This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

Graph Attention Fusion Network for Robust 3D Human Motion Tracking

Yen Pham*, Xiaohui Yuan*, Chengyuan Zhuang†, Zhenchun Wei‡, Tian Chen‡, Lichuan Gu§

***Abstract*—Accurate tracking of human motion from videos is critical for real-world applications, but the existing lifting methods face performance degradation with noisy inputs and occlusions. The limitations typically manifest as 3D motion jitter and structural inconsistencies, which require an extra refinement method to smooth out the corrupted outputs. This paper presents a dual-stream network that integrates semantic-rich spatial-temporal joint features with local joint velocity through a motion-aware fusion mechanism. GAtFuN devises graph attention modules to encode skeletal connectivity and enforce temporal coherence. This design ensures kinematically plausible and temporally smooth 3D motion tracking, even under occlusion or low-quality input. Extensive experimental results using public datasets demonstrate improved performance. Compared to the state-of-the-art networks, GAtFuN reduced the Mean Per Joint Position Error and Procrustes-Aligned Mean Per Joint Position Error by 7.8% and 6.8%, respectively. GAtFuN demonstrated superior temporal smoothness and the fewest errors on videos with occlusion, outperforming the compared methods.**

***Index Terms*—3D Tracking, Video, Attention, Fusion**

I. Introduction

Estimating accurate 3D human poses from 2D video remains a fundamental challenge in computer vision, especially under real-world conditions where 2D keypoint detections are often noisy, incomplete, or occluded [1], [2]. Existing 2D-to-3D lifting models typically assume clean input and focus on reconstructing plausible poses through spatial reasoning or temporal sequence modeling [3], [4]. Early temporal convolution approaches [3], [5] demonstrated that larger temporal receptive fields can improve motion consistency, while transformer-based methods such as PoseFormer [4], MixSTE [6], MotionBERT [7], and STRFormer [8] further strengthened long-range spatial-temporal reasoning through self-attention mechanisms. More recent approaches, such as FinePOSE [9], introduced semantic-aware pose representations to improve structural plausibility. However, despite these advances, existing systems often break down when confronted with corrupted 2D inputs, producing temporally unstable predictions.

Small perturbations in the 2D keypoints can cascade into severe motion jitter and structural collapse in the resulting 3D poses - an issue that remains unresolved by conventional lifting methods. One workaround has been to apply post-hoc motion refinement [10], [11]. SmoothNet [10] reduces jitter by filtering the predicted 3D sequences using velocity and acceleration priors. While effective for stabilization, such methods operate independently of the lifting process and assume that the 3D poses are already structurally sound without deformations. This architectural separation introduces latency and fails in scenarios where 2D keypoints are noisy to begin with. Similarly, transformer-based models [4], [6], [7] and diffusion-based generators [9], [12] offer strong spatial-temporal modeling but do not address motion refinement within the lifting pipeline. These approaches assume clean inputs and fail to correct hallucinated artifacts during inference.

This paper presents a Graph-based Attention Fusion Network (GAtFuN) for unified pose modeling, semantic prompting, and motion refinement. To the best of our knowledge, GAtFuN is the first 2D-to-3D lifting framework to jointly learn and integrate motion refinement directly within the 2D-to-3D pose estimation pipeline. Its architecture consists of a dual-stream design: one stream focuses on learning semantic-aware pose representations using transformer mechanisms, while the other stream performs motion refinement using graph attention to model and correct velocity transitions. These two streams are cross-fused at key stages, allowing the system to reason about motion dynamics and structural consistency jointly.

Our proposed method learns to denoise corrupted 2D keypoints, stabilize pose sequences, and enforce semantic coherence in a unified, end-to-end trainable model. Our contributions include the following: (1) a dual-stream 2D to 3D lifting architecture that fuses semantic-enriched transformer-based pose modeling with graph-based motion refinement; (2) a spatial graph attention module to refine the spatial relationships of joints using a skeletal adjacency matrix for motion consistency; (3) a temporal graph attention module that enforces the velocity of local joint transformations and prioritizes nearby frames for continuity.

The rest of this paper is organized as follows: Section II reviews related methods. Section III presents GAtFuN with a discussion of the network architecture and the fusion strategy. Section IV discusses our experimental results. Section V concludes this paper with a summary of our findings.

II. Related Work

Early methods improve joint-wise 3D estimation using local structure or short-term motion. Modulated GCN [13] introduces joint-specific feature transformations via weight modulation, enhancing relational accuracy between neighboring joints. Mobius-GCN [14] applies Möbius transformations to spectral graph convolutions to improve spatial reasoning on pose graphs. Temporal convolution-based methods [5], [15] model short-range temporal cues using dilated convolutions,

* University of North Texas, Denton, Texas, USA
† University of Puerto Rico, Rio Piedras, San Juan, Puerto Rico
‡ Hefei University of Technology, Anhui, China
§ Anhui Agricultural University, Anhui, China
This work is supported in part by NSF grants (#2424803 and #2617908)

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

filtering, and pooling, preserving temporal continuity in clean sequences. Anatomy-Aware [16] enforces bone-centric constraints to improve anatomically plausible joint predictions. These methods capture spatial or local temporal features, but are limited to short contexts and are vulnerable to noisy 2D inputs, which are common in real-world videos.

Transformer-based architectures expand the receptive field to capture long-range dependencies. PoseFormer [4], [17] introduces spatial-temporal attention but applies redundant sequence-to-frame inference. MixSTE [6] improves efficiency by alternating spatial and temporal transformers. MotionBERT [7] and STRFormer [8] use dual-stream attention and bidirectional temporal encoding to enhance temporal stability. Yet, these methods fail to integrate motion cues or semantic priors, making them vulnerable when tested in the wild.

Recent methods attempt to fuse global and local spatial-temporal modeling. HLTN [18] builds hierarchical representations over joint, limb, and body levels using non-uniform graph convolution and transformers. GLA-GCN [19] combines global GCN reasoning with localized joint-specific regression. MotionAGFormer [20] fuses attention and GCN in unified blocks. KTPFormer [21] injects anatomical and trajectory priors via specialized attention layers, and GraphMLP [22] proposes graph MLPs to model spatial relationships. While these architectures enhance representational capacity, they generally assume high-quality 2D input, leaving them fragile under noise or occlusion.

Attempts to refine motion during the 2D to 3D lifting, such as D3DP [12], introduce a denoising diffusion framework that generates structurally consistent 3D poses by reversing a noise process from clean 2D keypoints. While effective in capturing global pose structure, D3DP assumes distributional alignment between training and test inputs, which was not realistic in unconstrained environments. TMT [23] improves lifting by incorporating 3D velocity signals and motion segmentation during training. It builds a velocity-informed temporal adjacency matrix to model context-aware spatiotemporal dependencies. However, these motion features are only used during training, offering no inference-time correction of corrupted inputs.

Post-hoc temporal refinement methods like SmoothNet [10] and PoseFix [11] operate on 3D pose sequences after lifting. By modeling velocity and acceleration, SmoothNet [10] reduces jitter and enforces smoother trajectories. PoseFix [11] leverages natural language prompts to guide pose correction, aligning semantic descriptions of motion with 3D predictions through language-visual embeddings. While semantically richer than traditional filters, it remains a post-hoc module detached from the lifting process and depends on the availability of appropriate prompts and sufficiently interpretable motions. Both methods introduce inference latency and cannot recover structure or intent lost during 2D to 3D lifting.

III. Graph Attention Fusion Network

GAtFuN consists of interconnected, parallel streams (a pose modeling stream and a motion refinement stream) for 3D motion tracking from 2D keypoint sequences. As illustrated in Fig. 1, the architecture models both global joint dependencies and local motion dynamics by jointly reasoning over pose

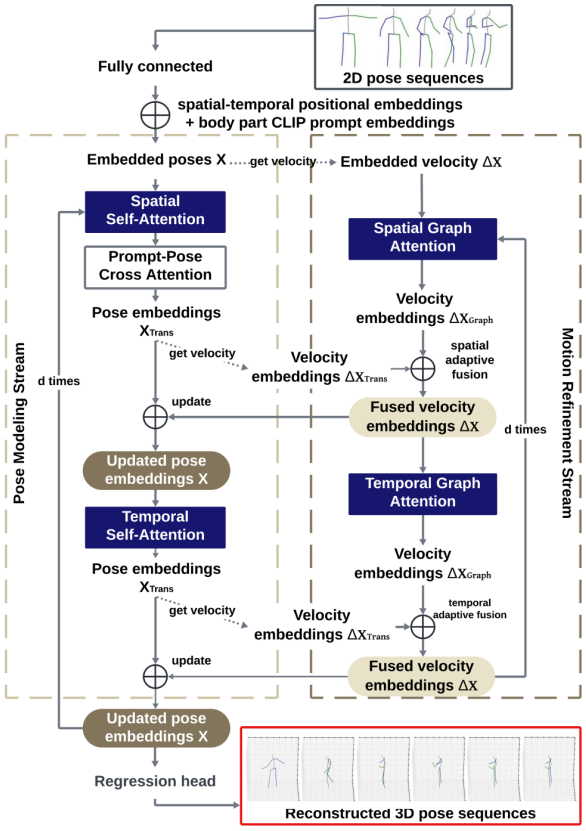


Fig. 1. The network architecture of GAtFuN.

and velocity. The input consists of 2D skeleton sequences $x \in \mathbb{R}^{B\times T\times N\times 2}$, where B is the batch size, T is the number of frames, and N is the number of joints. The projection to feature space of a dimension D produces initial pose embeddings $X \in \mathbb{R}^{B\times T\times N\times D}$. To preserve both the spatial structure and the temporal order, GAtFuN incorporates a spatial positional encoding $e^S_{\text{pos}} \in \mathbb{R}^{B\times 1\times N\times D}$ to encode joint-wise information, and a a temporal positional encoding $e^T_{\text{pos}} \in \mathbb{R}^{B\times T\times 1\times D}$ for frame-wise information. The encodings are added to X:

$$X \leftarrow X + e^S_{\text{pos}} + e^T_{\text{pos}}. \tag{1}$$

GAtFuN integrates contextual information by incorporating learnable text prompt embeddings P, using body-part and action-related textual cues (e.g., ”person”, ”action”, ”speed”, ”head”, ”body”, ”arm”, and ”leg”). A CLIP model [24] is pre-trained and each prompt token $\tilde{p}_k$ is augmented with a learnable residual r_k, allowing adaptation to the 3D pose task:

$$X \leftarrow X + P, \tag{2}$$

$$P = \{p_k\}_{k=1}^{K} \text{ and } p_k = \text{Concat}(r_k, \tilde{p}_k),$$

where $r_k \in \mathbb{R}^D$ allows the model to adapt the fixed CLIP embedding $\tilde{p}_k$ to the 3D pose estimation by injecting task-specific information, and K is the number of prompt tokens to construct the semantic prompt embedding. The pose modeling stream processes X through a Multi-Head Self-Attention (MHSA) module to capture global inter-joint relationships, followed by Prompt Learning (PL) [9] that modulates attention with textual

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

priors. It integrates a temporal MHSA module for the long-range dependencies.

