Robust Long-Term Test-Time Adaptation for 3D Human Pose Estimation Through Motion Discretization

Yilin Wen, Kechuan Dong, Yusuke Sugano

The University of Tokyo
{fylwen, kchdong, sugano}@iis.u-tokyo.ac.jp

Abstract

Online test-time adaptation addresses the train-test domain gap by adapting the model on unlabeled streaming test inputs before making the final prediction. However, online adaptation for 3D human pose estimation suffers from error accumulation when relying on self-supervision with imperfect predictions, leading to degraded performance over time. To mitigate this fundamental challenge, we propose a novel solution that highlights the use of motion discretization. Specifically, we employ unsupervised clustering in the latent motion representation space to derive a set of anchor motions, whose regularity aids in supervising the human pose estimator and enables efficient self-replay. Additionally, we introduce an effective and efficient soft-reset mechanism by reverting the pose estimator to its exponential moving average during continuous adaptation. We examine long-term online adaptation by continuously adapting to out-of-domain streaming test videos of the same individual, which allows for the capture of consistent personal shape and motion traits throughout the streaming observation. By mitigating error accumulation, our solution enables robust exploitation of these personal traits for enhanced accuracy. Experiments demonstrate that our solution outperforms previous online test-time adaptation methods and validate our design choices.

1 Introduction

Estimating 3D human body pose is essential for interpreting human behaviors. Given streaming video input, accurate pose estimation enables prompt evaluation of user performance, facilitating applications such as human-robot collaboration, physical training, and immersive interactions. Despite recent advances of learning-based methods (Kolotouros et al. 2019; Kocabas, Athanasiou, and Black 2020; Goel et al. 2023), pre-trained pose estimators often suffer from performance degradation in real-world scenarios that deviate from their training domains.

To address this issue, recent studies have explored online test-time adaptation. During test time, given the input unlabeled streaming video, these methods employ self-supervised learning to continuously update the model. This allows for online 3D pose estimation for each incoming batch while gradually adapting to the test domain, aiming to capture domain-specific patterns for enhanced accuracy over time. For example, BOA (Guan et al. 2021) and DynaBOA (Guan et al. 2022) adapt by supervising with ground-truth 2D poses and using images from the pre-training dataset as exemplars. CycleAdapt (Nam et al. 2023) further extends to practical scenarios by referring to 2D pose detections, and enhances adaptation by employing denoised motion estimations as pseudo labels to capture 3D geometry.

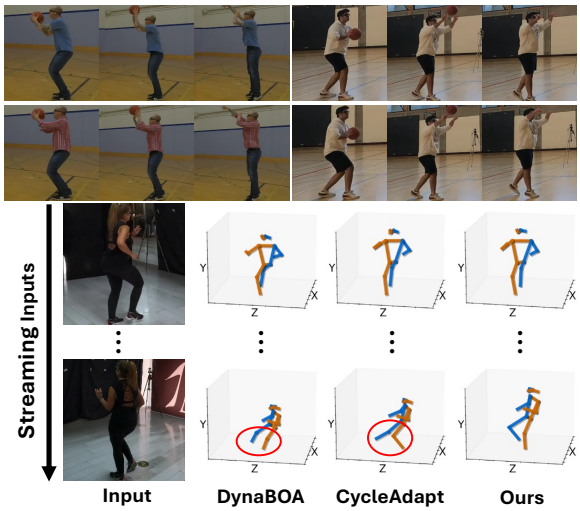

Figure 1: Illustration of personal shape and habitual motion traits across observations (upper) and error accumulation in existing works as adaptation progresses (lower).

Nonetheless, as shown in Fig. 1, we identify two drawbacks in existing works: First, these works can be sensitive to self-supervised signals derived from imperfect 2D pose detection and 3D pose estimations, which further leads to error accumulation as prediction errors compound over time, hindering accurate long-term adaptation. Second, individuals often exhibit consistent shape and habitual motion traits. Model adaptation should benefit from capturing these unique personal patterns when continuously observing and adapting to a single subject. However, the struggle with error accumulation positions continuous adaptation as also a risk, leaving personalized adaptation not thoroughly explored.

To address these limitations, we propose a novel solution

that integrates motion discretization and a soft-reset strategy, thus mitigating error accumulation in long-term adaptation. In this way, our solution enables enhanced online test-time adaptation through continuously observing and adapting to a personalized test domain. This test domain comprises streaming videos featuring a single individual who is not included in the pre-training dataset, with total recording durations potentially extending to tens of minutes. As shown in Fig. 2 and Alg. 1, our framework consists of two components, which are adapted alternately in a cyclic manner (Nam et al. 2023). The first component is a pose estimator that outputs 3D human poses corresponding to the input images. The second is an autoencoder-based motion denoising network that refines these estimations to generate 3D signals capturing the dynamics of human movements, which in turn regularize the adaptation of the pose estimator.

Our key innovations are threefold. First, we derive a discrete set of diverse anchor motions that capture plausible human movements. To achieve this, we perform unsupervised clustering on the latent space of the motion denoising network during pre-training, yielding a codebook of representative latent features. Each codebook entry is decoded into a distinct anchor motion representing a coherent motion pattern. During test time, these anchor motions provide regular supervision signals for regularizing the adaptation of the image-based pose estimator, thus mitigating the limitations of self-supervision using imperfect 3D estimations. Second, our motion discretization also enables a self-replay mechanism, where we adapt the motion denoising network using both incoming test-time estimations and pre-trained anchor motions. This ensures consistent decoding of regular anchor motions throughout the adaptation (Fig. 5), while eliminating the need to access the original pre-training data. Finally, we employ an efficient soft reset that periodically reverts the pose estimator to its exponential moving average during adaptation, thus reducing the impact of noisy updates while retaining critical test-time traits for robust adaptation.

