe self-occlusion, leading to increased noise in the 2D keypoint detections. The codebook, trained on large-scale high-quality poses, offers strong structural priors that help correct such noisy inputs, demonstrating its utility in handling challenging conditions.

## H Qualitative Results

In Figure 7, we present qualitative results of the *baseline* and the *baseline with codebook* on the Human3.6M dataset. It can be observed that equipping the baseline with the codebook leads to notable improvements across various action categories, particularly those involving significant self-occlusion or challenging conditions that result in low-quality 2D pose detections. Most of these improvements occur at the distal joints, where inter-joint constraints are weakest among all joints. The structural priors introduced by our codebook effectively compensate for this, demonstrating the superior robustness provided by the *UniCodebook*. In addition, we visualize both the joint-to-joint attention heatmap to observe how the model captures inter-joint dependencies, and the DCSA heatmap to examine which discrete tokens are attached to which joints or substructures.

## I Limitations and Future Work

While our method demonstrates strong performance across diverse pose estimation scenarios, several limitations remain.

- First, while the shared *UniCodebook* effectively captures anatomical structure, it is trained on static poses without a temporal prior. This is a deliberate design choice to focus on learning a discrete prior for anatomical structures and to maximize the use of single-pose data, which is significantly more abundant than sequential data. Nevertheless, a potential future work is to incorporate temporal consistency into the prior learning process, which may further enhance generalization to dynamic or motion-rich scenarios.
- Second, different datasets often contain varying human pose definitions, which means the *UniCodebook* must be retrained whenever the pose format changes. This limits its flexibility in cross-dataset applications. A possible direction to address this issue is to design a format-agnostic or adaptable prior that can generalize across different pose definitions without requiring full retraining, such as using a shared intermediate representation or learning a mapping layer.
- Third, the benchmark datasets we use lack reporting on subject demographics, particularly racial and ethnic diversity. Such biases primarily manifest in the image domain, which may degrade the performance of 2D pose detectors on underrepresented groups. Since our work focuses on 2D-to-3D pose lifting, which processes 2D coordinates rather than raw images, the direct impact of these appearance-based biases is mitigated. Nevertheless, we acknowledge that biases from upstream detectors can propagate, and we advocate for future work to create more demographically balanced datasets and establish fairer evaluation protocols.

## J Broader Impacts

Our work on calibration-free multi-view 3D human pose estimation (HPE) has the potential to contribute positively to society by enhancing the accessibility and scalability of applications that rely on accurate motion understanding, such as sports analysis, surveillance, action recognition, and autonomous driving—where input data often contains noise due to occlusions, lighting variations, and other challenging conditions.

However, as with any technology that involves the capture and interpretation of human motion, there are potential negative societal impacts. First, surveillance systems could misuse uncalibrated multi-view HPE models to monitor individuals across public or private spaces without their consent, raising serious privacy concerns. Moreover, if the system functions incorrectly (*e.g.*, due to out-of-distribution inputs or adversarial noise), it may produce flawed interpretations of behavior that could unfairly influence downstream applications, such as automated grading in education or risk assessment in security. Additionally, since our approach improves robustness in noisy environments, it could be repurposed for malicious uses such as tracking individuals across low-quality camera feeds.

To mitigate these risks, future work could incorporate safeguards such as usage restrictions, privacy-preserving pose estimation techniques, and transparency tools that allow users to understand how and where their data is being processed. It is essential to ensure that deployment contexts respect consent, data minimization, and fairness principles.

## References

- [1] Y. Cai, W. Zhang, Y. Wu, and C. Jin. Fusionformer: A concise unified feature fusion transformer for 3d pose estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 900–908, 2024.
- [2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 36(7):1325–1339, 2013.
- [3] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019.
- [4] J. Peng, Y. Zhou, and P. Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1123–1132, 2024.
- [5] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(11), 2008.
- [6] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13232–13242, 2022.

