PART-AWARE STRUCTURED REGISTER MODELING FOR 3D HUMAN POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current 2D-to-3D lifting approaches have achieved notable success in 3D human pose estimation. However, relying solely on 2D joint sequences leaves these methods susceptible to 2D detection noise and overlooks rich visual cues in RGB images that can help resolve depth ambiguities. Recent visually assisted pose lifting methods attempt to leverage such visual information, but largely focus on localized regions around predicted joints, overlooking high-level semantic visual representations and hierarchical structural context of the human body. To address these limitations, we propose Part-Aware Structured Register Modeling for 3D Human Pose Estimation (PSRM), a new framework comprising Global-to-Part Expert Adapter (GPEA) and Kinematic Register Attention (KRA). To bridge the granularity mismatch between frame-level visual context and joint-level pose features, GPEA progressively adapts and integrates visual cues through a cascade of global and part-specific experts. In addition, KRA introduces persistent global and part-level registers that iteratively interact with joint representations under kinematic constraints, enabling the aggregation and layer-spanning preservation of structural context. Extensive experiments on Human3.6M and MPI-INF-3DHP demonstrate that PSRM achieves state-of-the-art performance. Moreover, GPEA and KRA feature a highly versatile plug-and-play design, seamlessly integrating into representative pose lifting backbones and consistently boosting their precision.

1 INTRODUCTION

3D human pose estimation from monocular video aims to recover the 3D coordinates of human joints from video sequences. It is a fundamental task for applications such as action analysis (Liu & Yuan, 2018; Liu et al., 2017), human–computer interaction (Svenstrup et al., 2009), and virtual reality (Hagbi et al., 2010). Early methods (Li & Chan, 2014; Pavlakos et al., 2017; Mehta et al., 2017b; Sun et al., 2017) typically regress 3D human poses directly from monocular images, combining human perception and 3D structure recovery within a single model. Without an explicit intermediate representation, it is difficult to determine whether prediction errors arise from joint localization or 3D geometric reasoning. A two-stage formulation makes this separation explicit. Martinez et al. (Martinez et al., 2017) decompose the task into 2D joint detection and 2D-to-3D pose lifting, showing that 2D joint coordinates provide a compact and effective intermediate representation.

Despite substantial progress in 2D-to-3D pose lifting, the limited information available from monocular observations leaves depth ambiguity and human self-occlusion as major challenges in 3D pose recovery (Li et al., 2022a;b; Huang et al., 2025; Mehraban et al., 2024; Zhu et al., 2023). The task therefore remains underdetermined. Moreover, 2D pose detection reduces an image to a sparse set of joint coordinates, discarding human appearance, local texture, occlusion relationships, and other visual cues. When 2D joints are incorrectly detected or missing, the subsequent pose lifting model cannot refer back to the original image for information that can help correct these errors.

To recover visual information missing from 2D pose representations, Context-Aware PoseFormer (Zhao et al., 2023a) uses deformable attention to extract visual features around joints from multi-scale feature maps. APP (Zhang et al., 2024) subsequently extends this approach to video pose lifting by sampling local visual features from intermediate feature maps at locations guided by 2D joints and fusing them with temporal pose features. However, by focusing visual extraction on localized regions around predicted 2D joints, these methods often miss the broader context. Such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

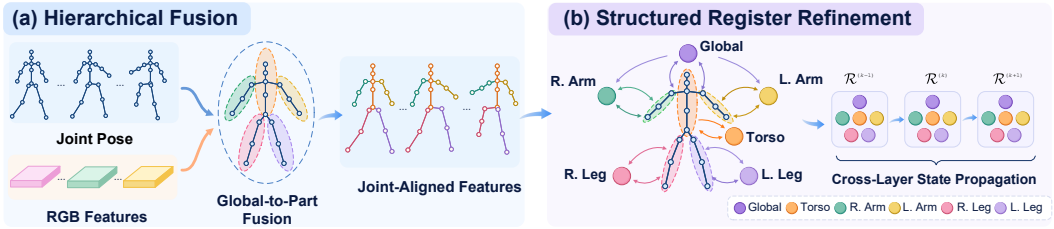

Figure 1: Illustration of hierarchical fusion and structured register refinement in our proposed PSRM. **(a)** Hierarchical fusion integrates RGB and pose features from global to part levels, producing joint-aligned representations. **(b)** Structured register refinement updates joint representations through interactions with global and part registers, whose states propagate across layers.

local sampling strategies can be vulnerable to severe occlusions and 2D detection errors, while failing to capture the overall body configuration and scene constraints. To address these limitations, we argue that high-level semantic representations from a visual encoder provide holistic whole-body appearance and scene context, which can complement sparse 2D joint coordinates in recovering 3D poses. Nevertheless, these representations describe the frame as a whole, whereas pose estimation requires features specific to individual joints. Exploiting this visual context therefore requires adapting the shared frame-level information to joint-level pose representations.