The velocity embeddings ΔX, $\Delta X \in \mathbb{R}^{B\times T\times N\times D}$, is computed as frame-wise differences in pose embeddings:

$$\Delta X_{(t)} = X_{(t)} - X_{(t-1)}, \quad \text{for } t = 2, \ldots, T, \tag{3}$$

where $\Delta X_{(1)} := \Delta X_{(2)}$. ΔX is processed with a Spatial Graph Attention (SGA) module to ensure plausible joint relations. A Temporal Graph Attention (TGA) module is used to refine motion patterns. At each refinement step, velocity embeddings ΔX_{Trans} from the pose modeling stream are derived from inter-frame token differences, then fused with the velocity embeddings ΔX_{Graph} from the motion refinement stream via adaptive weighting.

GAtFuN progressively refine the pose representations. In the first encoder layer, it fuses the outputs of the spatial MHSA and PL of the pose modeling stream with those from the spatial graph attention module of the motion refinement stream before temporal modeling. In deeper layers, spatial and temporal modules from both streams are alternately stacked for the refinement of X using the fused motion ΔX, which generates 3D motions with a regression head.

A. Spatial and Temporal Topologies and Attentions

Fig. 2 shows the idea of the attention modules, which explicitly use velocity. Each joint is marked with a solid dot,

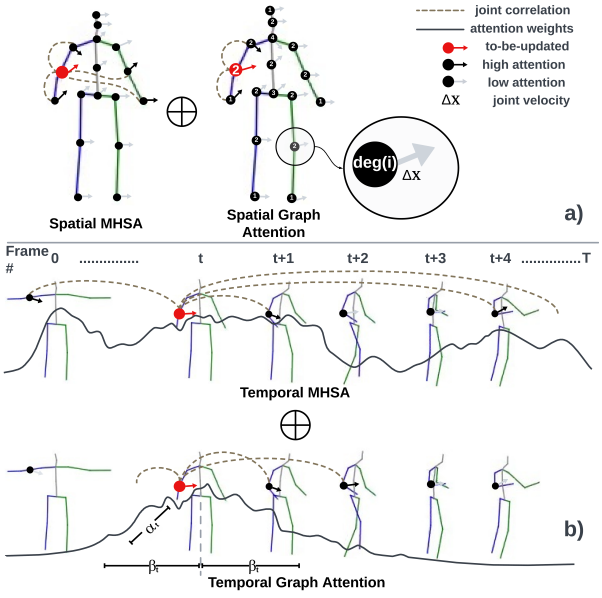


Fig. 2. a), Spatial fusion integrates velocity from the spatial MHSA and the spatial graph attention encoder. b), temporal fusion occurs between the temporal MHSA and the temporal graph attention encoder.

and the arrows depict its motion velocity. The red arrows highlight the velocity of the joint ΔX currently being updated, the black arrows indicate motions of joints of high attention, and the gray arrows represent motions of joints with low attention weights. The dashed lines illustrate the correlations among joints. The fusion operation is denoted with $\oplus$. Fig. 2(a) shows spatial MHSA (left) and spatial graph attention (right). The spatial MHSA uses global self-attention to capture joint dependencies. Our SGA, however, introduces a graph-based inductive affinity using a skeletal adjacency matrix. The attention is normalized using the degree of each joint, denoted by $\deg(i)$, that is, the number of directly connected joints (e.g., a wrist connects to an elbow, so the degree of the wrist is 1, whereas the neck has direct connections to four joints and hence its degree is 4). The spatial adjacency matrix prevents highly connected joints from dominating attention.

Fig. 2(b) shows the temporal MHSA (top) and temporal graph attention (bottom). The temporal MHSA models long-range joint dependencies using self-attention across frames. In contrast, TGA incorporates a self-attention-refined temporal adjacency matrix, which is modulated with a logistic decay function to attenuate weights for temporally distant frames. This allows the model to emphasize temporally local motions and, hence, prevents abrupt fluctuations that cause jitters.

B. Prompt-enriched Embeddings

Prompt-enriched features are integrated into both the pose and the motion refinement streams. These prompt features provide semantic priors, including contextual knowledge about joint identity, functional role, and motion intent. In the pose modeling, prompt embeddings are injected into the spatial transformer via a multi-head cross-attention (MHCA) module. These embeddings, generated from a CLIP model using body-part-related prompts (e.g., ”arm”, ”head”, ”leg”, ”action”), serve as structured priors that guide attention distribution across joint tokens. This disambiguates spatial relationships between joints. For instance, in cases of occlusion or overlapping joints (e.g., during seated or hand-to-head poses), the model infers anatomical plausibility by attending more strongly to relevant joints defined by prompt tokens.

The semantic cues are used to guide attention in SGA and TGA via dynamical weighing joint influence, which benefits sequences with missing or noisy joint velocities. For example, when a wrist is occluded, the model learns to prioritize motion from semantically associated joints besides nearest neighbors in the kinematic graph. Moreover, action-level prompts such as ”running” or ”turning” help the model suppress irrelevant noise (e.g., jitter in the spine) while enhancing critical dynamic cues (e.g., alternating foot strikes).

C. Pose Modeling Stream

In the pose modeling stream, a spatial MHSA encodes global spatial dependencies among joints in each frame, whereas a temporal MHSA models global temporal dependencies across frames for each joint. MHSA is expressed as

$$X_{\text{Trans}}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{4}$$

where W^O is the projection matrix and $\text{head}_i = \text{Softmax}\left(\frac{Q_i(K_i)^\top}{\sqrt{d_{\text{MHSA}}}}\right)V_i$, $Q_i = XW_i^Q$, $K_i = XW_i^K$, and $V_i = XW_i^V$ represent the query, key, and value for each attention head, respectively. W_i^Q, W_i^K, and W_i^V are learnable weights for each attention head. The dimensionality of the key vectors is $d_{\text{MHSA}} = D/h$, where D is the input dimension and h is the number of attention heads. The input of spatial MHSA is $X \in \mathbb{R}^{BT\times N\times D}$, and the input of temporal MHSA

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

is $X \in \mathbb{R}^{BN\times T\times D}$. The output is processed by an MLP, a residual connection, and Layer Normalization (LN).

A fine-grained PL is incorporated in the first encoder layer to enforce refinement of spatial and temporal pose features:

$$X_{\text{Prompt}}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_l)W^C, \quad (5)$$

where W^C is the projection matrix and $\text{head}_l = \text{Softmax}\left(\frac{Q_l(K_l)^\top}{\sqrt{d_{\text{MHCA}}}}\right)V_l$, $Q_l = X_{\text{Trans}}W_l^Q$, $K_l = PW_l^K$, and $V_l = PW_l^V$, with W_l^Q, W_l^K, and W_l^V as learnable. $d_{\text{MHCA}} = D/l$ is the dimensionality of the key vectors in cross-attention. An updated X_{Trans} is produced with LN from the output X_{Prompt}, followed by an MLP and a residual connection:

$$X_{\text{Trans}} \leftarrow \mathbf{G}(\mathbf{LN}(X_{\text{Prompt}})) + X_{\text{Trans}}, \quad (6)$$

where $\mathbf{G}$ is an MLP with GELU activation and dropout.

D. Motion Refinement

1) Spatial Graph Attention: Let $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$ denote the human skeleton, where $\mathcal{V}_s = \{1,2,\ldots,N\}$ are the joints, and $\mathcal{E}_s \subseteq \mathcal{V}_s \times \mathcal{V}_s$ are the connecting joints. An element of the spatial adjacency matrix A_s is computed as follows:

$$A_s(i,j) = \begin{cases} \frac{1}{\sqrt{\deg(i)\cdot\deg(j)}} & \text{if } i=j \text{ or } \overline{ij} \in \mathcal{E}_s, \\ 0 & \text{otherwise}, \end{cases} \quad (7)$$

where $\deg(i)$ is the degree of joint i. The aggregated spatial motion features are

$$\Delta\tilde{X}_s = \hat{A}_s\mathbf{V_s}(\Delta X_s), \quad (8)$$

where $\hat{A}_s = J^{-\frac{1}{2}}A_sJ^{-\frac{1}{2}}$, $\mathbf{V_s}$ is a linear transformation applied to $\Delta X_s \in \mathbb{R}^{BT\times N\times D}$, and $\hat{A}_s$ is the normalized adjacency matrix. J is a joint degree matrix and $J_{(ii)} = \sum_j A_{ij}$.

SGA leverages the skeletal structure with a spatial adjacency matrix that is aligned with each joint's degree. Following the idea of dominant and implicit joints [25], normalization based on joint degree ensures that highly connected joints, do not dominate the attention, while sparsely connected joints, e.g., wrist, contribute meaningfully.

2) Temporal Graph Attention: Let $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ represent a graph of joint trajectory, where $\mathcal{V}_t = \{1,2,\ldots,T\}$ denotes the joints and $\mathcal{E}_t \subseteq \mathcal{V}_t \times \mathcal{V}_t$ denotes the temporal connections. A temporal adjacency matrix A_t is constructed

$$A_t(i,j) = \frac{1}{1+\exp\left(\alpha_t(|i-j|-\beta_t)\right)}, \quad (9)$$

where $|i-j|$ is the distance between two joints. motion features $\Delta\tilde{X}_t$ is computed via self-attention:

$$\Delta\tilde{X}_t = (\hat{A}_t \odot \text{Softmax}(\Delta X_t \Delta X_t^\top))\mathbf{V_t}(\Delta X_t), \quad (10)$$

where $\hat{A}_t = J^{-\frac{1}{2}}A_tJ^{-\frac{1}{2}}$, $\mathbf{V_t}$ denotes linear transformation and $\odot$ denotes element-wise multiplication. The logistic decay prioritizes temporally closer frames while gradually diminishing the influence of distant ones. α_t determines the rate of decay. A small α_t allows attention weights to decay more gently, enabling frames in near distance to have a greater contribution. β_t defines the temporal range for the peak rate of attention influence before the decay initiates. A large β_t allows the model to consider a broad temporal attention window. This design promotes more consistent estimations across time by reflecting natural temporal correlations without over-smoothing joint movements.