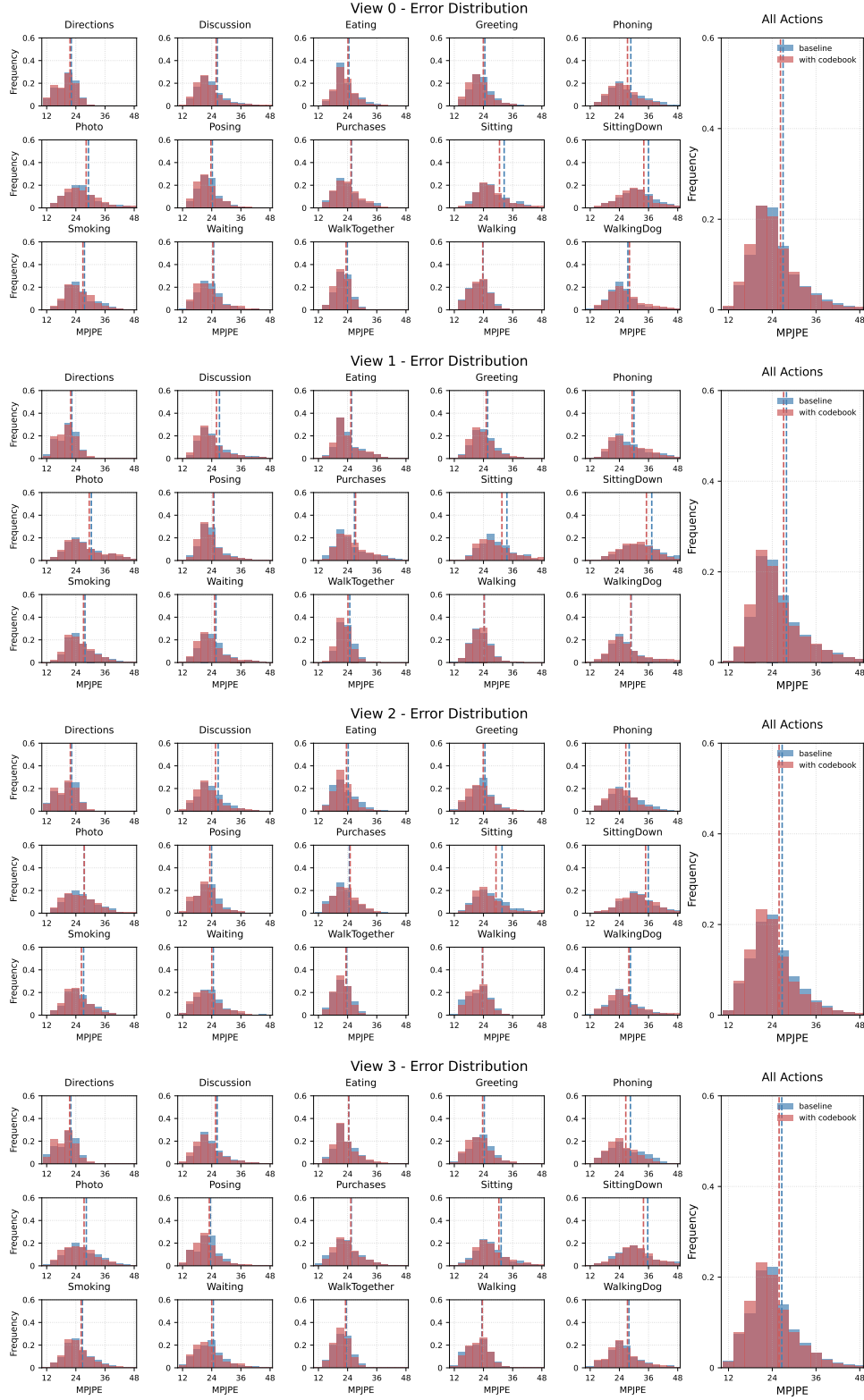


Figure 6: Comparison of MPJPE error distributions for the Human 3.6M dataset across view 0,1,2,3, showing model performance **without** codebook and **with** codebook. The dashed lines indicate mean MPJPE values (**blue** for baseline, **red** for codebook-enhanced). The x-axis starts from 10mm to focus on the meaningful error range. The models were trained using all 4 camera views.