Most video-based pose estimation methods represent the human body as a set of individual joints and model spatial and temporal dependencies through interactions among joint representations. Body-part modeling provides an additional level of structural organization by grouping kinematically related joints and capturing relationships specific to different body parts. Previous studies (Xu et al., 2024; Yang et al., 2026) demonstrate its benefits through part-specific representations and anatomically grouped feature modeling. These findings motivate us to maintain part-level context alongside whole-body context for joint refinement. However, existing approaches typically perform structural modeling within individual layers, lacking a persistent state to explicitly aggregate and retain hierarchical context across network layers. To overcome this limitation, we draw on register-based models, where dedicated tokens act as persistent states that progressively aggregate and preserve contextual information through layer-wise interactions (Wang et al., 2025; Kirby et al., 2026). We therefore introduce global and part-level registers that explicitly maintain whole-body and part-specific structural context across layers. Each register aggregates information from the full body or a designated kinematic group, while joint representations incorporate the resulting hierarchical context through attention.

To this end, we propose Part-Aware Structured Register Modeling (PSRM), a backbone-compatible framework for 3D human pose estimation comprising two dedicated modules: the Global-to-Part Expert Adapter (GPEA) and Kinematic Register Attention (KRA), as illustrated in Fig. 1. Specifically, GPEA addresses the granularity mismatch between frame-level visual features and joint-level pose representations through a global-to-part adaptation process. A Global Expert first adapts visual information under whole-body pose context, after which Part Experts further refine the adapted representations according to distinct body parts, establishing a hierarchical global-to-part organization in the joint representations. KRA then carries this hierarchical organization into subsequent pose lifting by introducing persistent global and part-level registers that maintain and progressively refine structural context across layers. To enable such structured refinement, we introduce Kinematic Joint Aggregation (KJA) and Inter-Register Isolation (IRI), where KJA constrains each register to aggregate information from its designated joint group and IRI blocks direct interactions between registers to preserve their distinct aggregation roles. Together, GPEA and KRA equip existing pose lifting architectures with joint-aligned visual information and hierarchical structural context. Both modules are designed in a plug-and-play manner and can be integrated into various representative pose lifting backbones. We instantiate PSRM with MotionAGFormer, MixSTE, and MotionBERT to demonstrate its backbone compatibility.

Our main contributions are as follows:

- We propose PSRM, a backbone-compatible framework for 3D human pose estimation that combines frame-level visual–pose adaptation with persistent hierarchical structural refinement.

- To jointly address cross-modal granularity mismatch and hierarchical structural modeling, we introduce two dedicated modules: GPEA progressively adapts frame-level visual context to joint representations via global and part experts, while KRA refines the resulting representations by organizing register–joint interactions under kinematic constraints and propagating whole-body and part-level context across layers.
- Extensive experiments on Human3.6M and MPI-INF-3DHP demonstrate state-of-the-art performance and validate the effectiveness of GPEA and KRA. Their plug-and-play design enables PSRM to be integrated with different pose lifting backbones, including MotionAGFormer, MixSTE, and MotionBERT.

2 Related Work

Pose-based 3D Human Pose Estimation. Pose-based methods (Martinez et al., 2017; Pavllo et al., 2019; Zhang et al., 2022; Mehraban et al., 2024) recover 3D human poses from 2D joint coordinates. For video inputs, they exploit spatial relationships between joints and temporal dependencies across frames to reduce depth ambiguity. VideoPose3D (Pavllo et al., 2019) uses dilated temporal convolutions to predict the 3D pose of a target frame from a sequence of 2D poses. PoseFormer (Zheng et al., 2021) introduces spatial and temporal Transformers to model joint relationships within frames and motion dependencies across frames. MixSTE (Zhang et al., 2022) adopts a sequence-to-sequence formulation, alternating between spatial modeling within frames and temporal modeling of individual joint trajectories. STCFormer (Tang et al., 2023) models spatial and temporal context in parallel to enhance spatio-temporal interactions. Despite their different architectures, these methods primarily derive structural and motion information from 2D joint sequences. When occlusion or detection errors lead to inaccurate joint localization, pose-only inputs provide limited evidence for correcting the resulting ambiguity. In contrast, PSRM complements 2D pose representations with visual features from the corresponding frames, allowing image content to contribute to 3D pose recovery.