The output of ΔX_{Graph} is obtained by combining the aggregated motion features with a residual connection

$$\Delta X_{\text{Graph}} = \Delta\tilde{X} + \mathbf{V}'(\Delta X), \quad (11)$$

where $\mathbf{V}'$ is another linear transformation. Similar to the pose modeling stream, ΔX_{Graph} is processed by an MLP, followed by a residual connection and LN.

E. Adaptive Motion Fusion

Motion representations ΔX_{Graph} are obtained from graph attention modules, while ΔX_{Trans} is derived from X_{Trans}, which is updated using Eq. (3). The refined motion representation ΔX is produced by integrating velocities

$$\Delta X = \gamma_1 \odot \Delta X_{\text{Trans}} + \gamma_2 \odot \Delta X_{\text{Graph}}, \quad (12)$$

where $\gamma_1, \gamma_2 = \text{Softmax}(\mathbf{W}(\text{Concat}(\Delta X_{\text{Trans}}, \Delta X_{\text{Graph}})))$ and $\mathbf{W}$ is a learnable linear transformation.

To ensure pose updates are aligned with the motion trajectory, the pose for a frame t is updated by adding the fused joint motion ΔX extracted by the pose modeling stream

$$X_{(t)} \leftarrow \begin{cases} X_{\text{Trans}}^{(1)}, & t=1, \\ X_{\text{Trans}}^{(t-1)} + \Delta X^{(t)}, & t=2,\ldots,T. \end{cases} \quad (13)$$

After iterations, a linear regression head is used to generate the final 3D output $Y \in \mathbb{R}^{B\times T\times N\times 3}$.

F. Training Loss

The loss function of our network is the weighted sum of Mean Per Joint Position Error loss L_W, Temporal Consistency Loss L_T, and Mean Per Joint Velocity Error L_V as follows:

$$Loss = L_W + \lambda_T L_T + \lambda_V L_V, \quad (14)$$

where L_W calculates the Euclidean distance to the ground truth while placing more emphasis on joints with larger errors [6], L_T encourages temporal coherence between frames [26], and L_V penalizes discrepancies in joint velocities [5]. λ_T and λ_V are scalar weights.

G. Datasets, Metrics and Settings

Our experiments are conducted using Human3.6M [27], MPI-INF-3DHP [28], and 3DPW [29]. Human3.6M includes 3D joint positions for each frame. MPI-INF-3DHP includes indoor and outdoor scenarios with varied clothing, lighting conditions, and occlusions. 3DPW captures in-the-wild challenges, e.g., occlusion, motion blur, and background clutter.

GAtFuN was trained in non-diffusion and diffusion-based settings using a batch size of 4. The weights λ_T and λ_V are 0.5 and 2.0, respectively. The joint weights w_n for L_W are adopted from MixSTE [6]. The temporal hyperparameter α_t is 0.125, and β_t is set to 5, 3 for $T = 243, 81$, respectively. K is 7 ("person", "action", "speed", "head", "body", "arm",

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

and "leg"). In the non-diffusion settings, GAtFuN was trained using an encoder depth of 8, attention heads of 8, and a channel dimension of 512. The initial learning rate is 6×10^{-5} with a decay of 0.995. In the diffusion settings, GAtFuN was trained with a single hypothesis and a single iteration, while testing with $H = 20$ and $I = 10$ to ensure a fair comparison. The initial learning rate remains 6×10^{-5} with a decay of 0.993. GAtFuN adopted AdamW optimizer with the momentum parameters as β_1, β_2 = 0.9, 0.999, and the weight decay as 0.1. The maximum number of timesteps is 1000.

IV. Experimental Results and Discussion

When Human3.6M was used, models were trained using subjects 1, 5, 6, 7, and 8, and subjects 9 and 11 for testing. For MPI-INF-3DHP, the training and test splits (TS1 to TS6) were adopted. A cross-dataset evaluation used models pretrained on Human3.6M (subjects 1, 5, 6, 7, and 8), where the 2D keypoints were extracted using a CPN [48]. The pretrained models were evaluated on the 60 videos of 3DPW, using 2D keypoints obtained from Detectron2 [49]. Our experiments are conducted using PyTorch on a GeForce RTX 4090 GPU.

A. Comparison of Motion Tracking

Table I reports the average MPJPE using Human3.6M. The methods are arranged in descending order according to Avg. The best performance is highlighted in bold-face fonts, and the second-best performance is underlined. Overall, GAtFuN achieved the best or highly competitive performance in all cases. The top two sections are results from using the non-diffusion settings. When CPN [48] was used, GAtFuN achieved the lowest MPJPE of 39.8. Although some methods attained smaller errors on a few actions, GAtFuN consistently achieved the smallest average errors across all. When finetuned SH [50] was used, GAtFuN again achieved the lowest MPJPE of 36.2. In the diffusion settings (the bottom section), GAtFuN achieved an MPJPE of 29.4, outperforming all compared methods. This shows a 7.8% reduction in error compared to FinePOSE [9] and a 4.5% reduction compared to MMCPose [47]. These indicate that integrating motion refinement into the lifting pipeline is more effective than relying solely on diffusion sampling or prompt-based enhancement.

Table II reports the average PMPJPE using Human3.6M. In the non-diffusion settings using CPN, GAtFuN achieved the best PMPJPE of 31.6mm, outperforming EHGFormer [38] (32.3mm) and DRPose [39] (32.4mm). Using SH, GAtFuN reduced the average PMPJPE to 30.0mm, which is superior to SasMamba [44] (33.6mm), LangPOSE [43] (31.3mm), PoseMamba [42] (31.5mm). In the diffusion settings, GAtFuN achieved a PMPJPE of 23.3mm, outperforming diffusion-based approaches, including FinePOSE [9] (25.0mm) and MMCPose [47] (23.7mm), for which GAtFuN reduced the error by approximately 6.8% and 1.7%, respectively.

Table III reports the average MPJPE on Human3.6M. In the non-diffusion settings, GAtFuN achieved an MPJPE of 14.5mm, outperforming PoseMamba [42] (14.8mm), SasMamba [44] (20.9mm), and DRPose [39] (20.7mm). The large improvement over transformer-based methods demonstrates the advantages of integrating graph-guided motion refinement into the lifting process. In the diffusion settings, GAtFuN achieved an MPJPE of 15.7mm. Compared to MMCPose [47], this is a 3.0% reduction in error. Notably, the diffusion-based GAtFuN reduced the error to nearly half of that obtained using detector-generated keypoints in Table I, indicating strong robustness under noisy detections with superior performance using clean motion trajectories. The gains of GAtFuN are not solely due to stronger detectors or diffusion sampling but arise from the unified modeling of semantic pose representations and motion-consistent graph refinement.

Table IV reports the MPJPE, PCK, and AUC using the MPI-INF-3DHP dataset. Ground truth keypoints were used as the input. In both diffusion and non-diffusion settings, GAtFuN exhibited superior performance compared to the state-of-the-art methods. GAtFuN achieved a PCK of 99.1%, AUC of 87.7%, and MPJPE of 14.1mm, which are much improved in contrast to TCPFormer [41], LangPose [43], and SasMamba [44]. In the case of diffusion settings, GAtFuN had a PCK of 99.3%, AUC of 83.1%, and MPJPE of 25.7mm, surpassing D3DP [12], FinePOSE [9], and DRPose [39].

Table V reports the errors using the 3DPW dataset. The models were trained with Human3.6M using a finetuned CPN [48] and SH [50] 2D keypoint detector, and tested on 3DPW using the off-the-shelf Detectron2 2D keypoint detector. GAtFuN demonstrated superior real-world performance in all cases. When SH was used as the keypoint detector in model training, GAtFuN struck a balance between spatial accuracy and temporal smoothness, while methods such as PoseMamba [42] achieved smooth motion trajectories by sacrificing accuracy. In the diffusion-based setting, GAtFuN led in spatial accuracy and effectively managed temporal inconsistencies. Even when prompt learning was excluded from our architecture, GAtFuN still outperformed FinePOSE [9] with much less parameter count in real-world scenarios.



Fig. 3. Intermediate results of D3DP, FinePOSE, and GAtFuN on Human3.6M. Red skeleton shows the ground truth.

Fig. 3 illustrates examples of intermediate poses of motion tracking of GAtFuN, as well as FinePOSE and D3DP using the Human3.6M dataset. The dashed circles highlight the improved regions by GAtFuN, which demonstrate precise joint positioning in actions such as 'Discussion', 'Posing', and 'Waiting', better capturing arm and leg extensions that align closely with ground truth. In self-occlusion-heavy actions like 'Directions'and 'Purchases', GAtFuN achieved accurate

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

TABLE I

AVERAGE MPJPE PER ACTION IN HUMAN3.6M. TOP AND MIDDLE: NON-DIFFUSION METHODS USING 2D INPUTS FROM FINETUNED CPN AND SH. BOTTOM SECTION: DIFFUSION METHODS (†) USING CPN INPUTS. ⋆, ‡: SEQUENCE LENGTH = 351, 729, RESPECTIVELY; OTHERWISE, 243.

