Test-time Personalizable Forecasting of 3D Human Poses
Supplementary materials

1. Details of Implicit Augmenter

The trainable implicit augmenter $\mathcal{A}_\phi^{(I)}$ can be considered as a diverse motion transformer with a similar network architecture of [4]. The main difference is that, instead of obtaining future activities, we aim to generate the diverse motion counterparts of the original observed sequence. Moreover, the training of $\mathcal{A}_\phi^{(I)}$ falls into the max-min adversarial learning scope, to ensure the diversity and preserve the semantic proximity of the $H = 8$ augmented samples. The network architecture of $\mathcal{A}_\phi^{(I)}$ is shown in Figure 1. We note that, the input of $\mathcal{A}_\phi^{(I)}$ is a sequence of T frames, and the output is $H = 8$ diverse augmentations with distinct properties. Similar to [4], the DCT and IDCT are the discrete cosine transform and its inverse version, respectively. Moreover, its pipeline is to first yield the lower-part of a human skeleton using $\mathcal{G}^{(1)}$, and then the upper-part is obtained by $\mathcal{G}^{(2)}$. Both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ are composed of 3 blocks, each of which contains 3 fully-connected GCN layers. Consistent with [4], $\mathbf{z}_h^{(1)}$ and $\mathbf{z}_h^{(2)}$ are the Gaussian noise samples[1].

2. Supplement to Implementation Details

Our helper and predictor networks involve the same architecture, derived from an existing deep end-to-end predictive model. At the H/P domain-generalizable learning stage, the mini-batch size is set to 32. We note that, for both H3.6M and GRAB, the model (including the implicit augmenter, and helper/predictor) is trained with 100 epochs, while for HumanEVA-I, it is 50 epochs. Instead of using the early-stopping strategy, we make the parameters of the last epoch as the base model. The whole model is implemented on PyTorch-1.9 framework, and trained on a single NVIDIA Tesla V100 GPU.

3. Results on HumanEva-I

For HumanEva-I dataset [5], we leverage the 3 subjects (S_1, S_2, S_3) as our dataset. To evaluate the personalization capability of the proposed H/P-TTP for unseen subjects, the specific data partitioning strategy is shown in the setup-2.3 in Table 1. Consistent with Table 4 and Table 5, both SoTA approaches of PGBIG and SPGSN are selected as the baselines. The detailed results are reported in Table 1. We observe that, equipped with our H/P-TTP, all baselines achieve better results. It evidences that the proposed H/P-TTP is indeed able to adapt to the unseen properties of new subjects, which is consistent with the conclusion of the H3.6M and GRAB datasets in the manuscript.

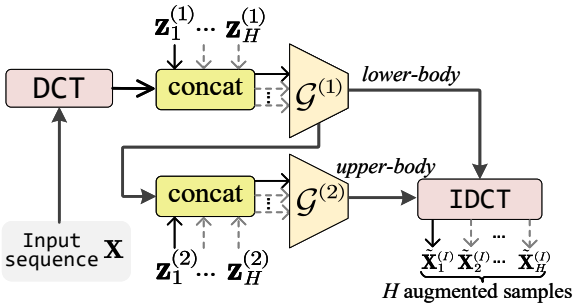

Figure 1. Network architecture of the implicit augmenter $\mathcal{A}_\phi^{(I)}$.

4. More Visualization

We supplement the visualization of all predicted poses of 2 unseen subjects from the H3.6M and one additional subject from the GRAB dataset. The specific visualization of the predicted poses is shown in Figure 2 and Figure 3.

5. Continue Personalization for More Subjects

Numerous experiments in the manuscript, as well as in this supplementary material, have investigated the personalized predictive ability of the proposed H/P-TTP for a single unseen subject. However, it may involve more unseen subjects with varied properties in the real-world application. Therefore, we further investigate the personalized ability of the proposed H/P-TTP for more unseen subjects. To be precise, we select the motion sequences of two subjects, S_1 and S_{10}, from the GRAB as the test set, and the disjointed 8 subjects as the training. Then, starting with the base model obtained from the training set, we alternately select a motion sample from S_1 and S_{10} to validate the personalized predic-

[1]To make a clear distinction, we use the blue color to quote the original manuscript, and the red color to indicate this supplementary material.

Unseen test subjects	MPJPE [mm] ↓				PMPJPE [mm] ↓				PCK@150mm [%] ↑				MPBLE [mm] ↓			
	PGBIG [3]	PGBIG [3] +H/P-TTP	SPGSN [2]	SPGSN [2] +H/P-TTP	PGBIG [3]	PGBIG [3] +H/P-TTP	SPGSN [2]	SPGSN [2] +H/P-TTP	PGBIG [3]	PGBIG [3] +H/P-TTP	SPGSN [2]	SPGSN [2] +H/P-TTP	PGBIG [3]	PGBIG [3] +H/P-TTP	SPGSN [2]	SPGSN [2] +H/P-TTP
S_1	96.4	83.3	82.1	**77.8**	67.9	64.6	67.2	**61.5**	77.9	79.3	82.1	**85.3**	26.2	23.1	24.0	**22.1**
S_2	90.1	87.4	88.5	**83.0**	74.2	70.9	74.3	**68.2**	74.2	76.4	79.6	**81.1**	23.4	21.0	20.3	**18.7**
S_3	86.7	82.0	78.8	**74.4**	73.5	70.0	69.6	**66.2**	74.8	77.2	75.3	**79.5**	24.0	23.2	22.4	**20.5**
Average	91.1	84.2	83.1	**78.4**	71.9	68.5	70.4	**65.3**	75.6	77.6	79.0	**82.0**	24.5	22.4	22.2	**20.4**