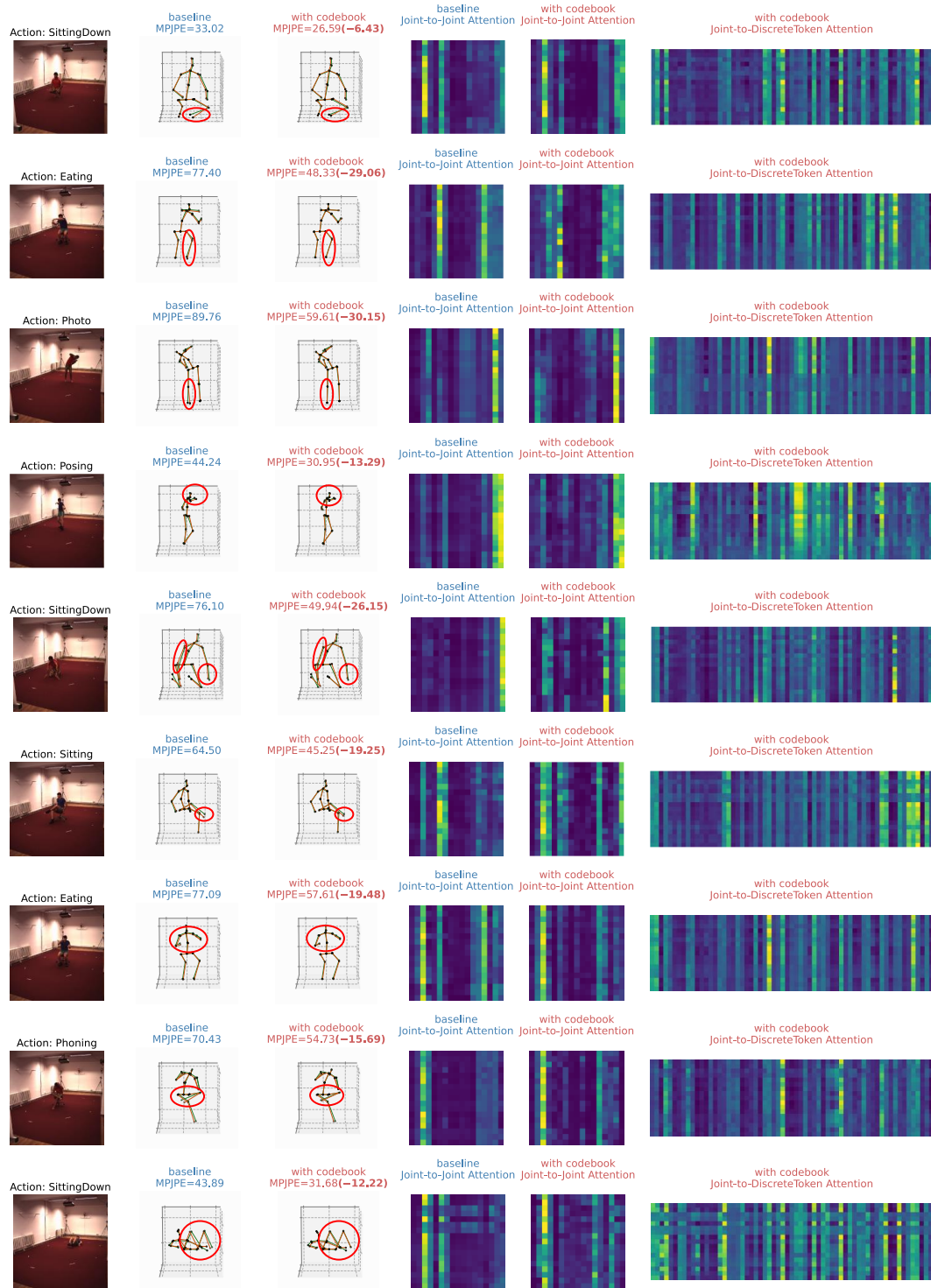


Figure 7: Qualitative comparisons of 3D human poses estimated by the **baseline** and the **baseline with codebook**. The **orange skeleton** denotes the prediction, while the **green skeleton** indicates the ground truth. Additionally, we visualize the joint-to-joint attention heatmap and DCSA heatmap (joint-to-DiscreteToken Attention in the figure) in the first spatial block. Both models are trained with 4 views, but for space efficiency, we only present the images and predictions from view 0.