Method	Det.	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD	Smoke	Wait	WalkD.	Walk	WalkT	Avg
PoseFormerv2 [17]	CPN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	45.2
GLA-GCN [19]	CPN	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
Strided Transformer [30]⋆	CPN	40.3	43.3	40.2	42.3	45.6	52.3	41.8	40.5	55.9	60.6	44.2	43.0	44.2	30.0	30.2	43.7
TC-MixSTE [31]	CPN	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
EHFusion [32]	CPN	37.7	42.6	39.0	40.0	44.6	53.1	41.1	39.0	53.4	59.6	43.8	40.7	42.0	29.3	29.8	42.4
P-STMO [33]	CPN	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
GLSTE [34]	CPN	38.1	41.7	39.1	39.5	43.7	49.7	40.4	39.7	54.2	58.7	42.8	42.1	42.0	28.7	28.5	41.9
MixSTE [6]	CPN	37.6	40.9	37.3	39.7	42.3	49.9	40.1	39.8	51.7	55.0	42.1	39.8	41.0	27.9	27.9	40.9
STRFormer [8]	CPN	37.8	40.2	36.8	38.3	42.0	50.2	38.6	39.2	51.9	56.1	41.9	39.3	41.0	27.7	28.1	40.6
STCFormer [35]	CPN	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
TMT [23]	CPN	37.3	40.5	37.6	40.4	40.6	50.4	38.5	39.6	51.8	53.8	41.9	38.8	41.2	28.1	27.5	40.5
Wang et al. [36]	CPN	36.9	40.2	36.8	39.4	41.7	49.8	39.4	39.3	51.1	54.6	41.6	39.1	40.3	27.3	27.6	40.3
Du et al. [37]‡	CPN	**36.1**	40.3	41.9	**33.3**	**35.7**	61.0	41.0	**33.6**	**37.4**	69.9	**38.4**	**36.7**	43.4	25.6	29.4	40.2
HLTN [18]	CPN	38.6	**38.8**	36.0	38.6	42.3	49.9	38.8	37.8	51.6	55.5	41.6	38.5	40.2	27.6	27.6	40.2
KTPFormer [21]	CPN	37.3	39.2	**35.9**	37.6	42.5	48.2	38.6	39.0	51.4	55.9	41.6	39.0	**40.0**	27.0	27.4	40.1
EHGFormer [38]	CPN	36.7	39.3	36.7	38.3	43.2	**45.9**	40.5	36.8	49.4	55.6	41.5	39.5	41.0	27.8	27.7	40.0
DRPose [39]	CPN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.9
GAtFuN	CPN	37.2	39.5	36.1	37.9	41.7	46.9	38.7	38.1	51.3	**54.5**	41.4	39.0	40.1	26.6	28.3	**39.8**
MotionBERT [7]	SH	36.3	38.7	38.6	33.6	42.1	50.1	36.2	35.7	50.1	56.6	41.3	37.4	37.7	25.6	26.5	39.2
RePOSE [40]	SH	34.6	36.8	**35.0**	**31.3**	38.8	**44.7**	**35.3**	35.5	47.5	50.9	38.4	**35.5**	34.9	**23.9**	25.0	36.5
EHGFormer [38]	SH	**34.1**	38.0	38.2	32.4	40.6	47.8	37.3	34.3	50.2	56.1	39.4	36.7	**26.3**	35.8	25.8	38.2
TCPFormer [41]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.9
PoseMamba [42]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37.1
LangPOSE [43]	SH	35.6	37.5	35.3	31.7	40.1	45.4	36.5	33.4	46.2	49.3	39.0	36.0	34.4	24.6	25.1	36.7
SasMamba [44]	SH	38.4	41.0	38.0	34.6	41.3	49.3	38.8	36.1	50.7	56.2	41.6	38.4	37.2	27.2	27.6	39.8
GAtFuN	SH	**34.1**	**36.2**	35.8	32.9	**36.8**	45.2	35.8	**33.5**	**46.3**	**49.8**	**37.5**	35.6	34.9	24.2	**24.1**	**36.2**
Diff3DHPE [45]†	CPN	37.3	40.6	36.3	38.0	41.8	46.6	38.3	39.4	51.3	53.0	42.1	39.5	40.7	27.3	28.2	40.0
DiffPOSE [46]†	CPN	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 [12]†	CPN	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
KTPFormer [21]†	CPN	30.1	32.1	29.1	30.6	35.4	39.3	32.8	30.9	43.1	45.5	34.7	33.2	32.7	22.1	23.0	33.0
FinePOSE [9]†	CPN	31.4	31.5	28.8	29.7	34.3	36.5	29.2	30.0	42.0	42.5	33.3	31.9	31.4	22.6	22.7	31.9
MMCPose [47]†	CPN	28.4	30.1	28.2	28.2	32.8	36.1	28.0	29.2	40.2	40.9	32.0	30.1	30.9	23.2	23.3	30.8
GAtFuN†	CPN	**26.6**	**28.8**	**26.4**	**27.0**	**32.0**	**34.9**	**27.9**	**27.9**	**40.6**	**38.5**	**30.8**	**29.5**	**29.2**	**20.0**	**20.6**	**29.4**

TABLE II

AVERAGE PMPJPE PER ACTION ON HUMAN3.6M.

Method	Det.	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
PoseFormerv2 [17]	CPN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	35.6
Strided Transformer [30]⋆	CPN	32.7	35.5	32.5	35.4	35.9	41.6	33.0	31.9	45.1	50.1	36.3	33.5	35.1	23.9	25.0	35.2
GLA-GCN [19]	CPN	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
EHFusion [32]	CPN	32.1	35.2	32.6	34.0	35.1	41.5	32.6	31.2	42.9	48.9	36.6	32.0	34.0	24.1	25.3	34.6
P-STMO [33]	CPN	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
GLSTE [34]	CPN	30.8	34.6	32.4	32.9	34.0	39.5	31.5	31.0	44.2	48.7	35.2	32.4	33.5	23.1	23.5	33.8
MixSTE [6]	CPN	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
STRFormer [8]	CPN	30.7	32.9	30.5	31.4	32.7	38.9	30.2	30.4	41.8	45.5	33.8	30.6	32.6	21.6	22.7	32.4
TMT [23]	CPN	30.6	32.8	30.1	31.8	33.2	38.9	31.4	30.3	41.8	**43.7**	33.9	30.5	33.0	21.8	**22.2**	32.4
EHGFormer [38]	CPN	30.7	32.7	30.4	31.6	34.1	**36.2**	31.8	29.2	40.1	45.5	34.2	30.8	32.1	22.3	23.1	32.3
Wang et al. [36]	CPN	30.2	32.5	29.8	31.0	32.5	38.8	30.4	30.3	41.0	43.9	33.7	30.3	33.0	21.2	**22.2**	32.1
HLTN [18]	CPN	40.6	**31.8**	29.4	30.9	32.8	38.3	30.2	29.0	41.1	44.7	33.6	30.0	31.8	21.7	22.3	31.9
KTPFormer [21]	CPN	30.1	32.3	29.6	30.8	32.3	37.3	30.0	30.2	41.0	45.3	33.6	29.9	31.4	21.5	22.6	31.9
TC-MixSTE [31]	CPN	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
STCFormer [35]	CPN	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
Du et al. [37]‡	CPN	**26.7**	34.1	33.2	27.6	**27.5**	47.0	**28.5**	**27.8**	**28.9**	57.9	**30.8**	**25.9**	35.4	21.0	24.9	31.8
DRPose [39]	CPN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.4
GAtFuN	CPN	29.7	**31.8**	29.4	**30.3**	31.0	36.9	31.0	29.9	42.0	**43.7**	31.6	30.8	**30.5**	21.8	23.3	**31.6**
MotionBERT [7]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32.9
RePOSE [40]	SH	29.1	30.9	**29.4**	26.6	32.1	35.8	**28.6**	30.6	39.9	44.9	33.3	**29.0**	29.6	**20.1**	**21.2**	30.7
EHGFormer [38]	SH	29.1	31.9	32.4	27.5	32.9	38.1	30.5	29.5	43.6	49.7	34.4	30.2	**22.3**	30.7	22.5	32.4
TCPFormer [41]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.7
PoseMamba [42]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.5
LangPOSE [43]	SH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.3
SasMamba [44]	SH	31.9	34.0	32.4	29.8	34.7	39.1	31.7	31.3	43.2	49.5	36.3	31.1	32.2	23.5	23.9	33.6
GAtFuN	SH	**28.8**	**30.1**	29.5	**24.8**	**29.4**	**34.1**	29.5	**29.2**	**38.3**	**43.8**	**30.4**	30.3	29.2	20.4	22.7	**30.0**
Diff3DHPE [45]†	CPN	29.8	33.0	29.2	30.8	32.0	36.7	29.5	30.0	40.4	41.6	33.8	30.6	32.2	22.0	23.1	31.6
D3DP [12]†	CPN	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 [21]†	CPN	24.1	26.7	24.2	24.9	27.3	30.6	25.2	23.4	34.1	35.9	28.1	25.3	25.9	17.8	18.8	26.2
FinePOSE [9]†	CPN	24.3	25.7	23.8	24.1	26.3	28.0	22.4	23.1	33.0	33.5	26.5	24.4	24.1	17.7	18.6	25.0
MMCPose [47]	CPN	**21.8**	24.2	22.2	22.8	25.1	**27.0**	**20.9**	**20.8**	31.8	32.3	25.6	**22.5**	23.3	17.1	17.9	23.7
GAtFuN†	CPN	21.9	**23.7**	**21.2**	**22.5**	**24.5**	27.3	22.0	21.5	**31.4**	**30.5**	**24.7**	22.6	**22.6**	**15.9**	**17.1**	**23.3**

reconstruction of the hidden body parts. In postures such as 'Smoking', where knee placement and hand-to-mouth alignment are subtle but critical, GAtFuN produced exceptional pose tracking by maintaining structural consistency.

Fig. 4 illustrates examples of intermediate poses of motion tracking of GAtFuN, FinePOSE and D3DP using a 3DPW video. The top row shows the video frames with ground truth. Temporal inconsistencies are clearly visible in the results of D3DP and FinePOSE, with erratic joint movements and abrupt pose shifts that result in body deformations and limb misalignment. GAtFuN maintained superior limb alignment, particularly during mid-stride or directional changes.

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

TABLE III
AVERAGE MPJPE PER ACTION OF METHODS ON THE HUMAN3.6M USING 2D GROUND TRUTH KEYPOINTS. TOP: NON-DIFFUSION-BASED. BOTTOM: DIFFUSION-BASED (†). ⋆: SEQUENCE LENGTH IS 351; OTHERWISE, 243.