We evaluate online personalized test-time adaptation on Ego-Exo4D (Grauman et al. 2024) and 3DPW (Von Marcard et al. 2018) datasets. In line with previous test-time adaptation approaches, we pre-train our model on the Human3.6M dataset (Ionescu et al. 2013). Results demonstrate our enhanced effectiveness compared to previous online test-time adaptation methods and verify our key designs. Our contributions are summarized as follows:

- We propose a novel solution for online test-time adaptation in 3D human pose estimation. Our solution highlights using motion discretization and integrates a soft-reset mechanism. In this way, we mitigate the inherent challenge of error accumulation in long-term self-supervised test-time adaptation.
- We examine online test-time adaptation by focusing on a personalized test domain. By mitigating error accumulation, our solution enables average performance gains through continuous adaptation on streaming videos featuring the same person.
- We demonstrate enhanced accuracy over existing online test-time adaptation methods.

2 Related Works

3D Human Body Pose Estimation Massive learning-based research has advanced 3D human body estimation given monocular RGB observation. This is achieved by exploiting the spatial dimensions to output body poses from single images (Kanazawa et al. 2018; Kolotouros et al. 2019; Moon and Lee 2020; Zhang et al. 2021), or further extending to video inputs and further harnessing temporal continuity for enhanced robustness (Kanazawa et al. 2019; Kocabas, Athanasiou, and Black 2020; Choi et al. 2021; Wei et al. 2022; Shen et al. 2023; You et al. 2023). These solutions primarily scale training data to enhance the generalizability of pre-trained models across various test scenarios, where recent research (Dwivedi et al. 2024; Goel et al. 2023) further leverages more powerful backbones such as ViT (Dosovitskiy et al. 2020). Additionally, recent studies also explore pose augmentation for enhanced accuracy (Gong, Zhang, and Feng 2021; Chai et al. 2023).

In contrast, another branch of solution addresses from the perspective of test-time adaptation, which directly updates the pre-trained model on unlabeled inputs prior to final estimation, thus mitigating the inherent train-test domain gap for enhanced performance (Joo, Neverova, and Vedaldi 2021; Mugaludi et al. 2021; Guan et al. 2021, 2022; Weng et al. 2022; Nam et al. 2023; Lin et al. 2025a,b). Specifically, when it comes to online test-time adaptation for streaming inputs, BOA (Guan et al. 2021) and DynaBOA (Guan et al. 2022) use a bilevel adaptation strategy that updates an image-based pose estimator with supervision of ground-truth 2D keypoints and 2D temporal constraints, while source domain images provide 3D exemplar guidance. However, their inability to capture the 3D regularity of test motions hinders accurate depth estimation. Noticing the phenomenon of error accumulation, Lin *et al*. (2025b) restore the updated model to its pre-trained weights at the end of each video and reinitialize with representative historical frames. CycleAdapt (Nam et al. 2023) introduces a motion denoising network to refine the image-based pose estimation. The denoised motion provides 3D pseudo labels for adapting the image-based pose estimator, thereby establishing cyclic adaptation between the two modules. However, its overreliance on imperfect estimations limits its ability to manage error accumulation in long-term adaptation.

Compared to existing research, we address error accumulation in continuous self-supervised adaptation by leveraging motion discretization and soft reset. This further enables us to harness personal traits throughout continuous adaptation on observations featuring the same person, resulting in enhanced performance.

Quantized Body Pose Representation Although human body poses are inherently represented as continuous values in 3D space, recent research has exploited discrete pose representations for both human pose estimation and motion generation. On one hand, a trend in human motion generation (Lucas et al. 2022; Zhang et al. 2023; Guo et al. 2024) employs VQ-VAE (Van Den Oord, Vinyals et al. 2017) to compress and quantize consecutive pose frames into discrete latent tokens. These discrete tokens are then integrated

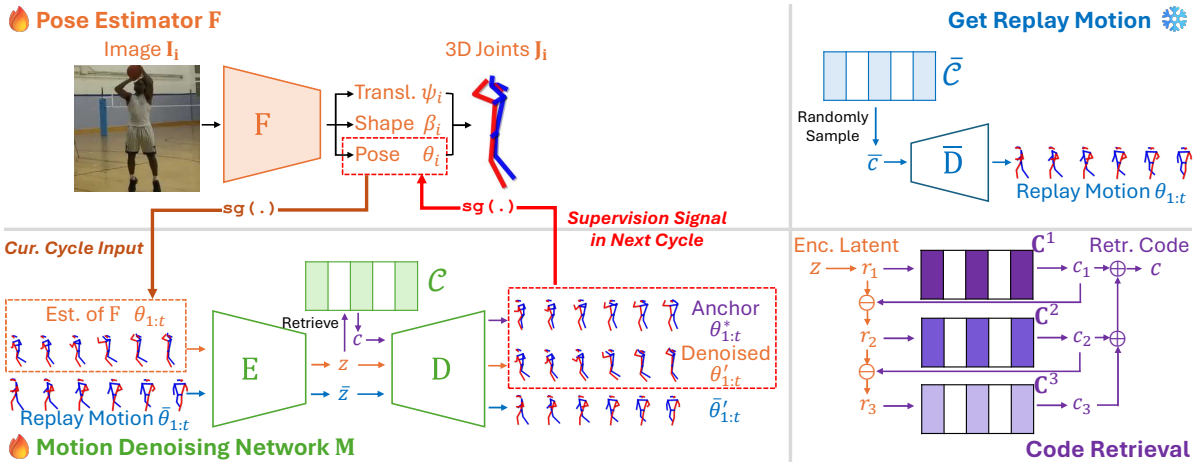

Figure 2: Framework Overview. During test time, the pose estimator F and motion denoising network M are alternately updated in a cyclic way. We employ motion discretization to regularize the adaptation of F and enable self-replay for adapting M.

into a GPT-like model, enabling long-term generation of valid and plausible poses through next-token prediction. On the other hand, recent research (Geng et al. 2023; Dwivedi et al. 2024) applies this quantization philosophy to facilitate image-based human pose estimation, as achieved by learning latent codebooks that capture the spatial relationships among groups of joints. This reduces pose estimation to a latent code prediction problem, which helps obtain valid outputs that accurately capture body pose physics.

