

Efficient Diffusion-Based 3D Human Pose Estimation with Hierarchical Temporal Pruning

Yuquan Bi, Hongsong Wang, Xinli Shi, *Senior Member, IEEE*, Zhipeng Gui, *Member, IEEE*, Jie Gui, *Senior Member, IEEE*, and Yuan Yan Tang, *Life Fellow, IEEE*

Abstract—Diffusion models have demonstrated strong capabilities in generating high-fidelity 3D human poses, yet their iterative nature and multi-hypothesis requirements incur substantial computational cost. In this paper, we propose an efficient diffusion-based 3D human pose estimation framework with a Hierarchical Temporal Pruning (HTP) strategy, which dynamically prunes redundant pose tokens across both frame and semantic levels while preserving critical motion dynamics. HTP operates in a staged, top-down manner: (1) Temporal Correlation-Enhanced Pruning (TCEP) identifies essential frames by analyzing inter-frame motion correlations through adaptive temporal graph construction; (2) Sparse-Focused Temporal MHSA (SFT MHSA) leverages the resulting frame-level sparsity to reduce attention computation, focusing on motion-relevant tokens; and (3) Mask-Guided Pose Token Pruner (MGPTP) performs fine-grained semantic pruning via clustering, retaining only the most informative pose tokens. Experiments on Human3.6M and MPI-INF-3DHP show that HTP reduces training MACs by 38.5%, inference MACs by 56.8%, and improves inference speed by an average of 81.1% compared to prior diffusion-based methods, while achieving state-of-the-art performance.

Index Terms—3D Human Pose Estimation, Diffusion Models, Temporal Pruning, Transformer.

I. INTRODUCTION

3D human pose estimation (HPE) from monocular videos is a fundamental task advancing rapidly in recent years for its significant applications in action recognition [1]–[5], human-robot interaction [6]–[8], and virtual reality [9]–[11]. Benefiting from the excellent performance of 2D pose detectors [12]–[17], the 2D-to-3D lifting pipeline [18]–[25] has become dominant due to its high precision and lightweight nature.

Corresponding author: Jie Gui and Hongsong Wang.

Yuquan Bi is with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China (e-mail: yuquanbi@seu.edu.cn).

H. Wang is with School of Computer Science and Engineering, Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Southeast University, Nanjing 210096, China (e-mail: hongsongwang@seu.edu.cn).

Xinli Shi is with the National Center for Applied Mathematics, Southeast University, Nanjing 211189, China (e-mail: xinli.shi@seu.edu.cn).

Z. Gui is with the School of Remote Sensing and Information Engineering and the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zhipeng.gui@whu.edu.cn).

Jie Gui is with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China, also with Purple Mountain Laboratories, Nanjing 211111, China, and also with the Engineering Research Center of Blockchain Application, Supervision and Management (Southeast University), Ministry of Education, Nanjing 210000, China (e-mail: guijie@seu.edu.cn).

Yuan Yan Tang is with the Department of Computer and Information Science, University of Macau, Macau, China, and also with Faculty of Science and Technology, UOW College Hong Kong, Hong Kong, China (e-mail: yytang@um.edu.mo).

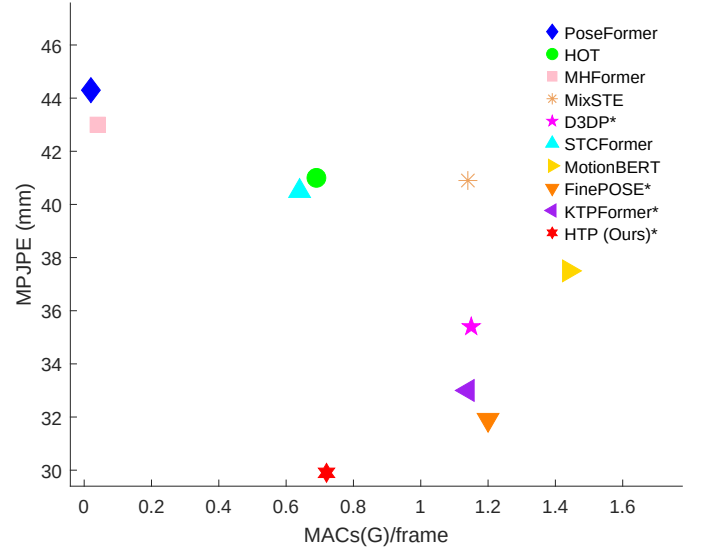


Fig. 1. MACs and MPJPE of different methods on the Human3.6M dataset. We achieve the best performance while demonstrating highly competitive MACs results. * indicates diffusion-based methods.

The 2D-to-3D lifting pipeline lacks depth priors and suffers from ambiguity. To mitigate this issue, recent works incorporate temporal correlations across video frames into the pose reconstruction process. For example, many transformer-based architectures [26]–[36] effectively capture long-range temporal dependencies by encoding the joint-level semantics of each video frame into pose tokens, achieving promising performance even on extremely long video sequences. However, the computational cost for spatial-temporal modeling in self-attention (SA) increases quadratically as the number of frames increases, resulting in substantial computational overhead.

Diffusion-based 3D HPE leverages transformer architectures to resolve depth ambiguity through the iterative refinement of high-fidelity 3D pose generation. These methods employ a transformer-based diffusion process, which requires K steps of iterative refinement to generate H pose hypotheses during inference. However, the inherent computational complexity of diffusion models, combined with the transformer-based SA mechanisms, leads to significant resource demands. For instance, processing a 243-frame video sequence with D3DP [37] requires 1.15G MACs per frame during training, but this increases to 228.8G per frame during inference with $H = 20$ and $K = 10$. Although adjusting H and K offers

a theoretical trade-off between accuracy and efficiency, the combined cost of diffusion steps and transformer operations makes it challenging to achieve both simultaneously.

A straightforward approach to reduce the primary computational bottleneck in 3D HPE is to eliminate redundant pose tokens in the temporal SA calculations. Existing methods typically adopt two disjoint strategies: (1) Frame-level pruning that discards adjacent frames under static redundancy assumptions, and (2) Semantic-level sparsification that clusters low-information tokens via heuristic criteria. While these single-stage strategies effectively reduce computation, they often overlook subtle yet crucial motion transitions. More importantly, such approaches are not well-suited for diffusion-based 3D HPE, where pose reconstruction unfolds iteratively across multiple noise levels. Simply applying naive pruning risks discarding informative content at intermediate steps, thereby compromising motion continuity and stability. To address these challenges, we propose an efficient diffusion-based 3D human pose estimation framework with Hierarchical Temporal Pruning (HTP), which operates across both frame-level and semantic-level stages to preserve essential motion dynamics throughout the denoising process. By selectively retaining key frames and salient pose tokens at each denoising iteration, HTP maintains the integrity of global motion patterns while reducing computational cost.

Specifically, we implement HTP through a structured, hierarchical pruning framework across both frame and semantic levels. First, the Temporal Correlation-Enhanced Pruning (TCEP) module analyzes temporal correlations across video frames. Each node represents a video frame, and we compute a dense correlation matrix to measure inter-frame similarity. Based on this, our Correlation-Enhanced Node Selection Algorithm constructs a dynamic temporal graph and selects a subset of nodes with strong temporal relevance as representative frames. A Sparse Binary Mask $\mathbf{M}$ is generated to store the retained temporal relationships. Second, based on the temporal correlations identified by TCEP, the Sparse-Focused Temporal Multi-Head Self-Attention (SFT MHSA) uses $\mathbf{M}$ to guide attention toward motion-relevant pose tokens. By restricting attention to key frames, SFT MHSA reduces computational overhead while preserving the model's ability to capture global temporal dependencies. Finally, we apply the Mask-Guided Pose Token Pruner (MGPTP), which integrates frame-level correlations from TCEP and sparse pose token representations from SFT MHSA. MGPTP discards redundant pose tokens while preserving tokens critical to motion fidelity using a density-aware strategy guided by the sparse mask $\mathbf{M}$. Together, these modules form a cohesive hierarchical denoising framework that enhances computational efficiency and preserves motion fidelity in diffusion-based 3D HPE. As shown in Fig. 1, HTP reduces training MACs by an average of 38.5%, setting a new standard in efficient diffusion-based 3D HPE. The contributions of this paper are as follows:

- We propose Hierarchical Temporal Pruning (HTP), a unified hierarchical pruning framework integrated into diffusion-based 3D HPE that reduces both frame- and token-level redundancy to improve efficiency, overcoming the limitations of previous single-stage strategies.

- TCEP, SFT MHSA, and MGPTP operate under a unified sparse constraint $\mathbf{M}$ to collaboratively reduce temporal redundancy and preserve motion-critical dynamics. All modules are plug-and-play and compatible with both diffusion- and transformer-based 3D HPE pipelines.
- Extensive experiments on Human3.6M and MPI-INF-3DHP show that HTP achieves state-of-the-art accuracy while reducing training MACs by 38.5%, inference MACs by 56.8%, and boosting FPS by 81.1% on average.

II. RELATED WORK

A. Transformer-Based 3D HPE

The Transformer, first proposed by Vaswani et al. [38], has demonstrated remarkable performance in computer vision (CV) tasks [39]–[43], as the self-attention mechanism has a strong ability to capture long-range dependencies. This characteristic makes it particularly well-suited for 3D HPE. PoseFormer [27] was the first to adopt the vision transformer as a backbone network for video-based 3D HPE. MixSTE [32] alternates between spatial and temporal transformer blocks to capture spatio-temporal features, providing 3D pose estimates for each frame in the input sequence. DualFormer [36] further enhances performance via dual-path attention across joints and frames. GKONet [25] incorporates geometric priors into a graph-guided transformer for structure-aware prediction. STCFormer [29] reduces computational complexity by separately modeling spatial and temporal components of input joints. MotionBERT [30] introduces a dual-stream spatial-temporal transformer to model long-range spatial-temporal relationships, and is finetuned for skeletal joint-based tasks. However, the quadratic complexity of self-attention in spatial-temporal modeling results in substantial computational overhead.