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD	Smoke	Wait	WalkD.	Walk	WalkT	Avg
P-STMO [33]	28.5	30.1	28.6	27.9	29.8	33.2	31.3	27.8	36.0	37.4	29.7	29.5	28.1	21.0	21.0	29.3
Strided Transformer [30] ⋆	27.1	29.4	26.5	27.1	28.6	33.0	30.7	26.8	38.2	34.7	29.1	29.8	26.8	19.1	19.8	28.5
GLSTE [34]	27.8	29.5	27.1	26.5	27.4	31.0	30.8	27.5	31.9	35.0	28.2	28.5	27.0	20.3	20.3	27.9
EHFusion [32]	25.0	25.8	25.3	25.3	28.6	33.6	28.3	23.9	32.3	31.7	27.3	25.5	24.2	20.0	20.5	26.5
MixSTE [6]	21.6	22.0	20.4	21.0	20.8	24.3	24.7	21.9	26.9	24.9	21.2	21.5	20.8	14.7	15.7	21.6
STCFormer [35]	20.8	21.8	20.0	20.6	23.4	25.0	23.6	19.3	27.8	26.1	21.6	20.6	19.5	14.3	15.1	21.3
TC-MixSTE [31]	21.4	21.8	20.3	20.1	20.1	23.5	24.2	21.6	26.8	24.5	21.0	21.2	20.7	14.7	15.3	21.1
GLA-GCN [19]	20.1	21.2	20.0	19.6	21.5	26.7	23.3	19.8	27.0	29.4	20.8	20.1	19.2	12.8	13.8	21.0
Wang et al. [36]	20.2	21.3	19.6	20.1	19.7	23.0	24.2	20.5	25.3	24.1	20.0	20.5	19.9	13.7	14.3	20.4
HLTN [18]	18.9	19.2	18.9	18.2	19.0	21.7	20.3	19.2	23.3	24.5	20.1	17.2	18.2	12.7	14.1	19.0
KTPFormer [21]	19.6	18.6	18.5	18.1	18.7	22.1	20.8	18.3	22.8	22.4	18.8	18.1	18.4	13.9	15.2	19.0
TMT [23]	18.7	18.2	18.7	18.1	18.8	21.5	19.8	18.3	23.8	23.4	18.8	17.0	17.7	12.7	13.9	18.6
EHGFormer [38]	18.3	18.5	17.0	17.9	18.8	19.4	19.7	17.5	22.2	21.8	18.4	16.9	17.3	12.9	14.0	18.0
MotionBERT [7]	16.7	19.9	17.1	16.5	17.4	18.8	19.3	20.5	24.0	22.1	18.6	16.8	16.7	10.8	11.5	17.8
RePOSE [40]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15.7
TCPFormer [41]	15.0	15.9	16.1	14.2	15.7	16.4	16.2	16.9	22.1	20.6	16.7	**13.3**	**13.8**	**9.2**	9.9	15.5
PoseMamba [42]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14.8
SasMamba [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.9
DRPose [39]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.7
GAtFuN	**13.6**	**14.5**	**14.8**	**13.4**	**14.2**	**14.2**	**16.1**	**14.6**	**18.9**	**19.1**	**16.3**	14.4	14.5	10.1	**9.1**	**14.5**
Diff3DHPE [45]†	20.4	20.8	19.4	19.3	20.5	21.6	21.6	20.9	24.2	23.3	21.1	19.3	19.4	14.5	16.1	20.2
D3DP [12]†	18.7	18.2	18.4	17.8	18.6	20.9	20.2	17.7	23.8	21.8	18.5	17.4	17.4	13.1	13.6	18.4
KTPFormer [21]†	18.8	17.4	18.1	17.7	18.3	20.6	19.6	17.7	23.3	22.0	18.7	17.0	16.8	12.4	13.5	18.1
FinePOSE [9]†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	16.7
MMCPose [47]†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	16.2
GAtFuN†	**16.1**	**14.2**	**15.1**	**14.7**	**16.2**	**17.5**	**16.6**	**15.5**	**19.8**	**19.4**	**15.9**	**14.8**	**15.3**	**11.4**	**12.8**	**15.7**

TABLE IV
COMPARISON ON MPI-INF-3DHP. ⋆, ‡: SEQUENCE LENGTH IS 27, 243, RESPECTIVELY; OTHERWISE, 81.

	Method	PCK ↑	AUC ↑	MPJPE ↓
Non-diffusion	P-STMO [33]	97.9	75.8	32.2
	GLA-GCN [19]	98.5	79.1	27.8
	GLSTE [34]	98.2	79.0	27.4
	PoseFormerv2 [17]	97.9	78.8	27.8
	STCFormer [35]	98.7	83.9	23.1
	TC-MixSTE [31]	98.7	79.5	27.6
	HLTN [18]⋆	98.2	82.3	22.1
	KTPFormer [21]	98.9	85.9	16.7
	RePOSE [40]	98.3	86.7	15.5
	TSwinPose [51]	98.6	84.3	24.9
	TMT [23] ⋆	-	-	48.3
	Wang et al. [36]	-	-	31.8
	EHGFormer [38] ⋆	**99.4**	75.4	35.9
	PoseMamba [42]	-	-	14.5
	TCPFormer [41]	99.0	**87.7**	15.0
	LangPose [43]	98.9	87.0	15.5
	SasMamba [44]	-	-	17.9
	DRPose [39]	95.7	66.0	47.5
	GAtFuN	99.1	**87.7**	**14.1**
Diffusion	D3DP [12]†‡	98.0	79.1	28.1
	DiffPose [46]†	98.0	75.9	29.1
	FinePOSE [9]†‡	98.9	80.0	26.2
	DRPose [39]†‡	98.5	79.1	28.1
	GAtFuN†‡	**99.3**	**83.1**	**25.7**

TABLE V
COMPARISON ON 3DPW (DETECTRON2 FOR KEYPOINT DETECTION AT INFERENCE).

	Method	DET	PMPJPE	MPJVE	AE	Param
Non-diffusion	GLA-GCN [19]	CPN	125.5	22.4	26.5	**1.30**
	PoseFormerv2 [17]	CPN	107.2	11.4	8.3	14.30
	STCFormer [35] ⋆	CPN	107.5	21.2	32.1	18.91
	KTPFormer [21]	CPN	102.5	10.2	8.3	33.65
	GAtFuN (w/o PL)	CPN	97.2	9.2	7.1	58.57
	GAtFuN	CPN	**95.4**	**8.4**	**6.3**	224.3
	MotionBERT [7]	SH	119.1	13.4	14.1	42.50
	MotionAGFormer [20]	SH	144.0	14.1	14.4	19.0
	RePOSE [40]	SH	118.2	11.2	9.7	**16.0**
	PoseMamba [42]	SH	127.7	9.2	**5.1**	26.50
	TCPFormer [41]	SH	141.1	14.2	14.7	35.10
	GAtFuN (w/o PL)	SH	101.2	8.9	7.1	58.57
	GAtFuN	SH	**96.9**	**7.8**	6.7	224.3
Diff.	D3DP [12]†	CPN	104.5	16.3	21.8	**34.84**
	FinePOSE [9]†	CPN	107.8	19.4	27.7	200.6
	GAtFuN (w/o PL)†	CPN	97.9	8.8	6.8	60.07
	GAtFuN†	CPN	**95.5**	**8.2**	**6.1**	225.8

B. Robustness to Occlusion and Noise

To evaluate the impact of occlusion on GAtFuN, synthetic occlusions are generated by placing a circular disk to block the human body. By changing the size of this disk and the number of frames with occlusion, we simulate spatial and temporal occlusions, respectively. The diameter of the disk varies with respect to the length of the diagonal line of the average human bounding box. This ensures the proportion of the occlusion. Four degrees of spatial occlusion were simulated. In degree one, the diameter of the disk is one-fifth (1/5) of the length of the diagonal line of the average bounding box. As we increase the size of the disk to be two-fifths or more, we have a degree of two or more, which introduces a more severe spatial occlusion. The disk was concentric with the bounding box of the human in the middle frame of the sequence. To simulate the temporal occlusion, we set four degrees of occlusion. The ratio of the number of frames including the disk and the total number of frames is used as the degree of occlusion. For example, degree one corresponds to a synthetic occlusion appearing in 49 consecutive frames in a video of 243 frames. The frames with synthetic occlusion were temporally centered.

Table VI reports the MPJPE for five actions under different of spatial and temporal occlusions. Across all cases, GAtFuN consistently achieves the lowest MPJPE compared with D3DP and FinePOSE, demonstrating stronger robustness to both spatially missing joints and temporally unavailable observations. Moreover, the performance advantage of GAtFuN is maintained, and often becomes more evident, as the occlusion severity increases. Without occlusion (S=0, D=0), GAtFuN

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

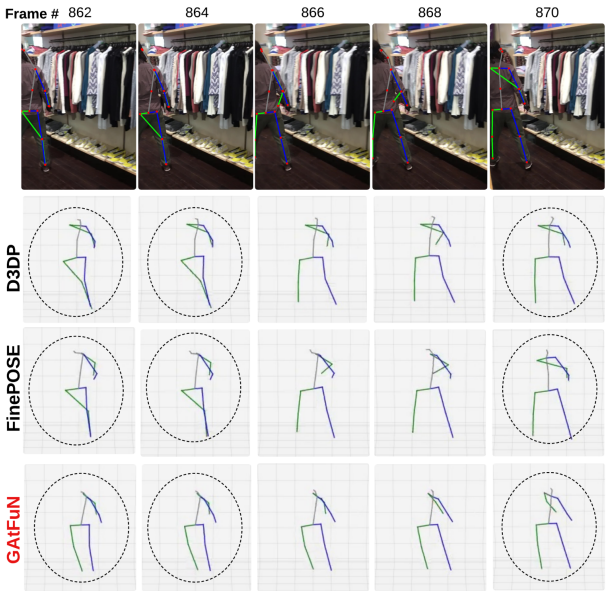


Fig. 4. Results of GAtFuN on the downtown_enterShop_00 video of 3DPW (keypoints extracted using Detectron2).

TABLE VI
AVERAGE MPJPE ON HUMAN3.6M WITH SPATIAL (S) AND TEMPORAL (D) OCCLUSIONS.