In comparison, our solution emphasizes motion discretization in self-supervised adaptation for human body pose estimation. Derived from unsupervised clustering in the latent motion space, our anchor motions model the spatiotemporal relationships among joints by capturing both pre-trained regularity and test-time motion traits, thus mitigating error accumulation during long-term adaptation.

3 Methodology

Our online test-time personalized adaptation defines its test domain as a streaming video $\mathcal{S}$, which consists of concatenated image sequences featuring the same individual who was unseen during pre-training. During test time, we adapt the model on each incoming unlabeled batch $\mathcal{V} \subset \mathcal{S}$, making on-the-fly prediction of 3D body joints $\mathbf{J} \in \mathbb{R}^{N\times 3}$ for each image $\mathbf{I} \in \mathcal{V}$.

As shown in Fig. 2, our framework consists of two components: a pose estimator F and a motion denoising network M. F maps an input image $\mathbf{I} \in \mathcal{V}$ to SMPL (Loper et al. 2015) pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and translation $\boldsymbol{\psi}$, from which the 3D joints $\mathbf{J}$ are subsequently regressed. M is a denoised autoencoder that captures temporal continuity across outputs of F and generates 3D motion signals to help adapt the F. F and M are adapted alternately to capture test-time appearance and motion traits in a cyclic manner (Nam et al. 2023).

We summarize our test-time pipeline in Alg. 1 and introduce our key designs in the following subsections. We first introduce our motion discretization, which is achieved through unsupervised clustering and benefits test-time adaptation in two ways: first, it regularizes the adaptation of F to mitigate the effects of self-supervision with noisy estimations; second, it enables a self-replay mechanism when adapting the motion denoising network M, addressing representation drift to ensure consistent decoding of regular anchor motions. We further introduce our soft-reset strategy and provide implementation details.

Algorithm 1: Our adaptation pipeline.

Input: Streaming video $\mathcal{S}$ featuring the test person, pre-trained pose estimator $\bar{\mathrm{F}}$, motion denoising network $\bar{\mathrm{M}} = (\bar{\mathrm{E}}, \bar{\mathrm{D}})$ and codebook $\bar{\mathcal{C}}$

Output: 3D body keypoints $\mathbf{J} \in \mathbb{R}^{N\times 3}$ for every image $\mathbf{I} \in \mathcal{S}$.

Initialize $\mathrm{F}, \mathrm{E}, \mathrm{D}, \mathcal{C}$ with $\bar{\mathrm{F}}, \bar{\mathrm{E}}, \bar{\mathrm{D}}, \bar{\mathcal{C}}$, respectively
while incoming $\mathcal{V} \subset \mathcal{S}$ **do**
  Set $\mathrm{F}_{pre} \leftarrow \mathrm{F}$. Prepare replay motion $\bar{\boldsymbol{\theta}}_{1:t}$ by Eq. (4).
  for cycle=0,...,c **do**
    # Adapt Pose Estimator F
    Retrieve $\boldsymbol{\beta}', \boldsymbol{\theta}'$ and anchor $\boldsymbol{\theta}^*$ obtained in previous cycle
    $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{J} \leftarrow \mathrm{F}(\mathbf{I})$ for $\mathbf{I} \in \mathcal{V}$
    Update F with L_F (Eq. (3)) ▷ *loss with anchor motion*
    # Adapt Motion Denoising Network M
    $\boldsymbol{z} \leftarrow \mathrm{E}(\boldsymbol{\theta}_{1:t}), \bar{\boldsymbol{z}} \leftarrow \mathrm{E}(\bar{\boldsymbol{\theta}}_{1:t})$
    For $\boldsymbol{z}$, retrieve nearest code $\boldsymbol{c}$ from $\mathcal{C}$ (Eq. (1)).
    Anchor $\boldsymbol{\theta}^*_{1:t} \leftarrow \mathrm{D}(\boldsymbol{c})$, denoised $\boldsymbol{\theta}'_{1:t} \leftarrow \mathrm{D}(\boldsymbol{z})$, $\bar{\boldsymbol{\theta}}'_{1:t} \leftarrow \mathrm{D}(\bar{\boldsymbol{z}})$
    Update E, D with L_M (Eq. (5)), $\mathcal{C}$ with $\bar{\boldsymbol{z}}$ ▷ *self-replay*
  end for
  # Get final prediction
  $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{J} \leftarrow \mathrm{F}(\mathbf{I})$
  Soft reset F by Eq. (6) ▷ *soft reset*
end while

Unsupervised Clustering for Motion Discretization

We perform unsupervised clustering on the latent space of the motion denoising network M, yielding a codebook $\mathcal{C}$ of discrete latent motion representations. Inspired by recent work on motion generation (Guo et al. 2024), our $\mathcal{C} = \{\mathbf{C}^i \in \mathbb{R}^{N_c\times d} | i = 1, ..., k\}$ adopts a residual design

with k layers, with each layer having N_c codes of dimension d. The anchor motion set is then obtained by decoding codebook codes, which capture diverse yet regular human movements spanning the motion space.

We construct $\mathcal{C}$ during the pre-training of M. While the encoder E and decoder D are optimized as a standard denoising autoencoder, $\mathcal{C}$ is updated using the exponential moving average with the encoded latent $\boldsymbol{z} \in \mathbb{R}^d$ of the input motion. Specifically, we initialize the residual $\boldsymbol{r}_1 = \boldsymbol{z}$. For each layer i, we recursively retrieve the nearest code $\boldsymbol{c}_i \in \mathbf{C}^i$ and update the residual $\boldsymbol{r}_{i+1}$:

$$\boldsymbol{c}_i = \underset{\boldsymbol{c} \in \mathbf{C}^i}{\arg\min} ||\boldsymbol{c} - \boldsymbol{r}_i||_2, \boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \boldsymbol{c}_i. \quad (1)$$

The selected codes $\boldsymbol{c}_i$ are then updated using the corresponding $\boldsymbol{r}_i$ via exponential moving average.