B. Diffusion-Based 3D HPE

Diffusion models are a class of generative models that progressively degrade observed data by adding noise, and then restore the original data through a reverse denoising process. These models have demonstrated promising results across various applications, such as image/video generation [44]–[46], super-resolution [45], [47], and Human Motion Generation [48]–[51]. Recently, several 3D HPE methods based on diffusion models [37], [52]–[55] have been proposed to generate high-fidelity 3D human poses, aiming to address the challenge of intrinsic depth ambiguity. D3DP [37], [53] integrates a denoiser based on MixSTE [32] to reconstruct noisy 3D poses by assembling joint-by-joint multiple hypotheses. FinePose [54] learns modifiers for different human body parts to describe human movements at multiple levels of granularity. KTPFormer [55] incorporates the anatomical structure of the human body and joint motion trajectories across frames as prior knowledge to learn spatial and temporal correlations. However, the reliance on iterative refinement (K steps) and multiple hypotheses (H samples) significantly increases the computational burden, posing greater efficiency challenges compared to transformer-based counterparts. Distinct from prior diffusion-based approaches [37], [54], [55] that primarily focus on enhancing generation fidelity, HTP is explicitly

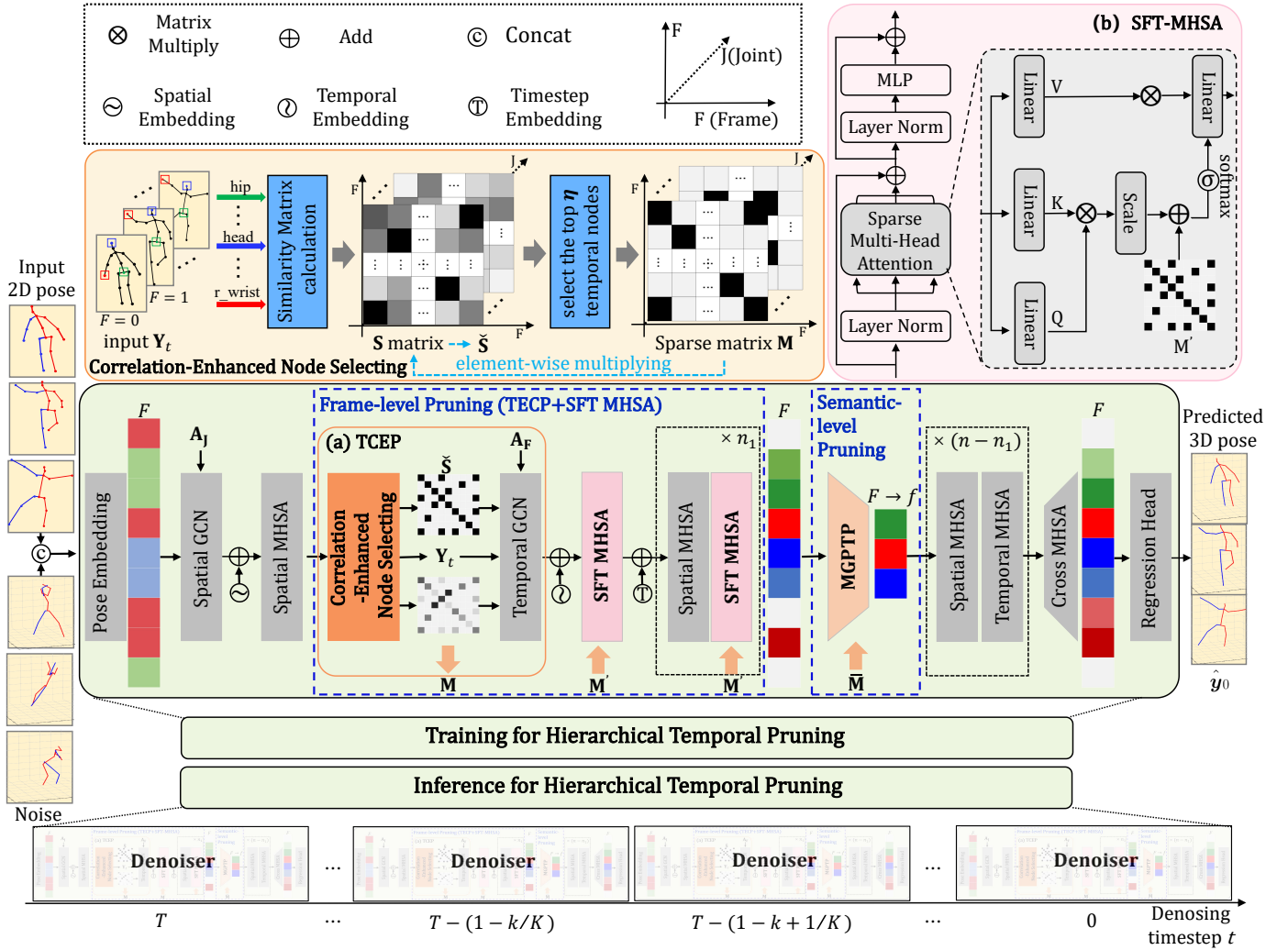


Fig. 2. **The architecture of the proposed HTP.** The framework is structured into two hierarchical pruning phases: (a) **Frame-level Pruning** and (b) **Semantic-level Pruning**. At each diffusion step, the input $[y_t \oplus x]$ is embedded and processed by the Spatial GCN [31] and Spatial MHSA. **In Phase (a)**, TCEP first infers a sparse temporal mask M and pruned features, which then guide the SFT MHSA in modeling frame-to-frame dependencies on the full sequence length F with reduced redundancy. **In Phase (b)**, MGPTP physically condenses the sequence from F to f by aggregating representative tokens. Finally, Cross MHSA restores the original length F for prediction. Note that M' and $\bar{M}$ denote variants of M adapted for SFT MHSA and MGPTP, respectively. Pose embedding, Spatial MHSA, and full Temporal MHSA are standard operations in 3D HPE.

designed to mitigate this computational bottleneck. By introducing a hierarchical temporal pruning strategy, it effectively optimizes the trade-off between efficiency and performance.

C. Improving Efficiency in 3D HPE

Enhancing computational efficiency in 3D human pose estimation is crucial for real-world applications, especially in resource-limited environments. While prior works have explored pruning strategies to improve temporal modeling efficiency, most adopt a single-level approach. Frame-level methods, such as DeciWatch [56], G-SFormer [57], and Uplift [58], reduce computational cost by sampling or scheduling representative frames based on temporal redundancy. Token-level methods, including HOT [59] and P-STMO [33], focus on semantic sparsification by removing low-saliency tokens or clustering similar features. While these single-strategy

approaches help reduce complexity, they often fail to account for interactions between temporal structure and semantic content. This can lead to suboptimal retention of motion-critical information, particularly in dynamic scenes or under iterative refinement processes like diffusion. In contrast, our approach integrates pruning at both frame and semantic levels within a unified framework. This hierarchical design enables more informed token selection and better preserves motion coherence throughout the denoising, providing a more robust and adaptive solution than methods relying on a single strategy.

III. PRELIMINARY OF DIFFUSION-BASED 3D HPE

Diffusion models are generative frameworks that characterize data distributions via a time-dependent diffusion process with two phases: (1) a *training phase*, where data is progressively perturbed by adding noise, and a denoiser is trained to

reverse this perturbation, and (2) *an inference phase*, where the trained denoiser reconstructs the original, uncorrupted data.

A. Training

Starting from the ground truth input 3D pose $\mathbf{y}_0$, a sequence of noisy samples $\{\mathbf{y}_t\}_{t=1}^T$ is generated, where T denotes the total number of timesteps. During this process, standard Gaussian noise is introduced to $\mathbf{y}_0$, progressively transforming it into a Gaussian distribution $\mathbf{y}_T \sim p_T$. Following DDPMs [60], the perturbation of $\mathbf{y}_t$ can be expressed as:

$$q(\mathbf{y}_t | \mathbf{y}_0) := \sqrt{\bar{\alpha}_t} \mathbf{y}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (1)$$

with $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. where $\{\beta_t\}_{t=1}^T$ denotes the variance schedule.

Subsequently, $\mathbf{y}_t$ is passed to the Denoiser conditioned on 2D keypoints $\mathbf{x}$ and timestep t to reconstruct the original 3D pose $\hat{\mathbf{y}}_0$. The entire framework is optimized through the standard MPJPE (Mean Per Joint Position Error) loss, which minimizes the average Euclidean distance between the predicted and ground-truth 3D joint positions:

$$\mathcal{L} = \frac{1}{J} \sum_{k=1}^J \|\mathbf{y}_0 - \hat{\mathbf{y}}_0\|_2. \quad (2)$$

B. Inference

During inference, the reverse diffusion process is applied to recover the original 3D pose $\mathbf{y}_0$ by iteratively denoising the noisy sample $\mathbf{y}_T \sim p_T$. Following D3DP [37], we adopt the multi-hypothesis strategy within the DDIM [61] framework, focusing on predicting the original input rather than the noise. The reverse process can be articulated as follows:

$$\mathbf{y}_{h,t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \hat{\mathbf{y}}_{h,0} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \hat{\epsilon}_t + \sigma_t \epsilon, \quad (3)$$

where t and $t-1$ are the adjacent timesteps in the subset $\tau \subset \{1, \dots, T\}$. $h \in \{0, \dots, H\}$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, and

$$\begin{aligned} \hat{\epsilon}_t &= (\mathbf{y}_{h,t} - \sqrt{\bar{\alpha}_t} \cdot \hat{\mathbf{y}}_{h,0}) / \sqrt{1 - \bar{\alpha}_t} \\ \sigma_t &= \sqrt{(1 - \bar{\alpha}_{t-1}) / (1 - \bar{\alpha}_t)} \cdot \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}}. \end{aligned} \quad (4)$$

Beginning by sampling H initial 3D poses $\mathbf{y}_{h,T}$ from a unit Gaussian and then feeding the samples into the denoiser to produce H viable 3D pose hypotheses $\mathbf{y}_{h,0}$. This process is repeated iteratively for K steps, with the timestep t updated as $T(1 - \frac{k}{K})$ at each iteration $k \in [1, K]$.

In line with other diffusion-based methods [37], [54], [55], we employ the Joint-Wise Reprojection-Based Multi-Hypothesis Aggregation (JPMA) technique [37] to aggregate and evaluate the final 3D pose predictions.