Visual-based 3D Human Pose Estimation. Visual-based methods (Pavlakos et al., 2017; Zhao et al., 2023a; Zhang et al., 2024) exploit image evidence for 3D pose estimation. Early approaches directly recover 3D poses from RGB images. Pavlakos et al. (Pavlakos et al., 2017) estimate joint locations with volumetric probability distributions, VNect (Mehta et al., 2017b) jointly predicts 2D and 3D joint information, and Sun et al. (Sun et al., 2018) connect heatmap representations with coordinate regression through differentiable integration. More recent methods retain explicit 2D poses and introduce visual features to assist pose lifting. Context-Aware PoseFormer (Zhao et al., 2023a) extracts joint context from intermediate feature maps of a pretrained 2D pose detector, while APP (Zhang et al., 2024) performs pose-aware sampling from multi-level detector features and integrates them with spatio-temporal pose representations. Unlike methods that derive visual cues from 2D pose detectors, our approach uses frame-level representations from a visual encoder, with GPEA adapting the shared visual context to individual joints through global and part experts.

Structural Priors for 3D Human Pose Estimation. Structural priors exploit the topology and kinematic relationships of the human skeleton to guide 3D pose estimation (Zhao et al., 2019; 2022; Cai et al., 2023; Peng et al., 2024). SemGCN (Zhao et al., 2019) represents poses as skeleton graphs and learns local and global joint relationships through semantic graph convolutions. GraFormer (Zhao et al., 2022) combines graph convolutions with Transformers to capture non-local dependencies while preserving skeletal adjacency. Beyond modeling connections between joints, HTNet (Cai et al., 2023) models human structure hierarchically at the levels of individual joints, body parts, and the whole body. KTPFormer (Peng et al., 2024) further incorporates kinematic topology and joint trajectory topology into spatial and temporal modeling. These methods mainly use structural priors to guide joint feature learning and information propagation. In contrast, KRA introduces global and part registers alongside joint representations to explicitly maintain hierarchical structural context across network layers, allowing information about the whole body and its parts to be progressively aggregated and reused during joint refinement.

3 Method

As illustrated in Fig. 2, PSRM follows a sequence-to-sequence pipeline that takes a 2D pose sequence and the corresponding visual features as input and predicts a 3D pose sequence of the same length.

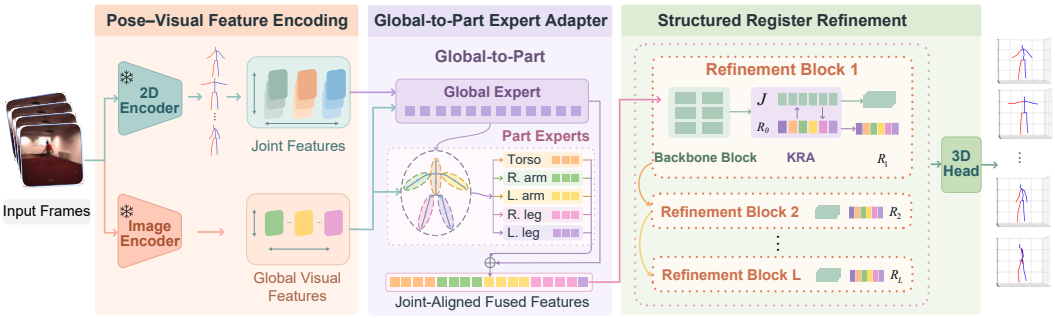

Figure 2: Overview of our proposed PSRM. Specifically, Pose–Visual Feature Encoding first extracts pose and visual representations from the input sequence. Subsequently, Global-to-Part Expert Adapter performs global-to-part visual–pose fusion. Finally, Structured Register Refinement integrates KRA into selected backbone layers for hierarchical structural modeling.