S	D	Method	Dir.	Disc.	Eat	Greet	Phone	Avg
0	0	D3DP	60.1	72.5	67.8	65.7	85.6	70.3
		FinePOSE	59.5	69.4	63.1	63.6	83.1	67.7
		GAtFuN	**54.8**	**65.2**	**58.9**	**60.4**	**71.3**	**62.1**
1	1	D3DP	63.6	78.3	69.9	68.3	88.7	73.8
		FinePOSE	62.4	73.9	64.7	65.8	84.8	70.3
		GAtFuN	**55.4**	**68.1**	**60.7**	**61.4**	**73.1**	**63.7**
	2	D3DP	64.3	82.1	69.3	69.5	92.5	75.5
		FinePOSE	63.5	77.9	65.3	68.1	85.7	72.1
		GAtFuN	**58.4**	**70.2**	**59.2**	**63.5**	**74.6**	**65.2**
	3	D3DP	65.2	85.3	69.5	72.4	95.2	77.5
		FinePOSE	64.2	83.1	65.8	69.5	87.8	74.1
		GAtFuN	**60.4**	**76.9**	**61.4**	**64.2**	**79.2**	**68.4**
	4	D3DP	66.9	87.1	71.2	73.8	96.9	79.2
		FinePOSE	65.2	84.2	66.9	71.2	89.4	75.4
		GAtFuN	**61.1**	**77.4**	**62.4**	**66.1**	**80.7**	**69.5**
2	1	D3DP	68.3	81.4	71.8	68.1	90.7	76.1
		FinePOSE	67.9	79.2	67.8	66.0	86.9	73.6
		GAtFuN	**63.7**	**70.4**	**62.6**	**62.0**	**73.8**	**66.5**
	2	D3DP	72.4	86.1	73.7	73.2	94.1	79.9
		FinePOSE	71.8	82.3	69.4	69.9	90.5	76.8
		GAtFuN	**59.7**	**73.8**	**64.7**	**63.9**	**78.2**	**68.1**
	3	D3DP	74.2	92.1	77.4	77.6	99.8	84.2
		FinePOSE	73.7	89.6	74.6	76.2	93.8	81.6
		GAtFuN	**65.4**	**79.0**	**69.3**	**68.7**	**83.1**	**73.1**
	4	D3DP	82.1	101.4	84.2	83.8	106.7	91.6
		FinePOSE	80.4	96.8	80.7	80.1	101.3	87.9
		GAtFuN	**71.9**	**85.6**	**74.0**	**73.1**	**92.3**	**79.4**
3	1	D3DP	117.2	102.6	113.2	83.8	105.4	104.4
		FinePOSE	113.6	97.9	110.7	78.9	101.1	100.4
		GAtFuN	**104.2**	**93.1**	**105.1**	**74.8**	**87.2**	**92.9**
	2	D3DP	127.4	120.5	140.4	113.9	121.1	124.7
		FinePOSE	119.3	119.0	137.8	112.9	115.6	121.0
		GAtFuN	**108.7**	**105.9**	**121.8**	**103.6**	**102.7**	**108.5**
	3	D3DP	156.1	145.1	169.1	140.5	123.9	146.9
		FinePOSE	157.7	135.8	171.5	123.6	121.6	142.0
		GAtFuN	**145.1**	**123.7**	**151.2**	**116.1**	**108.4**	**128.9**
	4	D3DP	179.4	173.4	188.5	176.2	163.6	176.2
		FinePOSE	165.8	154.2	183.1	147.9	146.7	159.5
		GAtFuN	**151.1**	**142.3**	**167.4**	**128.9**	**120.7**	**142.1**
4	1	D3DP	142.3	126.6	137.4	108.4	131.1	129.2
		FinePOSE	141.4	115.9	128.6	102.3	123.4	122.3
		GAtFuN	**127.3**	**108.8**	**114.1**	**94.1**	**109.3**	**110.7**
	2	D3DP	163.7	154.9	176.2	147.2	157.3	159.9
		FinePOSE	159.2	138.4	158.3	134.8	142.7	146.7
		GAtFuN	**148.2**	**129.8**	**142.6**	**124.9**	**127.3**	**134.6**
	3	D3DP	201.6	189.7	205.0	197.2	177.7	194.2
		FinePOSE	189.4	168.6	182.3	168.9	160.0	173.8
		GAtFuN	**170.3**	**154.7**	**164.2**	**149.5**	**148.4**	**157.4**
	4	D3DP	246.3	221.4	249.7	236.7	208.4	232.5
		FinePOSE	224.8	199.4	217.4	208.1	194.2	208.8
		GAtFuN	**205.3**	**182.2**	**202.1**	**180.5**	**179.4**	**189.9**

achieves an MPJPE of 62.1 mm, outperforming FinePOSE [9] and D3DP [33]. As the occlusion level increases, performance degrades across all methods. However, the increase in error of GAtFuN is consistently smaller. For example, under moderate occlusion (S=1, D=2), GAtFuN achieves an average MPJPE of 65.2 mm, compared with 72.1 mm for FinePOSE [9] and 75.5 mm for D3DP [12]. Under stronger occlusions (S=2, D=4), GAtFuN has an average MPJPE of 79.4 mm, while FinePOSE [9] and D3DP [12] degrade to 87.9 mm and 91.6 mm, respectively. Even in the most challenging setting (S=4, D=4), where both severe spatial occlusion and prolonged temporal gaps are present, GAtFuN yields the lowest MPJPE of 189.9 mm, compared with 208.8 mm for FinePOSE [9] and 232.5 mm for D3DP [12]. These indicate that GAtFuN degrades gracefully under increasing occlusion severity and is able to recover reliable poses from incomplete observations.

A similar trend is observed across individual actions. For example, "Phone" action involves hand-to-face interactions and self-occlusions that make pose estimation particularly challenging. Under the clean setting (S=0, D=0), GAtFuN reduces the MPJPE from 83.1 mm with FinePOSE [9] to 71.3 mm, yielding a 14.2% improvement. Significant gains (10%) remain under heavy occlusion, which highlights the effectiveness of the proposed motion refinement stream and graph-guided temporal reasoning in handling ambiguous. These demonstrate that the proposed method improves accuracy under clean conditions and substantially enhances robustness in the presence of severe spatial and temporal occlusions.

TABLE VII
PERFORMANCE UNDER DIFFERENT LEVELS OF NOISE.

	Method	Noise Variance					Degradation (ΔMPJPE)				Ave
		0	5	10	15	20	5-0	10-5	15-10	20-15	Deg
Non	MHFormer	35.8	38.5	41.6	44.2	47.9	2.7	3.1	2.6	3.7	3.03
	GAtFuN	**14.5**	**15.7**	**18.2**	**20.8**	**23.4**	**1.2**	**2.5**	**2.6**	**2.6**	**2.23**
Diff.	D3DP	18.4	19.1	23.6	26.1	33.2	0.7	4.5	2.5	7.1	3.70
	FinePOSE	16.7	18.0	21.8	24.7	29.8	1.3	3.8	2.9	5.1	3.28
	GAtFuN	**15.7**	**16.3**	**17.4**	**19.3**	**22.4**	**0.6**	**1.1**	**1.9**	**3.1**	**1.68**

Table VII reports the MPJPE when noise is introduced to the ground truth as inputs. "Noise Variance" reports the MPJPE when Gaussian noise of different variances is introduced. "Degradation" is the difference between adjacent noise variances, which shows the magnitude of impact when a greater amount of noise is introduced. "Ave Deg" reports the average degradation as noise variance increases. In all cases, GAtFuN exhibits the smallest degradation when a greater amount of noise is applied. This demonstrates a greater robustness to noise, i.e., less susceptible to occlusion.

C. Ablation Study

Experiments were conducted using Human3.6M in our ablation study. CPN [48] was used to detect keypoints. Table VIII reports the performance when different modules were included. The baseline model, with the spatial and temporal MHSA encoders, achieved an MPJPE of 36.3mm with the lowest MAC. Introducing PL reduces the MPJPE to 31.9mm, and when the pose representation was integrated with motion representation, GAtFuN delivered the best MPJPE of 29.4mm. When only SGA was used, the performance dropped by about 3%, which demonstrates the benefit of modeling dynamic joint relationships through graph reasoning. Among all modules,

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

PL increases the model size the most. The SGA and TGA introduce less than 50% of extra model size to the baseline.

TABLE VIII
PERFORMANCE OF INCLUDING VARIOUS MODULES.

S- + T-MHSA	PL	S-Graph	T-Graph	S-Graph	T-Graph	MPJPE	Params	MAC
✓						36.3	**34.84**	**0.59**
✓	✓					31.9	200.6	0.62
✓	✓	✓				31.4	213.2	0.83
✓	✓	✓	✓			30.8	225.8	1.09
✓	✓			✓		30.4	213.2	0.83
✓				✓	✓	32.9	60.07	1.06
✓	✓			✓	✓	**29.4**	225.8	1.09

To assess the utility of the adjacency matrix, the spatial (A_s) and temporal (A_t) adjacency matrices were replaced with matrices with equal weights: $A_s(i,j)$, $A_t(i,j)$ are 1 for $i = j$ or $\bar{ij} \in \mathcal{E}_*$ ($\mathcal{E}_*$ can be $\mathcal{E}_s$ or $\mathcal{E}_t$), and 0 otherwise. Table IX reports the MPJPE of GAtFuN using different adjacency matrices. The lowest errors were obtained when a normalized spatial adjacency matrix with node degrees and a normalized temporal adjacency matrix with a decay function were used. This demonstrates improvements to the uniform adjacent matrix. Additionally, TGA was compared to the case where the temporal graph retains only the k-nearest neighbors (KNN) as connected nodes after self-attention. While the KNN-based approach enables selective frame connections, it relies on a fixed number of neighbors, which restricts the adaptability to local temporal patterns for motion refinement, which is confirmed by the greater error at 31.4mm in the table.

TABLE IX
MPJPE OF GATFUN USING DIFFERENT ADJACENCY MATRICES.

unif. S-adj. mat. eq. weig.	norm. S-adj. mat. deg.	unif. T-adj. mat. eq. weig.	norm. T-adj. mat. decay	KNN of T-self-attn	MPJPE (CPN)	MPJPE (SH)
✓		✓			31.5	38.8
✓			✓		29.6	36.7
	✓	✓			31.0	37.3
	✓			✓	31.4	37.7
	✓		✓		**29.4**	**36.2**

TABLE X
MPJPE OF DIFFERENT PROMPT CONFIGURATIONS

Prompt Style	Non-diffusion			Diffusion		
	MHSA	GAtFuN	Imp.	MHSA	GAtFuN	Imp.
No Prompt	43.9	42.6	3.0%	37.6	32.9	12.5%
Prompt-M	42.0	40.5	3.5%	35.8	30.8	14.0%
Prompt-S	42.1	41.0	2.8%	36.4	30.9	15.1%
Prompt-C	41.8	40.4	3.3%	34.0	30.9	9.1%
Prompt-Al	41.9	40.1	4.3%	34.6	30.8	11.0%
Fine-grained	41.0	39.8	2.9%	25.0	23.3	6.8%

D. Prompt Styles

Table X presents the results of an ablation study of prompt. The Spatial-Temporal MHSA transformer is compared with GAtFuN under identical prompt settings. While MHSA uses spatial and temporal self-attention, GAtFuN additionally incorporates Spatial Graph Attention (SGA) and Temporal Graph Attention (TGA) for motion-aware structural refinement. GAtFuN outperforms MHSA in all prompt styles. Under diffusion without prompts (i.e., No Prompt), GAtFuN reduces MPJPE from 37.6 mm to 32.9 mm (12.5% improvement), which demonstrates a significant gain from the graph-attention modules. Also, performance improvements are substantial across all styles, ranging from 6.8% to 15.1% under diffusion settings. The largest gain (15.1%) is achieved with the semi-learnable prompt (Prompt-S), which demonstrates the synergy between semantic guidance and graph-based refinement. When using the fine-grained prompt (a prompt-driven framework built around CLIP semantics), GAtFuN improves MPJPE by 6.8%. CLIP-based prompts and the integration of SGA and TGA modules jointly contribute to the improved performance of GAtFuN by integrating complementary semantic information and spatial-temporal components.