IV. METHOD

A. Overview

Following the formulation in Sec. III, our framework estimates the clean 3D pose $\mathbf{y}_0$ from its noisy observation $\mathbf{y}_t$ conditioned on 2D keypoints $\mathbf{x}$. As illustrated in Fig. 2, the input

Algorithm 1 Generation and Utilization of $\mathbf{M}$

Require: Input Pose Tokens $\mathbf{Y} \in \mathbb{R}^{J \times F \times D}$, Neighbor Count η
Ensure: $\mathbf{M} \in \mathbb{R}^{J \times F \times F}$, $\mathbf{M}' \in \mathbb{R}^{J \times F \times F}$, $\bar{\mathbf{M}} \in \mathbb{R}^{F \times F}$

- 1: **Phase 1: Joint-wise Mask Generation (Section IV-B)**
- 2: **Initialize** $\mathbf{M} \leftarrow \{0\}^{J \times F \times F}$
- 3: **for** each joint index $j \in \{1, \dots, J\}$ **do**
- 4: $\mathbf{Y}_t^{(j)} \leftarrow \mathbf{Y}_t[j, :, :] \in \mathbb{R}^{F \times D}$ ▷ Extract temporal features
- 5: // 1.1 Compute Pairwise Similarity (Eq. 6)
- 6: $\mathbf{S}^{(j)} \leftarrow (\mathbf{Y}_t^{(j)} \cdot (\mathbf{Y}_t^{(j)})^\top) / \sqrt{D}$
- 7: // 1.2 Exclude self-loops and select Top- η Neighbors,
- 8: $\mathbf{S}_{\text{masked}}^{(j)} \leftarrow \text{MaskDiagonal}(\mathbf{S}^{(j)}, -\infty)$
- 9: $\mathcal{I}_{\text{top}} \leftarrow \text{TopK}(\mathbf{S}_{\text{masked}}^{(j)}, k = \eta, \text{dim} = -1). \text{indices}$
- 10: // $\mathcal{I}_{\text{top}}[p]$ contains the set of neighbor indices $\{q\}$ for p
- 11: // 1.3 Construct Symmetric Connectivity (Eq. 7)
- 12: **for** each frame $p \in \{1, \dots, F\}$ **do**
- 13: $\mathbf{M}[j, p, p] \leftarrow 1$ ▷ Self-loop ($p = q$)
- 14: $\mathcal{Q}_p \leftarrow \mathcal{I}_{\text{top}}[p]$ ▷ Get neighbor indices q for frame p
- 15: $\mathbf{M}[j, p, \mathcal{Q}_p] \leftarrow 1$ ▷ Forward: $q \in \text{Top}_p(\mathbf{S}^{(j)}, \eta)$
- 16: $\mathbf{M}[j, \mathcal{Q}_p, p] \leftarrow 1$ ▷ Symmetric: $p \in \text{Top}_q(\mathbf{S}^{(j)}, \eta)$
- 17: **end for**
- 18: **end for**
- 19: // $\mathbf{M}$ is now the stacked binary mask $\in \{0, 1\}^{J \times F \times F}$
- 20: **Phase 2: Application in SFT MHSA (Section IV-C)**
- 21: $\mathbf{M}' \leftarrow \text{Where}(\mathbf{M} == 1, 0, -\infty)$
- 22: **Phase 3: Application in MGTP (Section IV-D)**
- 23: // Aggregate across joints
- 24: $\bar{\mathbf{M}} \leftarrow \text{Mean}(\mathbf{M}, \text{dim} = 0) \in \mathbb{R}^{F \times F}$

pair is first projected into a high-dimensional representation and processed by the Spatial GCN [31] and the Spatial MHSA to encode skeletal topology. The output $\mathbf{Y}_t$ then undergoes our core Hierarchical Temporal Pruning (HTP) strategy, which operates in a coarse-to-fine manner:

Frame-Level Pruning: This phase focuses on filtering redundancy while maintaining full temporal resolution (F). First, the *Temporal Correlation-Enhanced Pruning (TCEP)* module initiates the hierarchy by establishing a sparse topology to filter static frames. Subsequently, the diffusion timestep embedding $\mathcal{F}(t)$ is injected into the feature map. The *Sparse-Focused Temporal MHSA (SFT MHSA)* performs transitional refinement: it models long-range dependencies strictly within this sparse structure. By enhancing the feature discriminability of the retained frames, it acts as a semantic bridge, preparing the representation for the subsequent hard pruning.

Semantic-Level Pruning: Advancing the hierarchy, the *Mask-Guided Pose Token Pruner (MGTP)* executes "hard-pruning" by physically compressing the sequence length from F to f . It aggregates the refined tokens from SFT MHSA into high-level descriptors, achieving deep semantic abstraction.

The condensed sequence $\bar{\mathbf{Y}}_t$ is then processed by $n - n_1$ standard encoder blocks for deep refinement. Finally, a Cross MHSA restores the full temporal resolution (from f back to F) for the final prediction $\hat{\mathbf{y}}_0$. During inference, we follow the reverse process detailed in Sec. III, employing a deterministic DDIM [61] sampler to recover 3D poses from pure Gaussian noise. Specifically, we iteratively refine H initial hypotheses over K steps and subsequently aggregate these diverse predictions to ensure robust reconstruction.

TABLE I

QUANTITATIVE COMPARISON WITH THE SOTA METHODS ON THE HUMAN3.6M DATASET. F : THE NUMBER OF INPUT FRAMES. CE: ESTIMATING CENTER FRAME ONLY. DETECTOR: USING CPN [12] AND SH [13] AS THE 2D KEYPOINTS DETECTOR TO GENERATE THE INPUTS, OR USING THE GROUND TRUTH 2D KEYPOINTS AS INPUTS. † INDICATES THE SCRATCH SETTING IN [30], AND ‡ INDICATES THE FINETUNE SETTING IN [30]. HTP (OURS) UTILIZES THE DEFAULT D3DP BACKBONE; VARIANTS LIKE HTP w/ MIXSTE DEMONSTRATE PLUG-AND-PLAY GENERALIZATION. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE FORMATS.

Method	Type	Publication	F	CE	Human3.6M (DET)			Human3.6M (GT)		MACs (G)	Params (M)
					Detector	MPJPE $\downarrow$	P-MPJPE $\downarrow$	Detector	MPJPE $\downarrow$		
TCN [62]	CNN	CVPR'19	243	–	CPN	46.8	36.5	GT	37.8	–	–
GLA-GCN [63]	GCN	ICCV'23	243	✓	CPN	44.4	34.8	GT	21.0	–	–
FTCM [64]	MLP	TCSVT'24	351	✓	CPN	45.3	35.3	GT	28.2	1.5	4.7
PoseMamba-X [65]	Mamba	AAAI'25	243	×	CPN	37.1	31.5	GT	–	109.9	26.5
SAMA-L [66]	Mamba	ICCV'25	243	×	CPN	36.9	31.3	GT	–	53.2	17.3
PoseFormer [27]	Transformer	ICCV'21	81	✓	CPN	44.3	36.5	GT	31.3	1.6	9.6
P-STMO [33]	Transformer	ECCV'22	243	✓	CPN	42.8	34.4	GT	29.3	1.7	7.0
PoseFormerV2 [28]	Transformer	CVPR'23	243	✓	CPN	45.2	35.6	GT	35.5	2.1	14.4
MHFormer [67]	Transformer	CVPR'22	351	✓	CPN	43.0	34.4	GT	30.5	9.6	24.7
MixSTE [32]	Transformer	CVPR'22	243	×	CPN	40.9	30.6	GT	21.6	139.0	33.8
STCFormer [29]	Transformer	CVPR'23	243	×	CPN	40.5	31.8	GT	21.3	78.2	18.9
MotionBERT [30] †	Transformer	ICCV'23	243	×	SH	39.2	–	GT	17.8	174.8	42.5
MotionBERT [30] ‡	Transformer	ICCV'23	243	×	SH	37.5	–	GT	16.9	174.8	42.5
HOT [59]	Transformer	CVPR'24	243	×	CPN	41.0	–	GT	–	83.8	35.0
TC-MixSTE [34]	Transformer	TMM'24	243	×	CPN	39.9	31.9	GT	21.1	–	–
DualFormer [36]	Transformer	TCSVT'24	351	×	CPN	42.8	34.4	GT	28.9	–	–
HTP w/ MixSTE	Transformer	–	243	×	CPN	39.9	29.9	GT	20.7	87.6	36.4
HTP w/ MotionBERT †	Transformer	–	243	×	SH	38.9	–	GT	17.7	101.7	47.6
Diffpose [52]	Diffusion	CVPR'23	243	×	CPN	36.9	28.7	GT	18.9	–	–
Diffpose [68]	Diffusion	ICCV'23	64	×	–	42.9	30.8	GT	–	–	–
D3DP [37]	Diffusion	ICCV'23	243	×	CPN	35.4	28.7	GT	18.4	139.1	34.8
FinePOSE [54]	Diffusion	CVPR'24	243	×	CPN	<u>31.9</u>	<u>25.0</u>	GT	16.7	146	200.6
KTPFormer [55]	Diffusion	CVPR'24	243	×	CPN	33.0	26.2	GT	18.1	139.1	36.3
HTP (Ours)	Diffusion	–	243	×	CPN	29.9	23.3	GT	16.7	87.7	37.5

B. Temporal Correlation-Enhanced Pruning (TCEP)

Learning fine-grained temporal correlations is crucial for accurate 3D human pose estimation, particularly in diffusion-based frameworks where iterative denoising must preserve subtle motion cues. However, constructing dense temporal connections for each pose token risks computational inefficiency and semantic drift. To address this within the *Frame-level Pruning* phase, we design the TCEP module to establish the structural foundation of our hierarchy. By explicitly modeling sparse temporal dependencies, TCEP constructs a preliminary sparse topology by dynamically selecting semantically correlated frames for each joint while filtering out irrelevant connections, thereby ensuring that subsequent feature learning is strictly focused on motion-critical regions.

Given the input $\mathbf{Y}_t$, TCEP first overlays the learnable global temporal topology matrix $\hat{\mathbf{A}}_F \in \mathbb{R}^{F \times F}$ onto the original adjacency matrix $\mathbf{A}_F \in \mathbb{R}^{F \times F}$, as outlined by [31], [55]. This combination can be expressed as:

$$\mathbf{A}_T = \frac{(\mathbf{A}_F + \hat{\mathbf{A}}_F) + (\mathbf{A}_F + \hat{\mathbf{A}}_F)'}{2}, \quad (5)$$

where $'$ denotes the matrix transpose to ensure symmetry. As illustrated in Fig. 2 top left, we apply a Correlation-Enhanced Node Selecting algorithm to dynamically select temporal correlation nodes and refine their connections. For

each joint $j \in \{1, \dots, J\}$, $\mathbf{Y}_t^{(j)} \in \mathbb{R}^{F \times D}$ denotes its temporal token sequence. We compute a scaled similarity matrix:

$$\mathbf{S}^{(j)} = \frac{\mathbf{Y}_t^{(j)} (\mathbf{Y}_t^{(j)})^\top}{\sqrt{D}} \in \mathbb{R}^{F \times F}, \quad (6)$$

where each entry $\mathbf{S}_{pq}^{(j)}$ quantifies the correlation between frame p and q .

To sparsify the topology, we suppress the diagonal (*i.e.*, self-attention) and retain only the top- η highest-scoring off-diagonal neighbors for each row p , forming the directed neighborhood set $\text{Top}_p(\mathbf{S}^{(j)}, \eta) \subset \{1, \dots, F\}$. These selections are then used to construct a binary mask $\mathbf{M}^{(j)} \in \{0, 1\}^{F \times F}$ by restoring self-loops and applying symmetric completion:

$$\mathbf{M}_{pq}^{(j)} = \begin{cases} 1, & p = q \text{ or } q/p \in \text{Top}_{p/q}(\mathbf{S}^{(j)}, \eta), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

The detailed procedure for generating this joint-wise symmetric mask is summarized in Algorithm 1 Phase 1. To enforce sparse support, we update the similarity matrix by setting non-selected entries to a large negative constant, forming the masked version $\check{\mathbf{S}}^{(j)}$ that satisfies:

$$\check{\mathbf{S}}_{pq}^{(j)} = \begin{cases} \mathbf{S}_{pq}^{(j)}, & \text{if } \mathbf{M}_{pq}^{(j)} = 1, \\ -\infty, & \text{otherwise,} \end{cases} \quad (8)$$

TABLE II
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE HUMAN3.6M DATASET UNDER MPJPE FOR VARIOUS ACTIONS. *Dir.*, *Disc.*, $\dots$, AND *WalkT*. CORRESPOND TO 15 ACTION CLASSES. *Avg* INDICATES THE AVERAGE MPJPE AMONG 15 ACTION CLASSES. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINE FORMATS.