Given a clip of T frames, a 2D pose estimator produces joint coordinates $\mathbf{X} \in \mathbb{R}^{B\times T\times J\times 2}$, while a visual encoder extracts frame-level features $\mathbf{F} \in \mathbb{R}^{B\times T\times D_F}$. The two inputs are first projected into pose representations $\mathbf{P} \in \mathbb{R}^{B\times T\times J\times C}$ and visual representations $\mathbf{V} \in \mathbb{R}^{B\times T\times C}$. GPEA then integrates the visual representation into the pose features through global and part experts, yielding joint-aligned features $\mathbf{H}^0 \in \mathbb{R}^{B\times T\times J\times C}$. These features are subsequently processed by the pose lifting backbone for spatial–temporal modeling, with KRA inserted at selected layers to update the joint representations together with the global and part registers. The updated register states are propagated to the next KRA interaction, while the refined joint features continue through the backbone. Finally, a regression head maps the output joint representations to $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times T\times J\times 3}$, yielding a 3D pose for each input frame. PSRM can be instantiated with different pose lifting backbones while retaining this overall pipeline.

3.1 Global-to-Part Expert Adapter

Visual and pose representations operate at different granularities: the former encodes frame-level appearance, whereas the latter describes individual joints. To adapt the shared visual information to joint representations while accounting for body structure, we propose the Global-to-Part Expert Adapter (GPEA), as illustrated in Fig. 3. GPEA follows a progressive global-to-part adaptation process. A Global Expert shared across joints first adapts the visual representation under whole-body pose context. Based on the globally adapted features, Part Experts with separate parameters further refine the representations according to part-specific pose configurations.

We next formulate this process for a single frame, omitting the batch and temporal indices for simplicity. Let $\mathbf{P}_j \in \mathbb{R}^C$ denote the feature of joint j, and $\mathbf{V} \in \mathbb{R}^C$ the corresponding frame-level visual feature. Separate LayerNorm operations are first applied to obtain $\bar{\mathbf{P}}_j$ and $\bar{\mathbf{V}}$. We further introduce a learnable joint identity embedding $\mathbf{E}_j \in \mathbb{R}^C$ to preserve joint-specific information during adaptation.

Global Expert. We first apply average pooling to all normalized joint features to obtain the global pose features $\mathbf{G} \in \mathbb{R}^C$:

$$\mathbf{G} = \frac{1}{J}\sum_{j=0}^{J-1} \bar{\mathbf{P}}_j. \tag{1}$$

For joint j, we concatenate its pose features, the global pose features, the visual features, and its joint identity embedding:

$$\mathbf{U}_{G,j} = \mathrm{Concat}\left(\bar{\mathbf{P}}_j, \mathbf{G}, \bar{\mathbf{V}}, \mathbf{E}_j\right), \tag{2}$$

where $\mathbf{U}_{G,j} \in \mathbb{R}^{4C}$. A shared two-layer MLP then computes the residual update:

$$\Delta\mathbf{H}_{G,j} = \mathrm{MLP}_G\left(\mathbf{U}_{G,j}\right), \qquad \mathbf{H}_{G,j} = \mathbf{P}_j + \alpha_G \Delta\mathbf{H}_{G,j}. \tag{3}$$

where $\mathrm{MLP}_G : \mathbb{R}^{4C} \to \mathbb{R}^C$ shares parameters across all joints, and α_G is a learnable residual scaling coefficient. Although the mapping is shared, its output depends on the features $\mathbf{P}_j$ and identity embedding $\mathbf{E}_j$ of each joint.

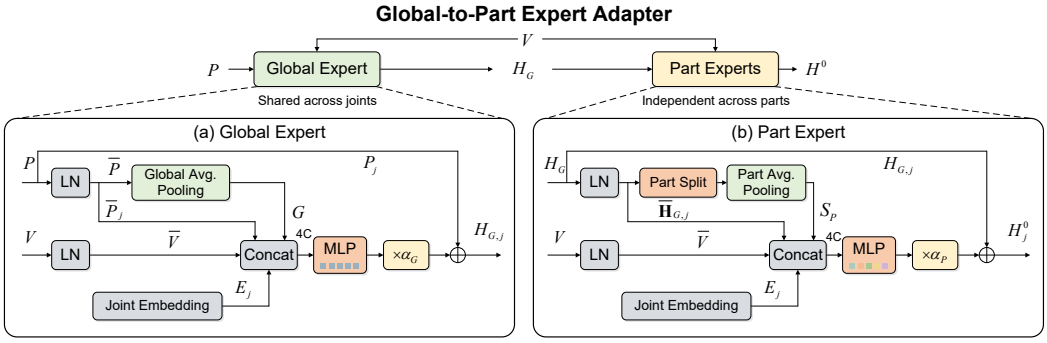

Figure 3: Architecture of GPEA. **(a)** The Global Expert first adapts shared visual information to joint representations under whole-body pose context. **(b)** Part Experts further refine the globally adapted representations using part-specific pose context. The Global Expert is shared across all joints, while the Part Experts are parameterized separately for different body parts.