Regularize with Motion Discretization to Adapt Pose Estimator

During test time, we obtain anchor motion and use it to regularize the adaptation of F. Denoting the output of F over t frames as $\boldsymbol{\theta}_{1:t}$, we encode it into the latent $\boldsymbol{z} = \mathrm{E}(\boldsymbol{\theta}_{1:t})$. $\boldsymbol{z}$ is then quantized to obtain $\boldsymbol{c} = \sum_{i=1}^{k} \boldsymbol{c}_i$. Each $\boldsymbol{c}_i$ is retrieved by Eq. (1), utilizing the latent regularity in measuring motion similarity. Both original latent $\boldsymbol{z}$ and quantized code $\boldsymbol{c}$ are fed in parallel to the decoder, yielding denoised motion $\boldsymbol{\theta}'_{1:t} = \mathrm{D}(\boldsymbol{z})$ and anchor motion $\boldsymbol{\theta}^*_{1:t} = \mathrm{D}(\boldsymbol{c})$.

The anchor motion then supervises the update of F in the next cycle, with the anchor loss defined as

$$L_{ach} = ||\boldsymbol{\theta} - \mathtt{sg}(\boldsymbol{\theta}^*)||, \quad (2)$$

where $\mathtt{sg}(.)$ denotes the stop-gradient operator. Similar to human motion generation works (Lucas et al. 2022; Zhang et al. 2023; Guo et al. 2024) that predict discrete latent motion codes for valid and realistic generation, we empirically find a comparable effect. While $\boldsymbol{\theta}'_{1:t}$ can be corrupted by self-supervision with noisy estimations (see Eq. (5) and Fig. 2), the discretized anchor motion $\boldsymbol{\theta}^*_{1:t}$ could filter out high-frequency noise while retaining core coherent motion patterns, thus regularizing the network update to facilitate long-term adaptation.

The overall loss for adapting F integrates L_{ach} with other signals introduced in CycleAdapt (Nam et al. 2023), *i.e.*, the detected 2D keypoints, the temporally averaged shape $\boldsymbol{\beta}'$, and the denoised motion $\boldsymbol{\theta}'_{1:t}$:

$$L_F = L_p + \lambda_1 L_s + \lambda_2 L_{2D} + \lambda_3 L_{ach}, \quad (3)$$

where $L_p = ||\boldsymbol{\theta} - \mathtt{sg}(\boldsymbol{\theta}')||, L_s = ||\boldsymbol{\beta} - \mathtt{sg}(\boldsymbol{\beta}')||$, and L_{2D} is the reprojection error that compares the 2D projections of the estimations with the 2D keypoints obtained from off-the-shelf detectors. $\lambda_1, \lambda_2, \lambda_3$ are weighting parameters.

Self-Replay with Motion Discretization to Adapt Motion Denoising Network

The online adaptation of M can suffer from representation drift, a phenomenon where latent codes gradually lose their ability to be decoded into consistent and regular anchor motions as adaptation progresses (*cf*. AQM (Caccia et al. 2020, Fig. 1)). As inspired by AQM (Caccia et al. 2020), our motion discretization further enables a self-replay mechanism to address this challenge.

To implement this, we construct a batch of replay motions during adaptation. As shown in Fig. 2-top right, we first sample a random code $\bar{\boldsymbol{c}} = \sum_{i=1}^{k} \bar{\boldsymbol{c}}_i$ from the pre-trained $\bar{\mathcal{C}}$, and then decode it with the pre-trained decoder $\bar{\mathrm{D}}$ to obtain the replay motion

$$\bar{\boldsymbol{\theta}}_{1:t} = \bar{\mathrm{D}}(\bar{\boldsymbol{c}}). \quad (4)$$

We then feed both replayed motions $\bar{\boldsymbol{\theta}}_{1:t}$ and test-time estimations $\boldsymbol{\theta}_{1:t}$ into M in parallel, with input frames randomly masked to encourage more robust adaptation. M is then updated in a self-supervised way, with the loss defined as

$$L_M = ||\bar{\boldsymbol{\theta}}'_{1:t} - \mathtt{sg}(\bar{\boldsymbol{\theta}}_{1:t})|| + ||\boldsymbol{\theta}'_{1:t} - \mathtt{sg}(\boldsymbol{\theta}_{1:t})||, \quad (5)$$

where $\bar{\boldsymbol{\theta}}'_{1:t}, \boldsymbol{\theta}'_{1:t}$ are the decoded outputs from the encoded latents $\bar{\boldsymbol{z}}, \boldsymbol{z}$ of the replay and test-time motions. We further update the codebook $\mathcal{C}$ using the replay latent $\bar{\boldsymbol{z}}$ via an exponential moving average with decay $\mu_{\mathcal{C}} = 0.999$, facilitating the synchronization of $\mathcal{C}$ with the evolving latent space.

In this way, we efficiently distill the pre-trained regularity without accessing the original training samples, thus addressing privacy concerns while enabling consistent and regular anchor motions (see Sec. 4 and Fig. 5) to aid in supervising the pose estimator F.

Soft Reset for the Pose Estimator

After adapting to each batch $\mathcal{V}$, we further perform a soft reset for the pose estimator F. Specifically, let F_{pre} and F respectively denote the weights before and after adapting on $\mathcal{V}$. We then set the weights of the pose estimator for subsequent adaptations as follows:

$$\mathrm{F} \leftarrow \mu_{\mathrm{F}} \mathrm{F}_{pre} + (1 - \mu_{\mathrm{F}}) \mathrm{F}, \quad (6)$$

where μ_{F} is the decay factor. This soft reset, implemented as an exponential moving average during adaptation, reduces the impact of noisy updates on individual batches while retaining critical traits learned from historical adaptation.