E. Network Design

Table XI presents our results of using different network architectures. The experiments were conducted on Human3.6M with CPN's 2D inputs. $\rightarrow$ denotes the process flow and $\oplus$ denotes adaptive fusion. Parentheses show the processing order. For example, network architecture 1 shows a two-stage processing structure. The input is processed in parallel by a spatial self-attention module (S-MHSA) (with pose feature X_{Trans} inputs) and SGA (with velocity features ΔX_{Graph}). The fusion of S-MHSA and SGA are processed with the temporal modules, in which features are processed by temporal MHSA followed by PL MHCA. The results of PL are fused with the results of the TGA.

As shown in the table, the placement of spatial and temporal modules has an impact on the performance. The network architecture 8 exhibited the best performance. Compared to the second best, the reductions of MPJPE and PMPJPE are 4.9% and 2%, respectively. The network performs better when PL is placed in the early stage, e.g., cases 4, 5, 6, and 8, which suggests that PL in a later stage limits its ability for feature encoding. In GAtFuN, spatial and temporal processing are distinctly separated, and adaptive fusion is applied to derive a rich yet specific set of features. This allows learning disentangled representations before cross-domain interaction.

TABLE XI
TRACKING ERROR USING DIFFERENT NETWORK ARCHITECTURES.

Network Architecture	MPJPE	PMPJPE
1. (S-MHSA $\oplus$ SGA) $\rightarrow$ ((T-MHSA $\rightarrow$ PL) $\oplus$ TGA)	33.5	25.5
2. (S-MHSA $\rightarrow$ (T-MHSA$\rightarrow$ PL)) $\oplus$ (SGA $\rightarrow$ TGA)	32.0	25.1
3. (T-MHSA $\oplus$ TGA) $\rightarrow$ ((S-MHSA $\rightarrow$ PL) $\oplus$ SGA)	31.9	24.5
4. ((T-MHSA $\rightarrow$ PL) $\rightarrow$ S-MHSA) $\oplus$ (TGA $\rightarrow$ SGA)	31.5	24.5
5. ((S-MHSA $\rightarrow$ PL) $\rightarrow$ T-MHSA) $\oplus$ (SGA $\rightarrow$ TGA)	31.3	24.5
6. ((T-MHSA $\rightarrow$ PL) $\oplus$ TGA) $\rightarrow$ (S-MHSA $\oplus$ SGA)	31.2	24.6
7. (T-MHSA $\rightarrow$ (S-MHSA $\rightarrow$ PL)) $\oplus$ (TGA $\rightarrow$ SGA)	30.9	23.8
8. ((S-MHSA $\rightarrow$ PL) $\oplus$ SGA) $\rightarrow$ (T-MHSA $\oplus$ TGA)	**29.4**	**23.3**

F. Attention

Fig. 5 illustrates the spatial and temporal attention maps. The spatial MHSA map shows pairwise attention among the 17 joints in a frame. The attention is broadly distributed, with substantial off-diagonal responses indicating dependencies between nonadjacent joints. Structured clusters around upper- and lower-body joints further suggest interactions among body parts. These patterns indicate that the Transformer branch captures global spatial dependencies for holistic pose reasoning. The temporal MHSA map shows attention across frames. Strong responses near the diagonal indicate that neighboring frames contribute substantially to temporal modeling and motion continuity. Vertical bands and other off-diagonal clusters indicate interactions with temporally distant frames. These

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

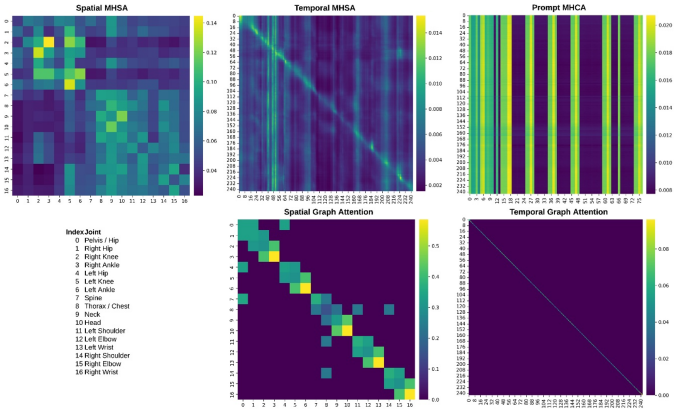


Fig. 5. Spatial and temporal attention maps.

patterns suggest that the Transformer captures local temporal continuity and longer-range dependencies, providing broader temporal context beyond simple local smoothing.

The prompt MHCA map represents cross-attention between pose features and prompts, with vertical axis as frames and horizontal axis as prompts. The prominent vertical bands indicate that a subset of prompt tokens receives consistently higher attention across the sequence. This suggests that the semantic conditioning provides persistent contextual information for pose modeling rather than varying independently over frames.

The spatial graph attention map represents joint-to-joint propagation over the 17 joints. The dominant near-diagonal clusters correspond primarily to anatomically connected joints, such as the shoulder–elbow–wrist chains. Because the graph weights are normalized by node degree, the contribution of each connected neighbor is adjusted according to the local graph structure. The limited off-diagonal responses indicate that the motion-refinement branch primarily performs topology-constrained message passing, which promots stable kinematic propagation and structural consistency. The temporal graph attention map represents motion propagation across frames. Its strong diagonal dominance indicates that the motion-refinement branch primarily relies on temporally local interactions. The relatively weak off-diagonal responses suggest that the graph branch emphasizes local temporal continuity rather than long-range retrieval of repeated or similar motion patterns. This localized propagation complements the broader temporal dependencies captured by the Transformer.

TABLE XII
ERRORS USING DIFFERENT TEMPORAL HYPERPARAMETER SETTINGS.

Metrics	α_t	β_t =3	5	7	9
MPJPE	0.125	30.3	**29.4**	29.5	30.9
PMPJPE		24.1	**23.3**	**23.3**	23.9
MPJPE	0.25	**29.4**	29.7	29.8	30.8
PMPJPE		**23.3**	23.5	23.4	23.6
MPJPE	0.5	30.8	**30.4**	31.1	31.7
PMPJPE		23.9	**23.5**	24.1	23.8

Table XII presents the impact of temporal hyperparameters in TGA. When α_t is 0.125 and β_t is 5, GAtFuN achieved the best performance, with an MPJPE of 29.4mm and a PMPJPE of 23.3mm. Small α_t, e.g., 0.125 or 0.25, often yields better performance regardless of the choice for β_t. This is attributed to the gradual decay in temporal attention, which allows frames farther apart to contribute effectively to the pose estimation. A large α_t imposes faster decay, diminishing the model's ability to capture long-term dependencies, which are critical for sequences with greater dynamic and complex motions. Moderate β_t values (5–7) yield better results by providing a sufficiently wide temporal range for attention without excessive temporal information. A large β_t, e.g., 9, slightly degrades performance, likely due to the model attending to an overly broad temporal window, introducing irrelevant contents.

TABLE XIII
ERRORS ON CHANNEL DIMENSION (D) AND SEQUENCE LENGTH (T).

T	D=128		D=256		D=512	
	MPJPE	P-MPJPE	MPJPE	P-MPJPE	MPJPE	P-MPJPE
27	34.8	27.3	33.2	26.1	31.8	24.9
81	32.5	25.5	31.0	24.3	30.2	23.9
243	31.4	24.7	30.4	23.7	**29.4**	**23.3**

G. Channel Dimensions and Receptive Fields

Table XIII reports the performance using different channel dimensions and temporal length. As the number of dimensions increases, errors decrease by as much as 4.62% for MPJPE and 4.71% for P-MPJPE. This suggests that higher-dimensional embeddings provide stronger representational capacity for modeling complex joint relationships and motions. With a fixed dimension, as the temporal length is increased, there is a consistent error drop. For example, when the channel dimension is 512, the MPJPE is 31.8 mm and 29.4 mm for T at 27 and 243, respectively.This implies that longer temporal contexts (i.e., larger T) enable a better capture of motion continuity and long-range dependencies, thereby enhancing temporal stability and reducing motion jitter.

H. Loss Function

Table XIV reports the results of GAtFuN using MPJPE loss and our proposed loss. Our proposed loss, including WMPJPE, temporal consistency (TC), and MPJVE, yielded better performance in the non-diffusion setting. In the diffusion setting, however, GAtFuN with MPJPE loss produced comparable performance. With the CPN detector, both settings achieved the same MPJPE (29.4 mm), while GAtFuN with MPJPE loss obtained a marginally lower P-MPJPE (23.2 mm vs. 23.3 mm). Table XIV also reports the results using ground-truth (GT) as inputs. The trajectory of the ground-truth is mostly smooth and noise-free, which makes motion refinement by GAtFuN limited. The results in the non-diffusion and diffusion cases show a similar pattern with overall smaller errors. These results suggest that the proposed loss yields an improved performance, particularly in non-diffusion settings.

TABLE XIV
PERFORMANCE USING DIFFERENT LOSS ON HUMAN3.6M.

Dector		Loss Function	MPJPE	P-MPJPE
Non-diff.	CPN	MPJPE	40.5	31.8
		Proposed loss	**39.8**	**31.6**
	SH	MPJPE	38.9	31.1
		Proposed loss	**36.2**	**30.0**
	GT	MPJPE	15.9	12.7
		Proposed loss	**14.5**	**12.3**
Diff.	CPN	MPJPE	**29.4**	23.2
		Proposed loss	**29.4**	**23.3**
	GT	MPJPE	**14.9**	**11.1**
		Proposed loss	15.7	12.0

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

I. Computational Cost

Table XV reports the computational costs (Para.: number of parameters; MAC: Multiply-Accumulate Operations; and FPS: Frame Per Second at Inference). In the non-diffusion settings, GAtFuN achieves the best MPJPE accuracy with the largest number of parameters among all methods. A similar pattern is observed in the diffusion settings. Although GAtFuN has a large parameter count due in part to prompt learning and the dual-stream design, parameter size does not directly translate to impractical runtime cost. The computation efficiency of GAtFuN is highly competitive. Specifically, the MACs of GAtFuN are about 1G, and the FPS is 2,726 and 24 in the non-diffusion and diffusion settings. The performance gain of GAtFuN arises from effective motion-aware spatial-temporal modeling rather than expensive post-hoc refinement. In contrast, FinePOSE [9] employs prompt learning and has a similar number of parameters yet achieves the worse MPJPE. This suggests that integration of prompt learning alone is insufficient to drive accuracy improvements. Additionally, while D3DP [12] improves performance through auxiliary refinement components such as Hypothesis Selection Network [52] (HSN) and SmoothNet [10], these additions significantly increase computation from 0.59G MACs to 2.01G and reduce inference speed from 29 FPS to 20 FPS, while still falling short of GAtFuN's MPJPE. The results demonstrate that GAtFuN provides a favorable balance between accuracy and efficiency compared to competing prompt-based or refinement-heavy approaches.