Method	Type	MPJPE ↓															Avg.
		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	
TCN [62]	CNN	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
DUE [22]	GCN	37.9	41.9	36.8	39.5	40.8	49.2	40.1	40.7	47.9	53.3	40.2	41.1	40.3	30.8	28.6	40.6
GLA-GCN [63]	GCN	41.3	44.3	40.8	41.8	45.9	54.1	42.1	41.5	57.8	62.9	45.0	42.8	45.9	29.4	29.9	44.4
FTCM [64]	MLP	42.2	44.4	42.4	42.4	47.7	55.8	42.7	41.9	58.7	64.5	46.1	44.2	45.2	30.6	31.1	45.3
PoseFormer [27]	Transformer	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
GraFormer [69]	Transformer	45.2	50.8	48.0	50.0	54.9	65.0	48.2	47.1	60.2	70.0	51.6	48.7	54.1	39.7	43.1	51.8
P-STMO [33]	Transformer	38.9	42.7	40.4	41.1	45.6	49.7	40.9	39.9	55.5	59.4	44.9	42.2	42.7	29.4	29.4	42.8
MixSTE [32]	Transformer	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
MHFormer [67]	Transformer	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
STCFormer [29]	Transformer	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5
MotionBERT [30] [‡]	Transformer	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	46.9	53.9	39.5	36.3	35.8	25.1	25.3	37.5
TC-MixSTE [34]	Transformer	36.9	40.9	36.3	39.0	41.6	48.7	38.4	39.3	50.3	54.9	40.6	38.0	40.6	26.5	26.0	39.9
DualFormer [36]	Transformer	38.9	43.1	39.2	41.4	45.1	50.7	41.5	41.2	51.7	60.5	43.4	41.4	43.0	29.9	31.0	42.8
DiffPose [52]	Diffusion	33.2	36.6	33.0	35.6	37.6	45.1	35.7	35.5	46.4	49.9	37.3	35.6	36.5	24.4	24.1	36.9
D3DP [37]	Diffusion	33.0	34.8	31.7	33.1	37.5	43.7	34.8	33.6	45.7	47.8	37.0	35.0	35.0	24.3	24.1	35.4
FinePOSE [54]	Diffusion	31.4	<u>31.5</u>	28.8	29.7	<u>34.3</u>	<u>36.5</u>	<u>29.2</u>	<u>30.0</u>	<u>42.0</u>	<u>42.5</u>	<u>33.3</u>	<u>31.9</u>	<u>31.4</u>	<u>22.6</u>	<u>22.7</u>	<u>31.9</u>
KTPFormer [55]	Diffusion	<u>30.1</u>	32.1	29.1	30.6	35.4	39.3	32.8	30.9	43.1	45.5	34.7	33.2	32.7	<u>22.1</u>	23.0	33.0
HTP (Ours)	Diffusion	28.5	30.0	26.4	27.2	31.5	36.0	28.8	27.7	39.5	39.1	30.7	29.1	30.7	21.7	22.3	29.9

this masked matrix ensures that only top- η correlations are retained during the subsequent softmax normalization.

The masked similarity $\check{\mathbf{S}}^{(j)}$ is normalized using the softmax function, and fused with the global adjacency $\mathbf{A}_T$ via Hadamard product to yield a joint-specific attention matrix:

$$\check{\mathbf{A}}_T^{(j)} = \mathbf{A}_T \odot \sigma_1(\check{\mathbf{S}}^{(j)}). \quad (9)$$

Finally, the pruned temporal graph is used to refine joint-wise tokens through linear projection and nonlinear activation with residual connection:

$$\mathbf{Y}_t'^{(j)} = \mathbf{Y}_t^{(j)} + \sigma_2(\check{\mathbf{A}}_T^{(j)} \mathbf{Y}_t^{(j)} \odot \mathbf{W}) \in \mathbb{R}^{F \times D}, \quad (10)$$

where $\mathbf{W} \in \mathbb{R}^{D \times D}$ is a shared linear layer, and $\sigma_2(\cdot)$ is the GELU activation function.

We then stack $\{\mathbf{Y}_t'^{(j)}\}_{j=1}^J$ and $\{\mathbf{M}^{(j)}\}_{j=1}^J$ to reconstruct $\mathbf{Y}_t' \in \mathbb{R}^{J \times F \times D}$ and the final temporal mask $\mathbf{M} \in \{0, 1\}^{J \times F \times F}$ for downstream modules.

C. Sparse-Focused Temporal MHSA (SFT MHSA)

Serving as an intermediate refinement within the *Frame-level Pruning* phase, we introduce the SFT MHSA module. Guided by the sparse temporal mask $\mathbf{M}$ generated by TCEP, it restricts attention computations to semantically correlated frames. Importantly, the module functions as a semantic bridge, enhancing the distinctiveness of the selected frames and ensuring that the tokens are semantically robust before undergoing physical compression in the next phase.

The architecture of the SFT MHSA module is illustrated in Fig. 2 (b). Given the temporally correlated pose token sequence $\mathbf{Y}_t' \in \mathbb{R}^{J \times F \times D}$, we first apply a LayerNorm and project the features into multi-head attention components: query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$). These are obtained

via linear projections using learnable weight matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V \in \mathbb{R}^{D \times D}$, respectively. The multi-head self-attention computation is then performed in parallel across h attention heads, each with dimension $d_k = D/h$. As shown in Algorithm 1 Phase 2, to enforce sparsity, we define an additive mask $\mathbf{M}' \in \mathbb{R}^{J \times F \times F}$ constructed by applying the following rule to each joint-specific binary mask $\mathbf{M}^{(j)} \in \{0, 1\}^{F \times F}$:

$$\mathbf{M}'_{pq} = \begin{cases} 0, & \text{if } \mathbf{M}_{pq}^{(j)} = 1, \\ -\infty, & \text{if } \mathbf{M}_{pq}^{(j)} = 0, \end{cases} \quad (11)$$

where $p, q \in \{1, \dots, F\}$. These joint-level masks are stacked to obtain $\mathbf{M}'$, ensuring that only relevant temporal connections are preserved during attention computation, while others are suppressed by assigning near-zero weights via softmax. The attention for each head is then computed as:

$$\text{head}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}} + \mathbf{M}' \right) \mathbf{V}_i, \quad (12)$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{J \times F \times d_k}$ are the i -th head-specific projections. The outputs of all heads are concatenated and projected via $\mathbf{W}_O \in \mathbb{R}^{D \times D}$, followed by a residual connection:

$$\tilde{\mathbf{Y}}_t = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}_O + \mathbf{Y}_t'. \quad (13)$$

The result is further processed through a LayerNorm-MLP block with GELU activation, and finalized with a second residual connection:

$$\tilde{\mathbf{Y}}_t' = \text{MLP}(\text{LN}(\tilde{\mathbf{Y}}_t)) + \tilde{\mathbf{Y}}_t. \quad (14)$$

D. Mask-Guided Pose Token Pruner (MGPTP)

Advancing to the *Semantic-level Pruning* phase, we propose the MGPTP module to transform the refined frame-level

TABLE III
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE HUMAN3.6M DATASET UNDER P-MPJPE FOR VARIOUS ACTIONS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE FORMATS.

Method	Type	P-MPJPE ↓															
		Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
TCN [62]	CNN	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	26.6	27.3	36.5
DUE [22]	GCN	30.3	34.6	29.6	31.7	31.6	38.9	31.8	31.9	39.2	42.8	32.1	32.6	31.4	25.1	23.8	32.5
GLA-GCN [63]	GCN	32.4	35.3	32.6	34.2	35.0	42.1	32.1	31.9	45.5	49.5	36.1	32.4	35.6	23.5	24.7	34.8
FTCM [64]	MLP	31.9	35.1	34.0	34.2	36.0	42.1	32.3	31.2	46.6	51.9	36.5	33.8	34.4	23.8	24.9	35.3
PoseFormer [27]	Transformer	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
P-STMO [33]	Transformer	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
MixSTE [32]	Transformer	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
MHFormer [67]	Transformer	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
STCFormer [29]	Transformer	29.3	33.0	30.7	30.6	32.7	38.2	29.7	28.8	42.2	45.0	33.3	29.4	31.5	20.9	22.3	31.8
TC-Mixste [34]	Transformer	29.5	32.8	28.9	31.6	32.8	37.9	29.8	29.1	41.8	44.3	33.5	30.6	32.2	21.2	22.2	31.9
DualFormer [36]	Transformer	31.4	34.9	32.5	34.3	35.1	39.4	33.0	32.0	42.9	48.9	36.2	32.9	33.5	23.7	25.1	34.4
D3DP [37]	Diffusion	27.5	29.4	26.6	27.7	29.2	34.3	27.5	26.2	37.3	39.0	30.3	27.7	28.2	19.6	20.3	28.7
KTPFormer [55]	Diffusion	<u>24.1</u>	<u>26.7</u>	<u>24.2</u>	<u>24.9</u>	<u>27.3</u>	<u>30.6</u>	<u>25.2</u>	<u>23.4</u>	<u>34.1</u>	<u>35.9</u>	<u>28.1</u>	<u>25.3</u>	<u>25.9</u>	<u>17.8</u>	<u>18.8</u>	<u>26.2</u>
HTP (Ours)	Diffusion	21.8	23.4	21.2	21.9	23.8	27.9	21.9	20.8	31.2	31.2	24.6	21.8	23.5	17.3	17.4	23.3

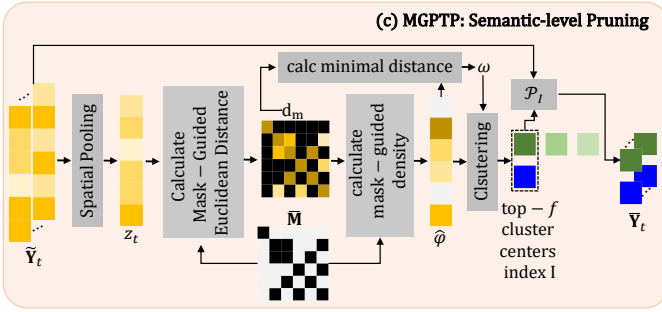


Fig. 3. The architecture of the MGPTP.

features into a compact semantic representation. Operating on the temporally refined tokens from SFT MHSA, MGPTP dynamically selects a compact subset of semantically informative frames by leveraging the learned attention masks. This mechanism aggregates temporal context into high-level semantic descriptors, maximizing computational efficiency while preserving essential motion cues.