Implementation Details

We include the network architecture and pre-training details in the appendix. We adopt the same pre-trained pose estimator F from CycleAdapt (Nam et al. 2023), which builds on a ResNet-50 (He et al. 2016) backbone. Our M processes inputs and outputs consisting of $t = 16$ frames at 15 fps, with a latent dimension of $d = 512$. The residual codebook $\mathcal{C}$ has $k = 3$ layers, each with $N_c = 512$ codes.

During test time, we segment $\mathcal{S}$ into sequential batches $\mathcal{V}$, each containing 160 frames at 30 fps. Most of the hyperparameters are adopted from the implementation of CycleAdapt (Nam et al. 2023). For each $\mathcal{V}$, we use the Adam optimizer (Kingma 2014) with an initial learning rate of 5×10^{-5}, which is further reduced to 1×10^{-6} using a cosine annealing (Loshchilov and Hutter 2016). The number of cycles is set to $c = 12$, with a mini-batch size of 32 for both F, M within each cycle. For F, we set $\lambda_1 = 0.001, \lambda_2 = 0.1, \lambda_3 = 0.3$ to balance different loss terms, and set $\mu_F = 0.95$ for our soft-reset strategy. When adapting M, we use a mini-batch size of 4 for replay motion.

	Ego-Exo4D								3DPW		
	Basketball		Soccer		Dance		All Scenarios				
	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPJPE	MPJPE-PA	MPVPE
Pre-trained F	215.8	122.7	198.6	114.7	198.1	111.0	205.8	116.5	230.3	123.4	253.4
w/ OpenPose (Cao et al. 2017)											
BOA (Guan et al. 2021)†	158.4	72.8	115.3	63.3	117.8	69.3	135.1	70.0	98.2	55.8	114.2
DynaBOA (Guan et al. 2022)†	173.0	73.4	129.1	68.2	141.0	70.9	153.3	71.6	139.7	63.8	155.1
CycleAdapt (Nam et al. 2023)	160.4	80.8	126.3	78.5	135.3	81.0	145.0	80.5	141.0	79.6	155.6
Ours	**141.6**	**71.3**	**105.3**	**61.3**	**106.4**	**67.0**	**121.5**	**68.1**	**83.9**	**51.6**	**100.3**
w/ ViTPose (Xu et al. 2022b,a)											
BOA (Guan et al. 2021)†	149.6	65.3	108.9	58.2	101.9	61.9	123.5	62.9	91.5	53.9	105.5
DynaBOA (Guan et al. 2022)†	154.8	64.3	114.8	58.6	108.5	61.3	129.4	62.2	118.4	56.1	134.5
CycleAdapt (Nam et al. 2023)	144.8	69.6	112.1	62.9	108.3	67.1	124.6	67.6	109.6	66.6	124.8
Ours	**137.4**	**63.4**	**102.2**	**56.8**	**99.8**	**59.5**	**116.4**	**60.8**	**85.0**	**53.3**	**100.4**

Table 1: Comparison with the pre-trained pose estimator and existing online test-time adaptation methods on Ego-Exo4D and 3DPW. † denotes leveraging original pre-trained data in adaptation.

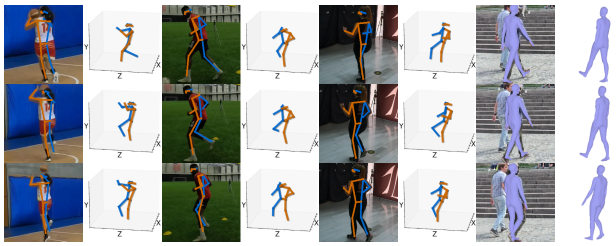

Figure 3: Our qualitative results. We show both 2D projections and 3D estimations in camera space.

	Ego-Exo4D		3DPW		
	MPJPE	MPJPE -PA	MPJPE	MPJPE -PA	MPVPE
Pre-trained F	205.8	116.5	230.3	123.4	253.4
SPIN (Kolotouros et al. 2019)	137.3	71.5	96.9	59.2	116.4
I2L-Mesh (Moon and Lee 2020)	136.6	73.8	93.2	57.7	110.1
PyMAF (Zhang et al. 2021)	135.5	70.2	92.8	58.9	110.1
HMR-2.0b (Goel et al. 2023) †	125.2	65.6	**81.3**	54.3	**94.4**
VIBE (Kocabas, Athanasiou, and Black 2020)	130.6	70.1	93.5	56.5	113.4
TCMR (Choi et al. 2021)	126.4	66.9	95.0	55.8	111.3
GLoT (Shen et al. 2023)	127.1	66.7	89.9	53.5	107.8
Ours (w/ OpenPose)	121.5	68.1	83.9	**51.6**	100.3
Ours (w/ ViTPose)	**116.4**	**60.8**	85.0	53.3	100.4

Table 2: Comparison with domain generalization methods that leverage more training datasets. All compared methods do not leverage Ego-Exo4D and 3DPW for pre-training. † indicates using a ViT-H/16 backbone.

4 Experiment

Dataset and Evaluation Protocol

Human3.6M is a large-scale dataset collected in controlled indoor settings (Ionescu et al. 2013). We follow previous test-time adaptation approaches (Guan et al. 2021, 2022; Nam et al. 2023) to pre-train ours on this dataset.

Ego-Exo4D has skilled human activities collected from cities worldwide (Grauman et al. 2024). We focus on basketball, soccer, and dancing scenarios, which present diverse body motion dynamics and action repetitions for capturing personal traits. Adapting to this dataset poses challenges due to the domain gap introduced by fisheye camera recordings and various motions. For evaluation, we select 30 participants and utilize their undistorted exocentric videos. The total recording durations of each participant range from 10 to 50 minutes. Detailed statistics are provided in the appendix.