Table 1. Average performance comparison (of both SoTA PGBIG [3] and SPGSN [2], on HumanEVA-I dataset [5]) of the end predicted pose (1000ms) over samples of each unseen subject S_x with $x = [1, 2, 3]$, and the corresponding average over all unseen subjects.

personalization	S_1	S_1	S_1 ($S_1 \leftrightarrow S_{10}$)	S_{10}	S_{10}	S_{10} ($S_1 \leftrightarrow S_{10}$)
	w/o TTP	single unseen subject	alternating unseen subjects	*w/o* TTP	single unseen subject	alternating unseen subjects
Method	SPGSN [2]	SPGSN [2] +H/P-TTP	SPGSN [2] +H/P-TTP	SPGSN [2]	SPGSN [2] +H/P-TTP	SPGSN [2] +H/P-TTP
MPJPE ↓	176.3	151.5	166.5	144.5	136.4	142.8
P-MPJPE ↓	149.6	137.0	145.0	129.0	117.3	124.9
PCK@150mm↑	64.1	68.6	65.7	67.8	70.4	65.7
MPBLE ↓	27.1	25.0	26.4	26.6	24.7	25.3

Table 2. Average performance (of the end predicted pose) for evaluating the continue personalization of two alternating unseen subjects (S_1 and S_{10}) from GRAB dataset. We abbreviate the test-time personalization to TTP for brevity.

tion ability for different subjects. We note that this alternating approach considers the continuously changing subjects. Compared to first evaluating the results on all samples on S_1 and then on all samples on S_{10}, this alternating setup is more challenging and realistic. We then report the average performance (under 4 protocols) over all motion samples of S_1 and S_{10} in Table 2.

From the result, we can derive the following key observations: 1) Either for the personalization to a single or more unseen subjects, for the vanilla baselines, once the proposed H/P-TTP is assembled, the overall performance is better. This confirms that the proposed H/P-TTP is indeed able to adapt to the properties and motion patterns of various unseen characters in the testing phase. 2) The personalization of multiple subjects is typically more challenging than that of a single subject, because of the continuously changing individual properties. However, the performance of our H/P-TTP is also better than the vanilla baselines, even for alternating unseen subjects, which is acceptable in practice. 3) The setup of continuously changing subjects can be easily extended to more (> 2) unseen subjects, which is not fundamentally different from the experimental setup of 2 alternating unseen subjects. The above analysis shows the scalability of our H/P-TTP for more unseen subjects in deployment environments.

6. Supplement to Ablation Studies

In this section, we supplement the ablation studies of the proposed H/P-TTP. All results are evaluated using the same experimental setting as the main manuscript.

In our work, in order to establish the relationship between helper and predictor features, and their output with respect to GT, we exploit the network layers before the penultimate layer as the feature *extractor* and the remaining as the *generator*. To verify it, we use different layers as **(9) the division points of the extractor and generator**. The results are shown in Table 2(left), where $-i$ layer means the penultimate i-th layer.

	division point	MPJPE [mm] ↓		μ	MPJPE [mm] ↓
SPGSN [2] +H/P-TTP	**-2**	103.4	SPGSN [2] +H/P-TTP	0.93	102.8
	-3	109.2		**0.95**	109.2
	-4	109.2		0.97	109.2

Table 3. Impact of the division points of the extractor and generator, and the momentum size μ on the final performance.

	fallback strategy	P	MPJPE [mm] ↓
SPGSN [2] +H/P-TTP	*w/o*	/	125.0
	w/	48	109.3
		60	106.5
		72	**102.8**
		84	114.2

Table 4. Impact of the fallback strategy, as well as the number of steps P on MPJPE of the SPGSN [2]+H/P-TTP.

Our proposed framework is a special case of the teacher-student network, and thus the exponential moving average (EMA) is used to update the helper with momentum $\mu = 0.95$ after updating the predictor. Therefore, we also investigate the **(10) impact of different momentum sizes** on the final performance, as shown in Table 2(right).

In addition, to avoid forgetting the pre-trained knowledge, we back off the parameters after $P = 72$ test-time personalization steps to the base model. To verify the effectiveness of the **(11) fallback strategy**, we run the experiments with and without it, as well as the different number of steps P. The results are shown in Table 3.

Figure 2. Detailed visualization of predicted poses. We show the walking (top) and direction (bottom) of the unseen test subject S_{11} from the H3.6M [1]. We observe that, with the help of the H/P-TTP, the predicted pose is more accurate and stable than the vanilla SPGSN.

Figure 3. Detailed visualization of predicted poses of the hammer-pass-1 activity of the unseen subject S_7 from GRAB dataset.

7. Limitation

One limitation of our H/T-TTP is that when estimating the adjusted learning rate at test time for the first M samples, the M-sized memory queue $\mathbb{M}$ has not been constructed yet. To be compatible with this situation, we randomly select M samples from the training set in advance, and take the corresponding h/p feature difference and p/gt outcome difference as the initialization of $\mathbb{M}$. Note that, h, p are the abbreviations for the "helper" and "predictor", and gt is the ground truth.

References

[1] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36:1325–1339, 2014. 3

[2] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-Parted Graph Scattering Networks for 3D Human Motion Prediction. *arXiv preprint arXiv:2208.00368*, 2022. 2

[3] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *CVPR*, pages 6437–6446, 2022. 2

[4] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *ICCV*, 2021. 1

[5] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, 2010. 1, 2