As shown in Fig. 3, given the pose tokens after SFT MHSA, $\tilde{Y}_t' \in \mathbb{R}^{J \times F \times D}$, we apply average pooling along the joint dimension to obtain frame-wise tokens $z_t \in \mathbb{R}^{F \times D}$. Simultaneously, the joint-wise temporal mask $M \in \{0, 1\}^{J \times F \times F}$ is aggregated across joints via the same pooling operation, producing a smoothed frame-wise attention mask $\bar{M} \in \mathbb{R}^{F \times F}$. This unified step ensures both the token representation and its corresponding temporal guidance share consistent semantic granularity across frames.

Then, a novel mask-guided density peaks clustering based on the k-nearest neighbors algorithm is employed to cluster pose tokens with high motion relevance, guided by $\bar{M}$. Specifically, for any pair of frame tokens $z_p, z_q \in z_t$, the mask-guided euclidean distance is defined as:

$$d_m(z_p, z_q) = \begin{cases} \frac{\|z_p - z_q\|_2}{\sqrt{D}}, & \bar{M}_{pq} = 1, \\ \Lambda, & \bar{M}_{pq} = 0, \end{cases} \quad (15)$$

where $\Lambda = \max_{p,q} \|z_p - z_q\|_2 / \sqrt{D} + \varepsilon$, $\varepsilon > 0$ is a small constant ensuring that pairs masked out by $\bar{M}$ are strictly farther than any valid pair, effectively excluding them during neighborhood search.

Let k be a fixed neighborhood size, and $NN_k(z_p)$ be the k -th nearest point to z_p according to d_m . Thus, the mask-guided k-nearest neighbors $KNN_m(\cdot)$ of z_p is defined as:

$$KNN_m(z_p) = \{z_q \in z_t | d_m(z_p, z_q) \leq d_m(z_p, NN_k(z_p))\}. \quad (16)$$

The local density φ_p of frame p is then computed via a Gaussian kernel over $KNN_m(z_p)$:

$$\varphi_p = \exp \left(-\frac{1}{k} \sum_{z_q \in KNN_m(z_p)} d_m(z_p, z_q)^2 \right). \quad (17)$$

To emphasize frames that maintain richer temporal connectivity, we aggregate mask support per frame $s_p = \sum_{q=1}^F \bar{M}_{pq}$, apply a stability-aware transformation:

$$\tilde{s}_p = \begin{cases} s_p, & \text{if } s_p > 0, \\ -\infty, & \text{otherwise.} \end{cases} \quad (18)$$

The mask-guided response density $\hat{\varphi}_p$ is then computed by:

$$\hat{\varphi}_p = \varphi_p \cdot \sigma_1(\tilde{s}_p), \quad (19)$$

where $\sigma_1(\cdot)$ denotes the softmax function applied across all $\tilde{s}_p$. Next, the minimal distance ω_p to higher-density neighbors is computed for each frame:

$$\omega_p = \begin{cases} \min_{q: \hat{\varphi}_q > \hat{\varphi}_p} d_m(z_p, z_q), & \text{if } \exists \hat{\varphi}_q > \hat{\varphi}_p, \\ \max_q d_m(z_p, z_q), & \text{otherwise.} \end{cases} \quad (20)$$

Following the clustering principle in [70], the saliency score of each token z_p is defined as $\omega_p \times \hat{\varphi}_p$. We then select the top- f cluster centers and obtain an ordered index set $I = \{i_1 < i_2 < \dots < i_f\}$ on the temporal axis. Finally, we define an

order-preserving temporal selection operator $\mathcal{P}_I$ applied to the original joint-wise token sequence, which extracts the temporal slices at positions I uniformly across all joints. The resulting sequence is

$$\tilde{\mathbf{Y}}_t = \mathcal{P}_I(\tilde{\mathbf{Y}}'_t) \in \mathbb{R}^{J \times f \times D}, \quad (21)$$

which retains the full spatial and feature dimensions while compressing the temporal axis. This pruned sequence maintains temporal coherence and semantic consistency across all joints, enabling more efficient downstream reasoning.

V. EXPERIMENTS

A. Datasets and Metrics

1) Human3.6M [71]: Human3.6M is the largest and most widely used indoor benchmark dataset for HPE tasks, comprising 3.6 million RGB images covering 15 activities performed by 11 actors. Videos are recorded at 50Hz using four synchronized and calibrated cameras. Following [32], [37], [67], our model is trained on 5 subjects (S1, S5, S6, S7, S8) and evaluated on 2 subjects (S9, S11). For evaluation metrics, we report the mean per joint position error (MPJPE) and Procrustes MPJPE (P-MPJPE).

2) MPI-INF-3DHP [72]: MPI-INF-3DHP is a recently popular dataset consisting of indoor and outdoor scenes. The training set contains 8 activities performed by 8 actors, while the test set covers 7 activities. Following the protocol in [33], we use the area under the curve (AUC), percentage of correct keypoints (PCK), and MPJPE as evaluation metrics.

B. Implementation Details

1) Training Details: We use CPN [12] as the 2D keypoint detector to generate the 2D inputs. The numbers of SFT MHSA n_1 are set to 3. The batch size is set to 4, with each sample containing a pose sequence of 243 frames. For temporal pruning, the pruning length f is selected based on dataset-specific temporal characteristics. For Human3.6M, which contains long and high-frame-rate sequences with substantial temporal redundancy, a pruning length of $f = 54$ provides an effective balance between removing redundant frames and retaining sufficient motion granularity. For MPI-INF-3DHP, where sequences are shorter and motion cues are more concentrated, a conservative pruning ratio of 3:1 ($f = 27$) is adopted to preserve essential fine-grained motion information. We adopt the AdamW [73] optimizer with the momentum parameters of $\beta_1, \beta_2 = 0.9, 0.999$, and a weight decay of 0.1. We train our model for 150 epochs and the initial learning rate is $6e^{-5}$ with a shrink factor of 0.993 after each epoch. For fair comparisons, we set the number of hypotheses $H = 1$ and iterations $K = 1$ during training, and $H = 20$ and $K = 10$ during inference, as in D3DP [37].

2) Implementation of Plug-and-Play Integration: To verify generality, we integrated HTP into Transformer-based frameworks. All variants were trained from scratch to ensure fair comparison. For MixSTE [32], the integration follows the identical architectural modifications as our main method. For the dual-stream MotionBERT [30], we inserted TCEP and an

initial SFT MHSA layer prior to the backbone. Inside the first $n_1 = 2$ DSTformer blocks, we replaced the original Temporal MHSA components with our SFT MHSA. The MGPTP module was inserted after the second block, allowing the remaining layers to process the efficient, pruned sequence.

3) Implementation of Efficiency Evaluation: We report the total Multiply-Accumulate operations (MACs), representing the cumulative computational cost aggregated across all K denoising iterations and H hypotheses. These values are computed via the THOP library, following the standard convention where 1 MACs $\approx$ 2 FLOPs. Inference speed (FPS) is measured on two NVIDIA GeForce RTX 4090 GPUs with batch size of 8 under FP32 precision. The FPS is calculated by dividing the total number of processed frames by the total wall-clock time of inference, excluding data loading overhead.

C. Quantitative Results

1) Human3.6M: Tab. I presents comparisons between our HTP and recent state-of-the-art (SOTA) 3D HPE methods on the Human3.6M dataset. Our approach achieves SOTA performance, with MPJPE of 29.9mm and P-MPJPE of 23.3mm using 2D poses detected by CPN [12] as inputs, and MPJPE of 16.7mm when using ground-truth 2D poses as inputs. With an 81% reduction in parameter count and a 40.0% decrease in computational cost, HTP surpasses the previous SOTA method FinePose [54] by 2.0mm in MPJPE and 1.7mm in P-MPJPE.

Additionally, to demonstrate the plug-and-play capability of our design, we integrate the proposed Hierarchical Temporal Pruning strategy into two Transformer-based frameworks: MixSTE [32] and MotionBERT [30]. On MixSTE, HTP reduces MACs by 37% while improving MPJPE by 1.0mm and P-MPJPE by 0.7mm. On MotionBERT, we observe similar gains, with a 42% reduction in MACs and improved pose accuracy across both metrics. We further compare HTP with emerging Mamba-based methods, PoseMamba-X [65] and SAMA-L [66]. While these state-space models offer competitive efficiency, HTP demonstrates superior reconstruction fidelity, outperforming SAMA-L by 7.0 mm in MPJPE.

Further, Tab.II and Tab.III report per-action results on the Human3.6M dataset under MPJPE and P-MPJPE metrics, using detected 2D keypoints [12] as input. Our HTP framework consistently achieves the lowest error across all 15 action categories, outperforming all prior state-of-the-art methods. Notably, significant improvements are observed in challenging categories such as “SitD”, “Walk”, and “WalkT”, highlighting its ability to handle diverse motion dynamics with precision.

2) MPI-INF-3DHP: Tab. IV presents comparisons between our method and recent SOTA 3D HPE approaches on the MPI-INF-3DHP dataset. Following [37], [54], our model is trained using ground-truth 2D poses as inputs. Compared to recent SOTA works [54], [55], HTP maintains comparable MPJPE while improving PCK by 0.5% and AUC by 0.5%. While the performance gain is less pronounced than in Human3.6M due to the reduced temporal redundancy available for pruning in shorter sequences, these results confirm that HTP remains robust and effective even under constrained temporal resolutions.

3) MACs and Speed: Tab. V presents a comprehensive comparison between HTP and recent diffusion-based 3D HPE

TABLE IV
QUANTITATIVE COMPARISON WITH SOTA METHODS ON THE MPI-INF-3DHP DATASET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINE FORMATS.

Method	Type	F	MPI-INF-3DHP		
			PCK $\uparrow$	AUC $\uparrow$	MPJPE $\downarrow$
TCN [62]	CNN	81	86.0	51.9	84.0
GLA-GCN [63]	GCN	81	98.5	79.1	27.8
FTCM [64]	MLP	81	98.0	79.8	31.2
PoseFormer [27]	Transformer	9	88.6	56.4	77.1
P-STMO [33]	Transformer	81	97.9	75.8	32.2
MixSTE [32]	Transformer	27	94.4	66.5	54.9
PoseFormerV2 [28]	Transformer	81	97.9	78.8	27.8
MHFormer [67]	Transformer	9	93.8	63.3	58.0
TC-MixSTE [34]	Transformer	81	98.7	79.5	27.6
DualFormer [36]	Transformer	9	97.8	73.4	40.1
Diffpose [52]	Diffusion	81	98.0	75.9	29.1
D3DP [37]	Diffusion	81	98.0	79.1	28.1
KTPFormer [55]	Diffusion	81	99.0	79.3	29.1
FinePOSE [54]	Diffusion	81	98.9	<u>80.0</u>	26.2
HTP (Ours)	Diffusion	81	99.5	80.5	<u>26.4</u>

TABLE V
MACs AND SPEED COMPARISON WITH DIFFUSION-BASED 3D HPE METHODS. ALL MODELS ARE EVALUATED UNDER THE SAME SETTING. BEST RESULTS ARE BOLDDED.