We report the average performance over participants by calculating the **Mean Per Joint Position Error (MPJPE)** in the root-aligned space and the **Procrustes-Aligned MPJPE (MPJPE-PA)** in millimeters (mm), across 17 keypoints defined in MS COCO format (Lin et al. 2014). We did not evaluate mesh reconstruction due to the lack of annotations.

3DPW is an in-the-wild dataset collected with a moving hand-held camera, which has recordings with various appearances and backgrounds for each participant (Von Marcard et al. 2018). We follow previous test-time adaptation studies (Guan et al. 2021, 2022; Nam et al. 2023) to evaluate on its test split, which includes 5 participants with a total number of 35,515 frames at 30 fps. For adaptation, we set the personalized test domain. We evaluate the estimated joints with **MPJPE** and **MPJPE-PA**, and further compare the output mesh with the ground truth to report the **Mean Per Vertex Position Error (MPVPE)** in mm.

Comparison with Related Works

Discussion on Online Test-time Adaptation We first focus on comparison with existing online test-time adaptation research that has publicly available code, including BOA (Guan et al. 2021), DynaBOA (Guan et al. 2022), and CycleAdapt (Nam et al. 2023). All compared methods are pre-trained on the Human3.6M dataset, employ a ResNet-50 backbone for the image-based pose estimator, adapt to the same video stream, and utilize the same 2D detections obtained by either OpenPose (Cao et al. 2017) or ViTPose (Xu et al. 2022a,b), ensuring a fair comparison. We also include the pre-trained pose estimator F as a baseline, which is directly applied to the test data using its pre-trained weights.

In Tab. 1, we respectively report quantitative comparisons on Ego-Exo4D and 3DPW. We show our qualitative results in Fig. 3, and provide qualitative comparisons with previous works in the appendix. Together, these results demonstrate our superior performance compared to existing online test-time adaptation methods. This is achieved because our designs mitigate challenges in long-term adaptation, allowing us to effectively exploit personal traits for robust estimation. Furthermore, we notice that using 2D detections from ViT-

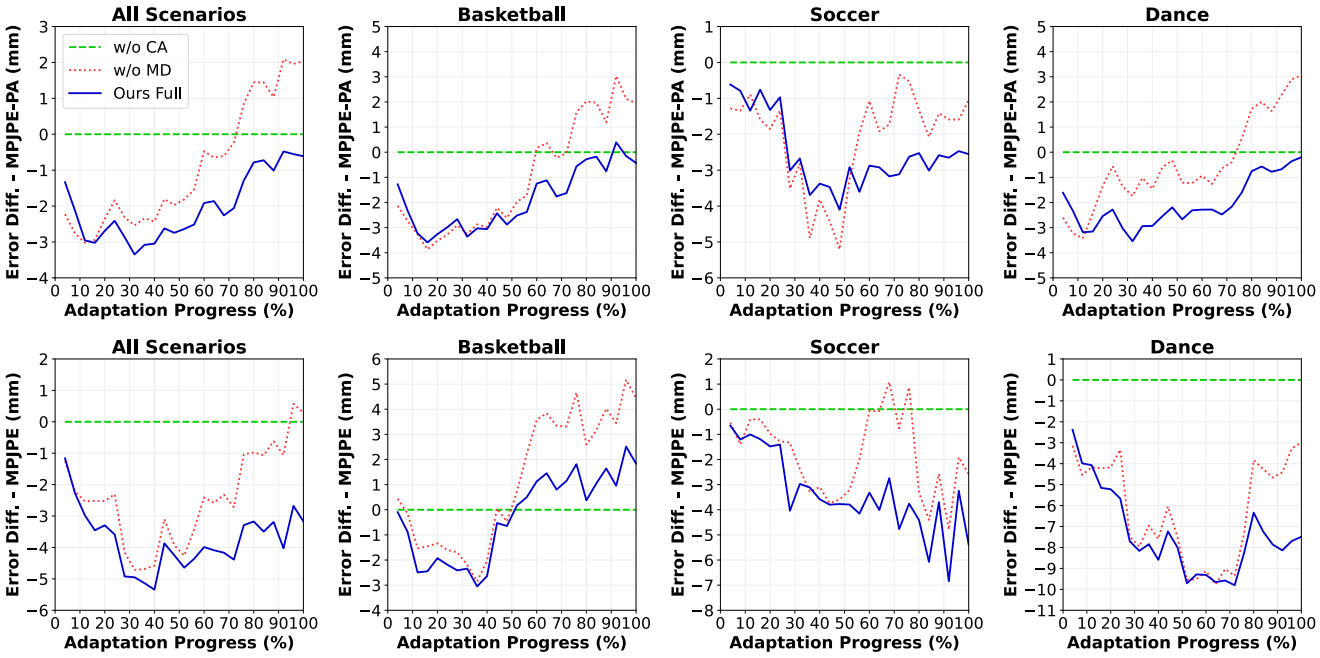

Figure 4: Error difference versus adaptation progress over time, with adaptation progress expressed relative to the total recording length for each participant. We plot the error difference relative to the baseline *w/o continuous adaptation* (dashed), which always starts from pre-trained weights for each $\mathcal{V}$. Our complete solution (solid) is compared against the variant of *w/o motion discretization* (dotted), which removes both L_{ach} and self-replay. A lower y-axis value indicates better performance.

Pose can enhance our performance on Ego-Exo4D but offers limited improvement on 3DPW. We attribute this to the observation that 3DPW often contains overlapping participants, which ViTPose struggles to separate effectively.

In comparison, we observe that DynaBOA and CycleAdapt suffer from severe error accumulation under self-supervision with imperfect estimations, resulting in decreased accuracy through continuous adaptation. Furthermore, BOA and DynaBOA report larger errors, especially in terms of MPJPE and MPVPE. This is primarily due to their heavy reliance on erroneous 2D supervision without adequate 3D guidance, leading to inaccurate depth estimation, as aligned with the discussion in (Nam et al. 2023).