TABLE XV
COMPUTATIONAL COST. H: HYPOTHESES. I: DENOISING ITERATION. †: POSE REFINEMENT.

	Method	H	I	Para. (M)	MACs (G)	FPS	MPJPE
Non-diff.	PoseFormer [4]	1	0	9.5	0.81	1952	44.3
	MHFormer [53]	3	0	24.7	4.81	598	43.1
	P-STMO [33]	1	0	**6.7**	0.87	3040	42.8
	MixSTE [6]	1	0	33.6	**0.57**	**4547**	41.0
	GAtFuN	1	0	224.3	1.08	2726	**39.8**
Diffusion	D3DP [12]	20	10	**34.8**	**0.59**	**29**	39.5
	D3DP + HSN [52]	20	10	35.4	1.98	24	39.1
	D3DP + HSN [52] †	20	10	35.8	2.01	20	38.9
	FinePOSE [9]	20	10	200.6	0.62	26	40.2
	GAtFuN	20	10	225.8	1.09	24	**38.3**

V. Conclusion

This paper presents a motion-aware, dual-stream network that integrates velocity and spatial-temporal fusion for tracking human motions in 3D. Compared to the state-of-the-art methods, GAtFuN reduced errors by 7.8% and 6.8% in MPJPE and PMPJPE on Human3.M, respectively. When ground truth keypoints were used, GAtFuN exhibited an MPJPE of 15.7mm, which outperformed FinePOSE by 6%. In non-diffusion settings, GAtFuN exhibits a superior MPJPE of 14.5mm. On both the MPI-INF-3DHP and 3DPW datasets, it demonstrated superior temporal smoothness, outperforming the compared methods, particularly under occlusion and motion blur. Our ablation study suggests that the proper structure of the SGA and TGA affects the performance. In addition, normalized adjacency matrices with node-degree weighting and logistic decay resulted in the most accurate model.

References

[1] Y. Zhou, J. Han, X. Yuan, Z. Wei, and R. Hong, "Inverse sparse group lasso model for robust object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1798–1810, 2017.

[2] X. Yuan, L. Kong, D. Feng, and Z. Wei, "Automatic feature point detection and tracking of human actions in time-of-flight videos," *IEEE/CAA J. of Automatica Sinica*, vol. 4, no. 4, pp. 677–685, 2017.

[3] R. Liu, J. Shen, H. Wang, C. Chen, S.-c. Cheung, and V. Asari, "Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction," in *CVPR*, 2020, pp. 5064–5073.

[4] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in *ICCV*, 2021, pp. 11 656–11 665.

[5] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *CVPR*, 2019, pp. 7753–7762.

[6] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in *CVPR*, 2022, pp. 13 232–13 242.

[7] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *ICCV*, 2023, pp. 15 085–15 099.

[8] X. Liu and H. Tang, "Strformer: Spatial–temporal–retemporal transformer for 3D human pose estimation," *Image and Vision Computing*, vol. 140, p. 104863, 2023.

[9] J. Xu, Y. Guo, and Y. Peng, "Finepose: Fine-grained prompt-driven 3D human pose estimation via diff. models," in *CVPR*, 2024, pp. 561–570.

[10] A. Zeng, L. Yang, X. Ju, J. Li, J. Wang, and Q. Xu, "Smoothnet: A plug-and-play network for refining human poses in videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 625–642.

[11] G. Delmas, P. Weinzaepfel, F. Moreno-Noguer, and G. Rogez, "Posefix: correcting 3d human poses with natural language," in *International Conference on Computer Vision*, 2023, pp. 15 018–15 028.

[12] 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 *ICCV*, 2023, pp. 14 761–14 771.

[13] Z. Zou and W. Tang, "Modulated graph convolutional network for 3D human pose estimation," in *ICCV*, 2021, pp. 11 477–11 487.

[14] N. Azizi, H. Possegger, E. Rodolà, and H. Bischof, "3D human pose estimation using möbius graph convolutional networks," in *ECCV*. Springer, 2022, pp. 160–178.

[15] Z. Liu, H. Wang, X. He, and L. Tao, "Robust 3D human pose estimation model based on temporal convolution," in *ICSIP*, 2020, pp. 456–462.

[16] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3D human pose estimation with bone-based pose decomp." *IEEE Trans. on Cir. and Sys. for Vid. Tech.*, vol. 32, no. 1, pp. 198–209, 2021.

[17] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation," in *CVPR*, 2023, pp. 8877–8886.

[18] X. Yan, J. Xie, M. Liu, H. Li, and H. Gao, "Hierarchical local temporal network for 2D-to-3D human pose estimation," *IEEE Internet of Things Journal*, vol. 12, no. 1, pp. 869–880, 2025.

[19] B. X. Yu, Z. Zhang, Y. Liu, S.-h. Zhong, Y. Liu, and C. W. Chen, "Glagcn: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video," in *ICCV*, 2023, pp. 8818–8829.

[20] S. Mehraban, V. Adeli, and B. Taati, "Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network," in *WACV*, 2024, pp. 6920–6930.

[21] J. Peng, Y. Zhou, and P. Mok, "Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation," in *CVPR*, 2024, pp. 1123–1132.

[22] W. Li, M. Liu, H. Liu, T. Guo, T. Wang, H. Tang, and N. Sebe, "GraphMLP: A graph mlp-like architecture for 3D human pose estimation," *Pattern Recognition*, vol. 158, p. 110925, 2025.

[23] Y. Wang, Z. Wang, M. Li, and H. Yan, "3D human pose estimation with two-step mixed-training strategy," in *WACV*, 2024, pp. 3332–3341.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, and G. Goh, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[25] L. Kong, X. Yuan, and A. M. Maharjan, "A hybrid framework for automatic joint detection of human poses in depth frames," *Pattern Recognition*, vol. 77, pp. 216–225, 2018.

[26] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3D human pose estimation," in *ECCV*, 2018, pp. 68–84.

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2026.3749495

[27] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[28] 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 *3DV*, 2017, pp. 506–516.

[29] T. Von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using imus and a moving camera," in *ECCV*, 2018, pp. 601–617.

[30] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang, "Exploiting temporal contexts with strided transformer for 3D human pose estimation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1282–1293, 2022.

[31] R. Wang, X. Ying, and B. Xing, "Exploiting temporal correlations for 3D human pose estimation," *IEEE Transactions on Multimedia*, vol. 26, pp. 4527–4539, 2023.

[32] J. Peng, Y. Zhou, and P. Mok, "Ehfusion: an efficient heterogeneous fusion model for group-based 3D human pose estimation," *The Visual Computer*, pp. 1–23, 2024.

[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 *ECCV*, 2022, pp. 461–478.

[34] Y. Wang, H. Kang, D. Wu, W. Yang, and L. Zhang, "Global and local spatio-temporal encoder for 3D human pose estimation," *IEEE Transactions on Multimedia*, vol. 26, pp. 4039–4049, 2023.

[35] Z. Tang, Z. Qiu, Y. Hao, R. Hong, and T. Yao, "3D human pose estimation with spatio-temporal criss-cross attention," in *CVPR*, 2023, pp. 4790–4799.

[36] Y. Wang, M. Li, and H. Yan, "Utilizing motion segmentation for optimizing the temporal adjacency matrix in 3D human pose estimation," *Neurocomputing*, vol. 600, p. 128153, 2024.

[37] S. Du, Z. Yuan, and T. Ikenaga, "Kinematics-aware spatial-temporal feature transform for 3D human pose estimation," *Pattern Recognition*, vol. 150, p. 110316, 2024.

[38] S. Zheng and W. Cao, "Ehgformer: An efficient hypergraph-injected transformer for 3D human pose estimation," *Image and Vision Computing*, p. 105425, 2025.

[39] Y. Wang, X. Liu, X. Wang, D. Wu, W. Yang, and H. Kang, "DRPose: A diffusion-based pose refinement framework for 3d human pose estimation," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 36, no. 5, pp. 6572–6585, 2026.

[40] Z. Sun, Y. Liang, Z. Ma, T. Zhang, L. Bao, G. Li, and S. He, "Repose: 3D human pose estimation via spatio-temporal depth relational consistency," in *ECCV*, 2024, pp. 309–325.

[41] J. Liu, M. Liu, H. Liu, and W. Li, "Tcpformer: Learning temporal correlation with implicit pose proxy for 3D human pose estimation," *AAAI*, pp. 5478–5486, 2025.

[42] 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," *AAAI*, pp. 3842–3850, 2025.

[43] L. Liao and R. Zheng, "Langpose: Language-aligned motion for robust 3d human pose estimation," in *Winter Conference on Applications of Computer Vision*, 2026, pp. 8073–8083.

[44] H. Cui, W. Hua, R. Huang, S. Jia, and T. Hayama, "Sasmamba: A lightweight structure-aware stride state space model for 3d human pose estimation," in *IEEE Winter Conference on Applications of Computer Vision*, 2026, pp. 2721–2730.

[45] J. Zhou, T. Zhang, Z. Hayder, L. Petersson, and M. Harandi, "Diff3dhpe: A diffusion model for 3D human pose estimation," in *ICCV*, 2023, pp. 2092–2102.

[46] J. Gong, L. G. Foo, Z. Fan, Q. Ke, H. Rahmani, and J. Liu, "Diffpose: Toward more reliable 3d pose estimation," in *CVPR*, 2023, pp. 13 041–13 051.

[47] X. Xu and J. Li, "MMCPose: Multimodal condition-driven 3d human pose estimation via diffusion models," *IEEE Transactions on Multimedia*, vol. 28, pp. 4373–4382, 2026.

[48] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *CVPR*, 2018, pp. 7103–7112.

[49] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[50] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 483–499.

[51] M. Li, H. Hu, J. Xiong, X. Zhao, and H. Yan, "Tswinpose: Enhanced monocular 3D human pose estimation with jointflow," *Expert Systems with Applications*, vol. 249, p. 123545, 2024.

[52] 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. on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 10 678–10 691, 2024.

[53] W. Li, H. Liu, H. Tang, P. Wang, and L. Van Gool, "Mhformer: Multi-hypothesis transformer for 3D human pose estimation," in *CVPR*, 2022, pp. 13 147–13 156.

Authorized licensed use limited to: University of North Texas. Downloaded on November 04,2026 at 02:27:16 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.