Method	MPJPE↓	Params (M)	Train	Inference	
			MACs/frame	MACs/frame	FPS↑
Setting: $K = 1, H = 20$					
D3DP [37]	38.8	34.8	0.58 G	22.9 G	772.7
FinePose [54]	40.0	200.6	0.60 G	22.9 G	723.7
KTPFormer [55]	39.5	36.3	0.58 G	23.6 G	705.8
HTP (Ours)	32.9	37.5	0.36 G	10.0 G	2277.5
Setting: $K = 10, H = 20$					
D3DP [37]	35.4	34.8	0.58 G	228.8 G	79.6
FinePose [54]	31.9	200.6	0.60 G	236.2 G	73.8
KTPFormer [55]	33.0	36.3	0.58 G	228.8 G	73.5
HTP (Ours)	29.9	37.5	0.36 G	99.8 G	137.0

methods across multiple sampling configurations ($K=1, 10$) with a fixed inference horizon ($H=20$). Under the low-sampling regime ($K=1$), HTP achieves a substantial acceleration of $3\times$ in inference FPS while reducing per-frame MACs by over 56%, along with the lowest MPJPE of 32.9. As K increases, HTP consistently achieves both lower error and lower inference cost compared to all prior methods, demonstrating its robustness to different sampling budgets. Notably, even with $K=1$, our method surpasses the accuracy of several baselines operating at $K=10$ (e.g., 32.9 MPJPE vs. 35.4 of D3DP), while reducing MACs from 457.6G to 20.0G and boosting FPS from 142.2 to 2443.9—representing a $16\times$ speedup. HTP thus achieves strong generalization and cost-efficiency across budget settings, enabling real-world deployment.

4) Efficiency Comparison with Non-Diffusion Baselines: We compare our diffusion-based HTP with representative Transformers in Tab. VI. Compared to the seq-to-frame baseline PoseFormer [27] (1.62 G/frame), our HTP ($n_1 = 1$) is drastically more efficient (0.50 G/frame) and accurate (40.6 mm vs. 44.3 mm). Against the efficient seq-to-seq STC-

TABLE VI
EFFICIENCY COMPARISON WITH NON-DIFFUSION BASELINES ON HUMAN3.6M.

Method	Type	MPJPE $\downarrow$	Params (M)	MACs/frame
PoseFormer [27]	Transformer	44.3	9.5	1.62 G
PoseFormerV2 [28]	Transformer	45.2	9.5	2.10 G
STCFormer [29]	Transformer	40.8	18.9	0.32 G
HTP w/ MixSTE	Transformer	39.9	36.4	0.36 G
Setting: $K = 1, H = 1$				
D3DP [37]	Diffusion	40.0	34.6	1.14 G
HTP (ours) w/ $n_1=3$	Diffusion	39.8	37.5	0.72 G
HTP (ours) w/ $n_1=1$	Diffusion	40.6	37.5	0.50 G

TABLE VII
ABLATION STUDY ON THE LOCATION OF MGPTP MODULE ON THE HUMAN3.6M DATASET.

f	n_1	MPJPE $\downarrow$	MACs (G)
54	1	33.0	60.6
54	2	32.0	74.2
54	3	29.9	87.7
54	4	30.8	101.2

TABLE VIII
EFFECT OF THE SEQUENCE LENGTH AFTER MGPTP MODULE ON THE HUMAN3.6M DATASET.

f	n_1	MPJPE $\downarrow$	MACs (G)
27	3	31.2	79.7
54	3	29.9	87.7
81	3	30.1	95.6
108	3	32.1	103.6

Former [29] (0.32 G/frame), our method achieves higher accuracy (40.6 mm vs. 40.8 mm) at a comparable computational scale. Moreover, our high-fidelity setting ($n_1 = 3$) further reduces the error to 39.8 mm while maintaining an affordable cost (0.72 G/frame), effectively positioning diffusion models as a high-performance competitor to lightweight transformers.

D. Sensitivity Analysis and Ablation Study

We conduct comprehensive sensitivity and ablation studies on the Human3.6M dataset to validate the design of HTP. Specifically, we examine: (1)–(3) key architectural hyperparameters, including the pruned sequence length f , module placement n_1 , and neighborhood sizes (η, k); (4) the flexibility of adjusting parameters (η, n_1) at inference time to balance efficiency; (5)–(6) the incremental contribution of each module and the specific role of the sparse mask $\mathbf{M}$ in guiding attention and pruning; and (7) the impact of input sequence length F on the trade-off between performance and computational cost.

1) Sequence length after MGPTP Module: The number of representative pose tokens retained by the MGPTP module plays a crucial role in performance. As shown in Tab. VII, increasing the sequence length does not lead to the anticipated performance improvements. This is because retaining too many pose tokens during clustering can amplify the impact of discrete or less informative tokens, which disrupts the coherent understanding of overall motion patterns. Thus, we set $f = 54$ in all experiments to achieve an optimal balance.

2) Location of MGPTP: In Tab. VIII, we explore different settings for the number of SFT MHSA layers n_1 , which effectively determines the placement of the MGPTP module. Adjusting n_1 allows for a trade-off between computational efficiency and performance. From the results, we observe that lower values of n_1 lead to reduced performance, as pruning pose tokens too early limits the network's ability

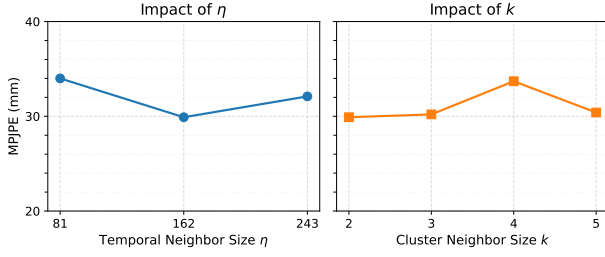
Fig. 4. The sensitivity analysis of η and k .

TABLE IX
ADJUSTING THE TEMPORAL
NODE NUMBER η IN TCEP
DURING INFERENCE.

η	MPJPE	FPS
243	30.3	242.5
162	29.9	244.6
81	34.4	254.3

TABLE X
ADJUSTING THE NUMBER OF
SFT MHSA LAYERS n_1
DURING INFERENCE.

n_1	MPJPE	MACs	FPS
1	29.9	99.8	244.6
2	29.7	122.0	204.5
3	29.1	228.8	176.5

to learn aggregated motion information through SFT MHSA. Conversely, setting n_1 too high can result in attention collapse, which negatively impacts performance. We find that $n_1 = 3$ provides the best balance between efficiency and accuracy, and thus use this setting for all experiments.

3) Hyperparameters η and k : As visualized in Fig. 4, we analyze the sensitivity of η (in TCEP) and k (in MGPTP). For η , performance peaks at $\eta = 162$; lower values (e.g., 81) fail to capture sufficient context, while full sequences ($\eta = 243$) introduce redundancy that negates pruning benefits. Regarding k , a compact neighborhood ($k = 2$) yields the lowest MPJPE (29.9 mm). Increasing k tends to over-smooth local density estimates, obscuring distinct motion states. Consequently, we adopt $\eta = 162$ and $k = 2$ as the default configuration.

4) Inference Hyperparameters: We further evaluate the impact of adjusting hyperparameters during the inference phase. As shown in the Tab. IX, for the temporal node number η , the optimal performance is achieved when the inference value matches the training setting ($\eta = 162$). Regarding the number of SFT MHSA layers n_1 (Tab. X), although matching the training depth ($n_1 = 3$) yields the marginal best MPJPE, it incurs higher computational costs. We thus adopt $n_1 = 1$ as the default inference configuration for efficiency, achieving a significant FPS boost and MACs reduction with only a negligible performance trade-off compared to $n_1 = 3$.

5) Effect of Each Module: Tab. XI investigates the role of each module through incremental configurations. We begin with a baseline that extends D3DP [37] by adding the Spatial GCN from [31]. In **Setting1**, we introduce the TCEP module without applying its generated sparse mask, resulting in a significant MPJPE reduction of 2.8 with only a minor increase in parameters (0.8M) and a marginal computational cost of 0.017 G MACs. This highlights TCEP's effectiveness in capturing essential motion cues by dynamically selecting temporal dependencies. In **Setting2**, we further adopt the sparse mask from TCEP to replace full temporal attention with global SFT MHSA. However, this setting performs worse than Setting1,

TABLE XI
ABLATION STUDY ON DIFFERENT DESIGNS OF HTP.

Setting	TCEP	SFT MHSA	MGPTP	MPJPE↓	Param	MACs
Baseline				34.7	35.5	143.375
Setting1	✓			31.9	36.3 ↑0.8	143.392 ↑0.017
Setting2	✓	✓		33.3	36.3 ↑0.0	143.392 ↑0.000
Setting3	✓		✓	31.6	37.5 ↑2.0	87.648 ↓55.727
HTP (Ours)	✓	✓	✓	29.9	37.5 ↑2.0	87.648 ↓55.727

TABLE XII
ANALYSIS OF THE IMPACT OF THE SPARSE MASK $\mathbf{M}$.

Setting	Temporal MHSA		MGPTP		MPJPE↓
	w $\mathbf{M}$	w/o $\mathbf{M}$	w $\mathbf{M}$	w/o $\mathbf{M}$	
Setting4		✓		✓	32.7
Setting5		✓	✓		31.6
Setting6	✓			✓	31.3
HTP (Ours)	✓		✓		29.9

indicating that overly aggressive sparsification may lead to attention collapse and hinder temporal reasoning. These results suggest that *moderate* application of sparse self-attention is essential to maintain global context.

In **Setting3**, we remove the sparse mask from Setting2 and instead apply the MGPTP module to prune redundant tokens. By utilizing the sparsity mask generated by TCEP to compress the sequence length, this setting achieves a net reduction of 55.7 G MACs, demonstrating that the efficiency gains from hierarchical pruning vastly outweigh the minimal overhead of the pruning decision modules. Finally, the full HTP configuration combines selective sparse attention with MGPTP, achieving the optimal balance between reconstruction accuracy and computational cost.

6) Impact of the Sparse Mask $\mathbf{M}$: As shown in Tab. XII, we evaluate the role of the mask $\mathbf{M}$ generated by TCEP. Compared to the baseline without mask guidance (**Setting 4**), applying $\mathbf{M}$ solely to MGPTP (**Setting 5**) or SFT MHSA (**Setting 6**) yields moderate gains. However, the full HTP configuration, which leverages $\mathbf{M}$ in both modules, outperforms Setting 5 and Setting 6 by 1.7mm and 1.4mm, respectively. This demonstrates that sparse attention and mask-guided pruning are mutually reinforcing: selective attention enhances feature discriminability for pruning, while accurate token retention preserves global context. Thus, $\mathbf{M}$ serves as a critical structural bridge, integrating the modules into a cohesive pruning pipeline.