We also discuss the run-time cost in the appendix, demonstrating competitive efficiency offered by our proposed solution. In contrast, BOA and DynaBOA require storing pre-trained images for exemplar guidance, which increases storage demands and run-time due to the exemplar retrieval.

Discussion with Domain Generalization Methods In Tab. 2, we conduct an empirical comparison with domain generalization methods that directly apply the pre-trained model to the test domain. These baselines pre-train on extensive datasets beyond Human3.6M, where HMR-2.0 (Goel et al. 2023) further employs more powerful ViT backbones.

Compared to the reported baselines that utilize a ResNet-50 backbone, our solution demonstrates better performance when using ViTPose detections. Additionally, with OpenPose detections, our solution outperforms these baselines on the 3DPW dataset and achieves a lower MPJPE on the Ego-Exo4D. Compared to HMR-2.0, our solution shows better accuracy on Ego-Exo4D when using 2D detections obtained from ViTPose, and achieves competitive results on 3DPW. These comparisons highlight the challenge of applying a pre-trained pose estimator to test domains that differ significantly from the pre-trained ones, while demonstrating the effectiveness of our solution in bridging this train-test domain gap. We further refer the readers to the appendix for a more detailed discussion.

w/ Soft-Reset on F	w/ Motion Discretization		MPJPE	MPJPE-PA
	w/ L_{ach} (θ^*)	w/ Self-Replay		
			144.3	80.6
✓			122.9	69.2
	✓	✓	138.2	74.8
✓	✓	✓	**121.5**	**68.1**
✓			122.9	69.2
✓	✓		123.2	69.2
✓		✓	122.9	69.2
✓	✓	✓	**121.5**	**68.1**

Table 3: Ablation study and discussion of key components on Ego-Exo4D. In Fig. 4, we further demonstrate that motion discretization benefits long-term adaptation.

Ablation Study and Discussions

We further examine our key designs on Ego-Exo4D, where participants have long recordings. For all compared methods, we use 2D detections obtained from OpenPose. Due to limited space, we report the overall results across all participants here and provide per-scenario results in the appendix.

w/ L_p (θ')	w/ L_{ach} (θ^*)	MPJPE	MPJPE-PA
✓		122.9	69.2
	✓	123.1	69.8
✓	✓	**121.5**	**68.1**

Soft-Reset Decay μ_F	MPJPE	MPJPE-PA
0	138.2	74.8
0.9	122.7	69.2
0.95	**121.5**	**68.1**
1.00	125.3	70.2

w/ Cont. Adapt.	MPJPE	MPJPE-PA
	125.3	70.2
✓	**121.5**	**68.1**

Table 4: Analysis of 3D motion supervision signals, soft-reset decay and continuous adaptation on Ego-Exo4D.

Discussion on Motion Discretization In Fig. 4 and Tab. 3, we demonstrate the benefits of motion discretization in long-term adaptation. We further find that leveraging motion discretization for both self-replay and anchor loss (*i.e.*, L_{ach} in Eq. (2)) synergistically achieves optimal performance, as reported in the lower part of Tab. 3. This is achieved because our self-replay addresses representation drift to enable decoding regular and realistic anchor motion throughout adaptation. These anchor motions could then provide faithful references to mitigate the negative influence of self-supervision with imperfect estimations. Conversely, using L_{ach} alone increases error due to noisy anchors (see Fig. 5).

In Tab. 4-left, we further examine the 3D motion supervision signal in the adaptation of F. In addition to the anchor motions $\boldsymbol{\theta}^*$ derived from the codebook, we observe that the denoised motion $\boldsymbol{\theta}'$, obtained by decoding the encoded latent $\boldsymbol{z}$, is also beneficial (*i.e.*, L_p in Eq. (3)). While the anchor motions $\boldsymbol{\theta}^*$ better preserve regular patterns under long-term self-supervision, $\boldsymbol{\theta}'$ captures fine-grained details that may be lost in discretization. Consequently, their complementary roles enable robust and accurate estimation.

Moreover, in the appendix, we demonstrate the benefits of using a residual codebook and updating it with replayed latent $\bar{\boldsymbol{z}}$ during adaptation. We also verify the effectiveness of our motion discretization by comparing it to an alternative solution that soft resets M.

Discussion on Soft Reset In Tab. 3-upper, we demonstrate the benefits of the proposed soft reset, where we further discuss the decay μ_F in Tab. 4-middle. Our soft reset enhances the performance by leveraging test-time traits captured from historical adaptation while mitigating the effects of noisy adaptation on individual batches. This is evident when compared with two baselines: $\mu_F = 0$, which updates solely based on weights from previous adaptation, and $\mu_F = 1$, which always resets F to its pre-trained weights after adaptation on each batch $\mathcal{V}$. Moreover, we empirically find that a conservative choice of $\mu_F = 0.95$ is beneficial.

Discussion on Continuous Personalized Adaptation As reported in Fig. 4 and Tab. 4-right, we achieve enhanced performance on average by exploiting personalized traits through continuous adaptation. Here, we set the baseline as resetting the model to its pre-trained weights before adapting to each incoming batch $\mathcal{V}$.

Specifically, as shown in Fig. 4, the performance gain from continuous adaptation is evident in soccer and dancing scenarios. However, we observe limited benefits for long-term adaptation in basketball scenarios, as some participants are often severely distorted in videos due to the camera setups. This further enlarges domain gaps and leads to significant batch-wise errors, presenting difficulties for continuous adaptation. Detailed analysis is provided in the appendix.