7) Impact of Input Sequence Length F : We analyze the trade-off between sequence length and performance in Tab. XIII. Following prior methods [29], [32], [37], [54], we adopt $F = 243$ as the default to maximize long-range temporal context. Empirical results show that reducing F to 162 or 81 degrades MPJPE. While shorter sequences reduce VRAM consumption and allow for larger batch sizes, our tests indicate that larger batch sizes yield minimal gains in

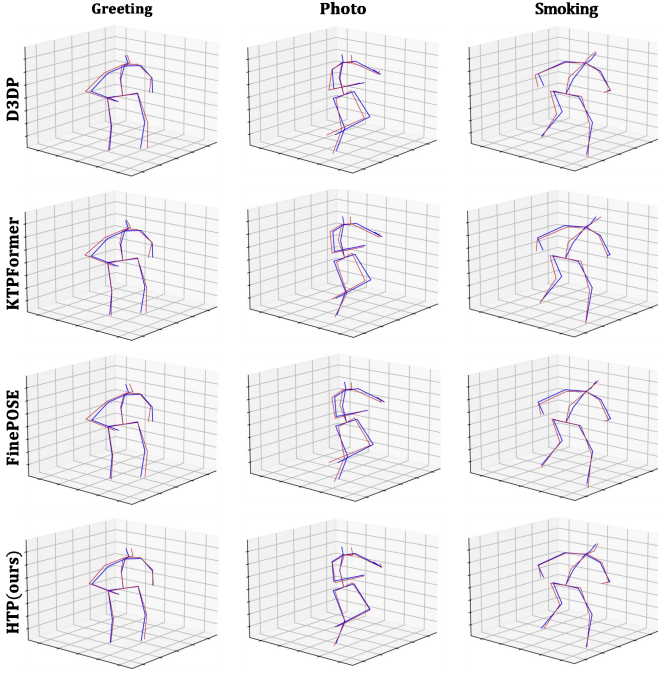


Fig. 5. Qualitative comparisons of our HTP with previous state-of-the-art methods [37], [54], [55] on the Human3.6M dataset. Solid blue line: ground-truth 3D pose. Solid red line: estimated 3D pose.

TABLE XIII
IMPACT OF INPUT SEQUENCE LENGTH F .

F	Batch Size	MPJPE↓	Inference	
			MACs/frame	FPS↑
81	4	34.4	119.2	139.2
81	32	34.4	119.2	140.6
162	4	34.3	119.2	130.2
162	12	34.3	119.2	130.8
243	4	29.9	99.8	137.0

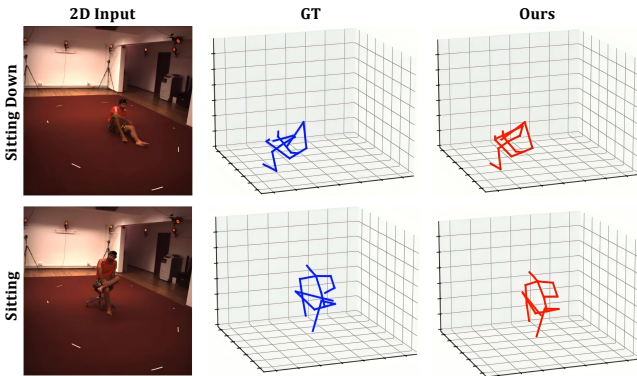


Fig. 7. Visualization of a failure case under severe self-occlusion.

inference FPS. Given that HTP’s pruning strategy already ensures high efficiency for 243-frame inputs, we prioritize the accuracy benefits of the longer sequence, recommending

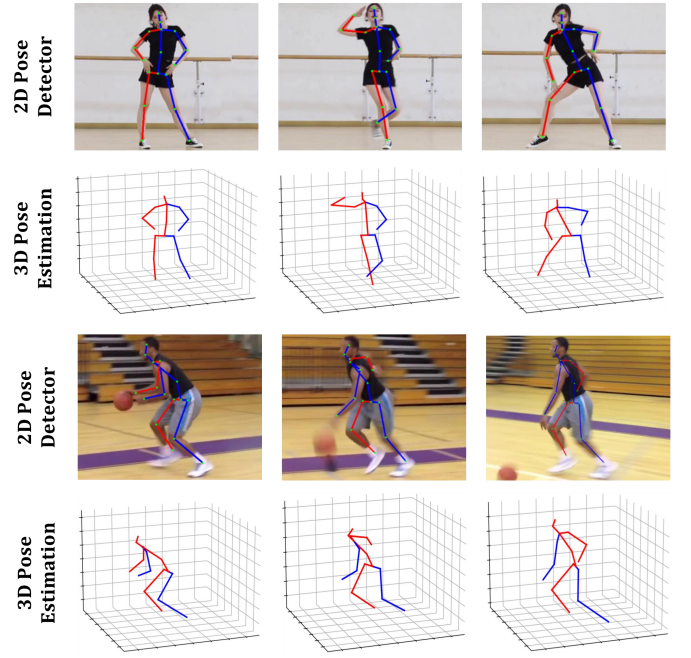


Fig. 6. Qualitative results of the proposed method on in-the-wild videos. Setting $H=5$, $K=5$. HRNet [14] was used as the 2D keypoint detector to generate 2D poses.

shorter versions only for memory-constrained environments.

E. Qualitative Analysis

1) **Qualitative results comparison:** Fig. 5 compares HTP with state-of-the-art diffusion-based methods [37], [54], [55] on Human3.6M ($H = 20$, $K = 10$). Across actions ranging from simple poses (e.g., “Photo”) to complex articulations (e.g., “Sitting down”, “Smoking”), HTP consistently yields 3D estimations that align closer to the ground truth. Notably, our method demonstrates superior fidelity in limb joints (e.g., elbows and wrists) and maintains better structural plausibility in challenging poses compared to baselines, effectively mitigating the joint distortions observed in D3DP and KTPFormer.

2) **In-the-wild Videos:** To validate generalization, we evaluate HTP on wild videos using 2D poses detected by HRNet [14]. As shown in Fig. 6, our method exhibits exceptional robustness, maintaining high accuracy even in challenging scenarios with severe self-occlusion and rapid motion.

3) **Qualitative Analysis of Frame Retention:** Fig. 8 visualizes the adaptive pruning behavior of HTP on “Walking” and “Sitting”. By selecting only 54 representative frames from the 243-frame input, HTP dynamically allocates computational resources based on motion complexity. Specifically, the model retains a higher density of frames during rapid transitions (blue dashed box) to capture fast-moving dynamics, while aggressively pruning frames during stable or slow-motion phases (cyan/green dashed boxes). This content-aware strategy effectively balances temporal sparsity with representational completeness, maintaining accurate 3D pose estimation while significantly reducing computational overhead.

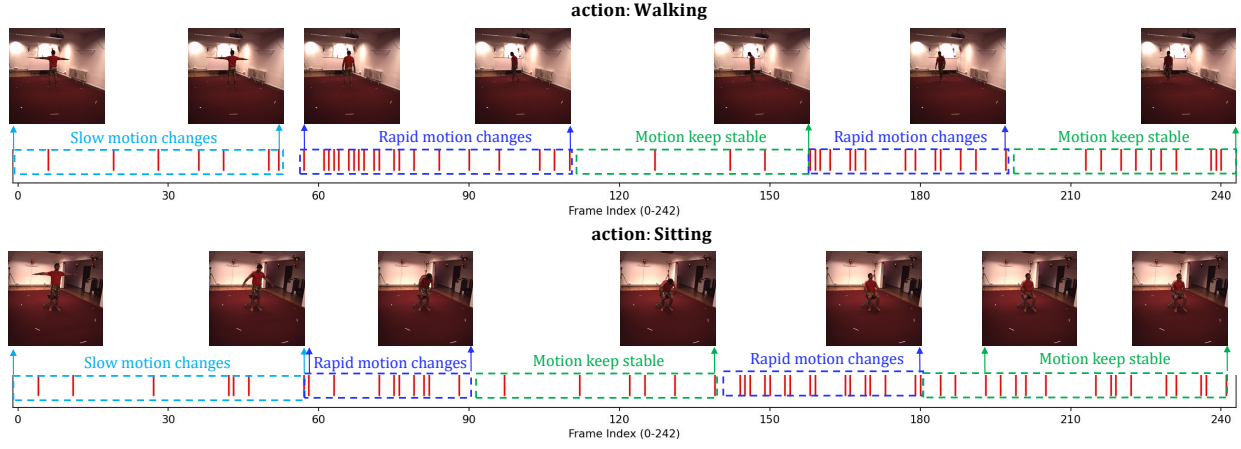


Fig. 8. **Qualitative Analysis of Frame Retention in HTP.** For each action, the timeline at the bottom shows the frame index (from 0 to 242), with red vertical lines marking the frames retained by HTP.

VI. LIMITATIONS AND FUTURE WORKS

Despite HTP's substantial efficiency gains, specific challenges present avenues for improvement. First, as shown in Fig. 7, severe self-occlusions can cause the pruning mechanism to inadvertently discard critical frames needed for resolving complex articulations. Second, as a 2D-to-3D lifting framework, HTP's performance is bounded by 2D input quality, limiting gains in noisy outdoor scenarios. Future work will address these by exploring occlusion-aware attention and spatial uncertainty modeling to enhance robustness.

VII. CONCLUSIONS

In this paper, we address the efficiency challenges in diffusion-based 3D human pose estimation with Hierarchical Temporal Pruning (HTP). HTP progressively reduces redundancy by selectively pruning pose tokens across both frame and semantic levels while preserving critical motion dynamics. Starting with temporal correlations in TCEP, focusing attention on SFT MHSA, and refining through MGPTP's semantic-level clustering, HTP selectively retains the most motion-critical pose tokens. Extensive experiments on Human3.6M and MPI-INF-3DHP datasets demonstrate that HTP achieves state-of-the-art accuracy while substantially reducing computational cost and improving inference speed.

REFERENCES

- [1] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with spatio-temporal visual attention on skeleton image sequences," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 29, no. 8, pp. 2405–2415, 2019.
- [2] H. Wang, W. Weng, J. Wang, F. Zhao, G.-S. Xie, X. Geng, and L. Wang, "Foundation model for skeleton-based human action understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 1, pp. 3247–3257, 2026.
- [3] J. Kuang, H. Wang, C. Han, Y. Zhang, and J. Gui, "Zero-shot skeleton-based action recognition with dual visual-text alignment," *Pattern Recognition*, vol. 171(1), pp. 1–15, 2025.
- [4] H. Wang, B. Yu, J. Li, L. Zhang, and D. Chen, "Multi-stream interaction networks for human action recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 32, no. 5, pp. 3050–3060, 2022.
- [5] J. Zhang, Y. Jia, W. Xie, and Z. Tu, "Zoom transformer for skeleton-based group activity recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 32, no. 12, pp. 8846–8859, 2022.
- [6] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3D human pose estimation in rgbd images for robotic task learning," in *Conference on Robotics and Automation*, 2018, p. 1986–1992.
- [7] M. Svenstrup, S. Tranberg, H. J. Andersen, and T. Bak, "Pose estimation and adaptive robot behaviour for human-robot interaction," in *Conference on Robotics and Automation*, 2009, pp. 3571–3576.
- [8] H. Wang, X. Ma, J. Kuang, and J. Gui, "Heterogeneous skeleton-based action representation learning," in *Conference on Computer Vision and Pattern Recognition*, 2025, pp. 19 154–19 164.
- [9] N. Hagbi, O. Bergig, J. El-Sana, and M. Billinghamurst, "Shape recognition and pose estimation for mobile augmented reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 10, pp. 1369–1379, 2010.
- [10] J. S. Yoon, L. Liu, V. Golyanik, K. Sarkar, H. S. Park, and C. Theobalt, "Pose-guided human animation from a single image in the wild," in *Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 039–15 048.
- [11] Y. Wang and Y. Peng, "Mars: Learning modality-agnostic representation for scalable cross-media retrieval," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 32, no. 7, pp. 4765–4777, 2022.
- [12] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Conference on Computer Vision and Pattern Recognition*, 2018, p. 7103–7112.
- [13] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*, 2016, pp. 483–499.
- [14] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.
- [15] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, 2023.
- [16] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [17] S. Liu, Y. Li, and G. Hua, "Human pose estimation in video via structured space learning and halfway temporal evaluation," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 29, no. 7, pp. 2029–2038, 2019.
- [18] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks," in *International Conference on Computer Vision*, 2019, pp. 2272–2281.
- [19] W. Hu, C. Zhang, F. Zhan, L. Zhang, and T.-T. Wong, "Conditional directed graph convolution for 3D human pose estimation," in *ACM International Conference on Multimedia*, 2021, pp. 602–611.
- [20] Y. You, H. Liu, T. Wang, W. Li, R. Ding, and X. Li, "Co-evolution of pose and mesh for 3D human body estimation from video," in *International Conference on Computer Vision*, 2023, pp. 14 963–14 973.
- [21] A. Zeng, X. Sun, F. Huang, M. Liu, Q. Xu, and S. Lin, "Snet: Improving generalization in 3D human pose estimation with a split-and-recombine

- approach,” in *European Conference on Computer Vision*, 2020, pp. 507–523.
- [22] J. Zhang, Y. Chen, and Z. Tu, “Uncertainty-aware 3D human pose estimation from monocular video,” in *ACM International Conference on Multimedia*, 2022, p. 5102–5113.
- [23] G. Wei, C. Lan, W. Zeng, and Z. Chen, “View invariant 3D human pose estimation,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 12, pp. 4601–4610, 2020.
- [24] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, “Anatomy-aware 3D human pose estimation with bone-based pose decomposition,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 32, no. 1, pp. 198–209, 2022.
- [25] M. Hu, C. Liu, S. Li, Q. Yan, Q. Fang, and Q. Chen, “A geometric knowledge oriented single-frame 2D-to-3D human absolute pose estimation method,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 12, pp. 7282–7295, 2023.
- [26] H. Chen, J. He, W. Xiang, Z. Cheng, W. Liu, H. Liu, B. Luo, Y. Geng, and X. Xie, “Hdformer: High-order directed transformer for 3D human pose estimation,” in *International Joint Conference on Artificial Intelligence*, 2023, pp. 581–589.
- [27] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, “3D human pose estimation with spatial and temporal transformers,” in *Conference on Computer Vision and Pattern Recognition*, 2021, p. 11656–11665.
- [28] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, “Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8877–8886.
- [29] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, “3D human pose estimation with spatio-temporal criss-cross attention,” in *Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4790–4799.
- [30] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, “Motionbert: A unified perspective on learning human motion representations,” in *International Conference on Computer Vision*, 2023, pp. 15 085–15 099.
- [31] S. Mehraban, V. Adeli, and B. Taati, “Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network,” in *Winter Conference on Applications of Computer Vision*, 2024, pp. 6920–6930.
- [32] 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 *Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 232–13 242.
- [33] W. Shan, Z. Liu, X. Zhang, S. Wang, S. Ma, and W. Gao, “P-stmo: Pre-trained spatial temporal many-to-one model for 3D human pose estimation,” in *European Conference on Computer Vision*, 2022, p. 461–478.
- [34] R. Wang, X. Ying, and B. Xing, “Exploiting temporal correlations for 3D human pose estimation,” *IEEE Trans. Multimedia*, vol. 26, pp. 4527–4539, 2023.
- [35] Y. Zhong, G. Yang, D. Zhong, X. Yang, and S. Wang, “Frame-padded multiscale transformer for monocular 3D human pose estimation,” *IEEE Trans. Multimedia*, vol. 26, pp. 6191–6201, 2023.
- [36] L. Zhou, Y. Chen, and J. Wang, “Dual-path transformer for 3D human pose estimation,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 5, pp. 3260–3270, 2023.
- [37] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, “Diffusion-based 3D human pose estimation with multi-hypothesis aggregation,” in *International Conference on Computer Vision*, 2023, p. 14761–14771.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszko-reit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, “Attention is all you need,” in *Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [39] J. Li, Y. Bi, S. Wang, and Q. Li, “Cfml-net: A context-aware feature representation learning anchor-free network for pedestrian detection,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 9, pp. 4948–4961, 2023.
- [40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
- [41] Y. Wang, X. Ou, J. Liang, and Z. Sun, “Deep semantic reconstruction hashing for similarity retrieval,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 1, pp. 387–400, 2021.
- [42] C. Jiang, M. Dong, X. Shi, and J. Gui, “Diversifying counterattacks: Orthogonal exploration for robust clip inference,” in *AAAI Conference on Artificial Intelligence*, 2026.
- [43] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, 2020.
- [44] G. Kim, H. Shim, H. Kim, Y. Choi, J. Kim, and E. Yang, “Diffusion video autoencoders: Toward temporally consistent face video editing via disentangled video encoding,” in *Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6091–6100.
- [45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.
- [46] Y. Wang, Y. Wu, Z. Dai, C. Tian, J. Long, and J. Chen, “Noisy correspondence rectification via asymmetric similarity learning,” in *AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 384–21 392.
- [47] H. Chunga, B. Sim, and J. C. Ye, “Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction,” in *Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 413–12 422.
- [48] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, “Physdiff: Physics-guided human motion diffusion model,” in *International Conference on Computer Vision*, 2023, pp. 16 010–16 021.
- [49] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, “Motiondiffuse: Text-driven human motion generation with diffusion model,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, 2024.
- [50] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, “Motiongpt: Human motion as a foreign language,” in *Neural Information Processing Systems*, 2023, pp. 20 067–20 079.
- [51] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt, “Mofusion: A framework for denoising-diffusion-based motion synthesis,” in *Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9760–9770.
- [52] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, “Diffpose: Toward more reliable 3D pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2023, p. 13041–13051.
- [53] W. Shan, Y. Zhang, X. Zhang, S. Wang, X. Zhou, S. Ma, and W. Gao, “Diffusion-based hypotheses generation and joint-level hypotheses aggregation for 3D human pose estimation,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 11, pp. 10 681–10 691, 2024.
- [54] J. Xu, Y. Guo, and Y. Peng, “Finepose: Fine-grained prompt-driven 3D human pose estimation via diffusion models,” in *Conference on Computer Vision and Pattern Recognition*, 2024, pp. 561–570.
- [55] J. Peng, Y. Zhou, and P. Y. Mok, “Ktformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1123–1132.
- [56] A. Zeng, X. Ju, L. Yang, R. Gao, X. Zhu, B. Dai, and Q. Xu, “Deciwatch: A simple baseline for 10x efficient 2D and 3D pose estimation,” in *European Conference on Computer Vision*, 2022, p. 607–624.
- [57] M. Cui, K. Zhang, and Z. Sun, “Graph and skipped transformer: Exploiting spatial and temporal modeling capacities for efficient 3D human pose estimation,” *arXiv preprint arXiv:2407.02990*, 2024.
- [58] M. Einfalt, K. Ludwig, and R. Lienhart, “Uplift and upsample: Efficient 3D human pose estimation with uplifting transformers,” in *Winter Conference on Applications of Computer Vision*, 2023, pp. 2903–2913.
- [59] W. Li, M. Liu, H. Liu, P. Wang, J. Cai, and N. Sebe, “Hourglass tokenizer for efficient transformer-based 3D human pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2024, pp. 604–613.
- [60] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Adv. Neural Inf. Proces. Syst.*, vol. 33, p. 6840–6851, 2020.
- [61] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.
- [62] D. Pavlo, C. Feichtenhofer, D. Grangier, and M. Auli, “3D human pose estimation in video with temporal convolutions and semi-supervised training,” in *Conference on Computer Vision and Pattern Recognition*, 2019, p. 7753–7762.
- [63] B. X. Yu, Z. Zhang, Y. Liu, S. Hua, Z. Zhong, Y. Liu, and C. W. Chen, “Gla-gcn: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video,” in *International Conference on Computer Vision*, 2023, pp. 8818–8829.
- [64] Z. Tang, Y. Hao, J. Li, and R. Hong, “Ftcm: Frequency-temporal collaborative module for efficient 3D human pose estimation in video,” *IEEE Trans. Circuit Syst. Video Technol.*, vol. 34, no. 2, pp. 911–923, 2024.
- [65] Y. Huang, J. Liu, K. Xian, and R. C. Qiu, “Posemamba: Monocular 3D human pose estimation with bidirectional global-local spatio-temporal

- state space model,” in *AAAI Conference on Artificial Intelligence*, 2025, pp. 3842–3850.
- [66] Y. Lu, J. Wang, J. Gao, R. Gong, C. Cai, and K.-H. Yap, “A structure-aware and motion-adaptive framework for 3D human pose estimation with mamba,” in *International Conference on Computer Vision*, 2025, pp. 7958–7968.
 - [67] W. Li, H. Liu, H. Tang, P. Wang, and L. V. Gool, “Mhformer: Multi-hypothesis transformer for 3D human pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2022, p. 13147–13156.
 - [68] K. Holmquist and B. Wandt, “Diffpose: Multi-hypothesis human pose estimation using diffusion models,” in *International Conference on Computer Vision*, 2023, pp. 15 977–15 987.
 - [69] W. Zhao, W. Wang, and Y. Tian, “Graformer: Graph-oriented transformer for 3D pose estimation,” in *Conference on Computer Vision and Pattern Recognition*, 2022, p. 20438–20447.
 - [70] M. Du, S. Ding, and H. Jia, “Study on density peaks clustering based on k-nearest neighbors and principal component analysis,” *Knowledge-Based Systems*, vol. 99, p. 135–145, 2016.
 - [71] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, 2013.
 - [72] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, “Monocular 3D human pose estimation in the wild using improved cnn supervision,” in *International Conference on 3D Vision*, 2017, p. 506–516.
 - [73] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.