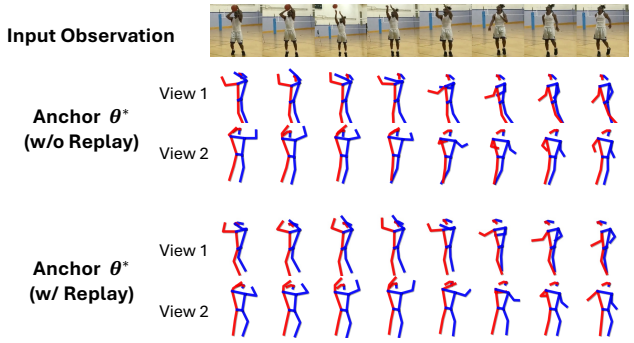

Figure 5: Visualization of the retrieved anchor $\boldsymbol{\theta}^*$ for the input video depicting basketball shooting, after adapting over 40 mins of observation. Our self-replay mechanism facilitates decoding of realistic and regular anchor motions (*e.g.* leg pose) throughout the adaptation process.

5 Conclusion

We propose a novel solution to address the challenges inherent in long-term online adaptation with self-supervision, which leverages motion discretization obtained through unsupervised clustering and incorporates an efficient soft-reset mechanism. Specifically, our motion discretization not only regularizes the adaptation process but also supports a self-replay mechanism. We further examine online test-time adaptation on a personalized domain, where our solution enables harnessing the benefits of continuous personalized adaptation and outperforms existing online test-time adaptation methods. Evaluations demonstrate our effectiveness and verify our key designs.

Limitations and Future Work We adopt a pose estimator with ResNet-50 backbones to output poses in local frames, aligning with prior works to illustrate the error accumulation challenge and demonstrate the effectiveness of our solution. Nonetheless, we acknowledge recent advancements with ViT backbones (Goel et al. 2023) and global human trajectory recovery from moving cameras (Shin et al. 2024). We also notice advanced techniques that could better exploit test-time traits while keeping stable adaptation, such as adaptive parameters for loss weights or EMA schedules. We consider integrating these advancements as potential extensions to our solution. Additionally, we recognize our current limitations in handling distorted appearances, such as those observed in the Ego-Exo4D basketball participants.

Acknowledgements

This research is supported by JSPS KAKENHI Grant Number JP25K03134, Toyota Foundation Grant Number D24-ST-0030, and The Telecommunications Advancement Foundation.

References

Caccia, L.; Belilovsky, E.; Caccia, M.; and Pineau, J. 2020. Online learned continual compression with adaptive quantization modules. In *International conference on machine learning*, 1240–1250. PMLR.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chai, W.; Jiang, Z.; Hwang, J.-N.; and Wang, G. 2023. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14655–14665.

Choi, H.; Moon, G.; Chang, J. Y.; and Lee, K. M. 2021. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1964–1973.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dwivedi, S. K.; Sun, Y.; Patel, P.; Feng, Y.; and Black, M. J. 2024. TokenHMR: Advancing Human Mesh Recovery with a Tokenized Pose Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1323–1333.

Geng, Z.; Wang, C.; Wei, Y.; Liu, Z.; Li, H.; and Hu, H. 2023. Human Pose as Compositional Tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14783–14794.

Gong, K.; Zhang, J.; and Feng, J. 2021. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8575–8584.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Guan, S.; Xu, J.; He, M. Z.; Wang, Y.; Ni, B.; and Yang, X. 2022. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 5070–5086.

Guan, S.; Xu, J.; Wang, Y.; Ni, B.; and Yang, X. 2021. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10472–10481.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339.

Joo, H.; Neverova, N.; and Vedaldi, A. 2021. Exemplar finetuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, 42–52. IEEE.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131.

Kanazawa, A.; Zhang, J. Y.; Felsen, P.; and Malik, J. 2019. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5614–5623.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5253–5263.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2252–2261.

Lin, Q.; Chen, R.; Gu, K.; and Yao, A. 2025a. Semantics-aware Test-time Adaptation for 3D Human Pose Estimation. *arXiv preprint arXiv:2502.10724*.

Lin, Q.; Gu, K.; Yang, L.; and Yao, A. 2025b. Online Test-time Adaptation for 3D Human Pose Estimation: A Practical Perspective with Estimated 2D Poses. *arXiv preprint arXiv:2503.11194*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6).

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lucas, T.; Baradel, F.; Weinzaepfel, P.; and Rogez, G. 2022. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, 417–435. Springer.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 752–768. Springer.

Mugaludi, R. R.; Kundu, J. N.; Jampani, V.; et al. 2021. Aligning silhouette topology for self-adaptive 3D human pose recovery. *Advances in Neural Information Processing Systems*, 34: 4582–4593.

Nam, H.; Jung, D. S.; Oh, Y.; and Lee, K. M. 2023. Cyclic test-time adaptation on monocular video for 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14829–14839.

Shen, X.; Yang, Z.; Wang, X.; Ma, J.; Zhou, C.; and Yang, Y. 2023. Global-to-local modeling for video-based 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8887–8896.

Shin, S.; Kim, J.; Halilaj, E.; and Black, M. J. 2024. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2070–2080.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Wei, W.-L.; Lin, J.-C.; Liu, T.-L.; and Liao, H.-Y. M. 2022. Capturing humans in motion: Temporal-attentive 3d human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13211–13220.

Weng, Z.; Wang, K.-C.; Kanazawa, A.; and Yeung, S. 2022. Domain adaptive 3d pose augmentation for in-the-wild human mesh recovery. In *2022 International Conference on 3D Vision (3DV)*, 261–270. IEEE.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022a. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems*.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022b. ViTPose+: Vision Transformer Foundation Model for Generic Body Pose Estimation. *arXiv preprint arXiv:2212.04246*.

You, Y.; Liu, H.; Wang, T.; Li, W.; Ding, R.; and Li, X. 2023. Co-evolution of pose and mesh for 3d human body estimation from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14963–14973.

Zhang, H.; Tian, Y.; Zhou, X.; Ouyang, W.; Liu, Y.; Wang, L.; and Sun, Z. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11446–11456.

Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14730–14740.