Part Experts. We divide the J joints into five disjoint body parts: torso/head, left arm, right arm, left leg, and right leg. The set of joints in part p is denoted by $\mathcal{J}_p$, with

$$\mathcal{J}_p \cap \mathcal{J}_q = \varnothing, \quad p \neq q, \qquad \bigcup_{p=1}^{5} \mathcal{J}_p = \{0, \ldots, J-1\}. \tag{4}$$

We first normalize the output of the Global Expert, obtaining $\bar{\mathbf{H}}_G = \mathrm{LN}(\mathbf{H}_G)$. For part p, we average the normalized joint features after global fusion to obtain the fused part features $\mathbf{S}_p \in \mathbb{R}^C$:

$$\mathbf{S}_p = \frac{1}{|\mathcal{J}_p|} \sum_{j \in \mathcal{J}_p} \bar{\mathbf{H}}_{G,j}. \tag{5}$$

For each joint $j \in \mathcal{J}_p$, the corresponding Part Expert concatenates the current joint features, the part features, the visual features for each frame, and the joint identity embedding:

$$\mathbf{U}_{p,j} = \mathrm{Concat}\left(\bar{\mathbf{H}}_{G,j}, \mathbf{S}_p, \bar{\mathbf{V}}, \mathbf{E}_j\right), \tag{6}$$

where $\mathbf{U}_{p,j} \in \mathbb{R}^{4C}$. The Part Expert for part p then computes

$$\Delta\mathbf{H}_{p,j} = \mathrm{MLP}_p\left(\mathbf{U}_{p,j}\right), \qquad \mathbf{H}_j^0 = \mathbf{H}_{G,j} + \alpha_p \Delta\mathbf{H}_{p,j}, \quad j \in \mathcal{J}_p. \tag{7}$$

where $\mathrm{MLP}_p : \mathbb{R}^{4C} \to \mathbb{R}^C$, and α_p is a learnable residual scaling coefficient for part p. The five experts share an architecture, with parameters shared within each part and separate across parts.

Finally, we combine the outputs of all Part Experts in the original joint order to obtain $\mathbf{H}^0 \in \mathbb{R}^{B\times T\times J\times C}$ and feed it into the spatio-temporal pose lifting backbone.

3.2 Kinematic Register Attention

Register tokens (Darcet et al., 2024) provide additional computational states that aggregate information through attention and carry it across layers. This property makes them suitable for maintaining whole-body and part-level context for joint refinement. However, unconstrained registers lack explicit structural roles, with no mechanism ensuring that different registers capture whole-body context or information specific to distinct body parts. We therefore introduce Kinematic Register Attention (KRA) to organize register–joint interactions according to human kinematic structure. Specifically, Kinematic Joint Aggregation (KJA) specifies which joints each register can aggregate information from, while Inter-Register Isolation (IRI) blocks direct attention between registers to preserve their distinct aggregation roles.

KRA is inserted at selected backbone layers, whose indices are denoted by $\mathcal{L}_{\mathrm{KRA}} = \{\ell_1, \ell_2, \ldots, \ell_K\}$. At the k-th KRA interaction, the module takes the joint representations from the current backbone layer, $\mathbf{J}^{\ell_k} \in \mathbb{R}^{B\times T\times J\times C}$, and the register states from the previous interaction, $\mathbf{R}^{k-1} \in \mathbb{R}^{B\times T\times 6\times C}$. The six learnable initial register embeddings are defined as

$$\mathbf{R}_{\mathrm{init}} = [\mathbf{r}_G, \mathbf{r}_{TH}, \mathbf{r}_{RA}, \mathbf{r}_{LA}, \mathbf{r}_{RL}, \mathbf{r}_{LL}] \in \mathbb{R}^{6\times C}, \tag{8}$$

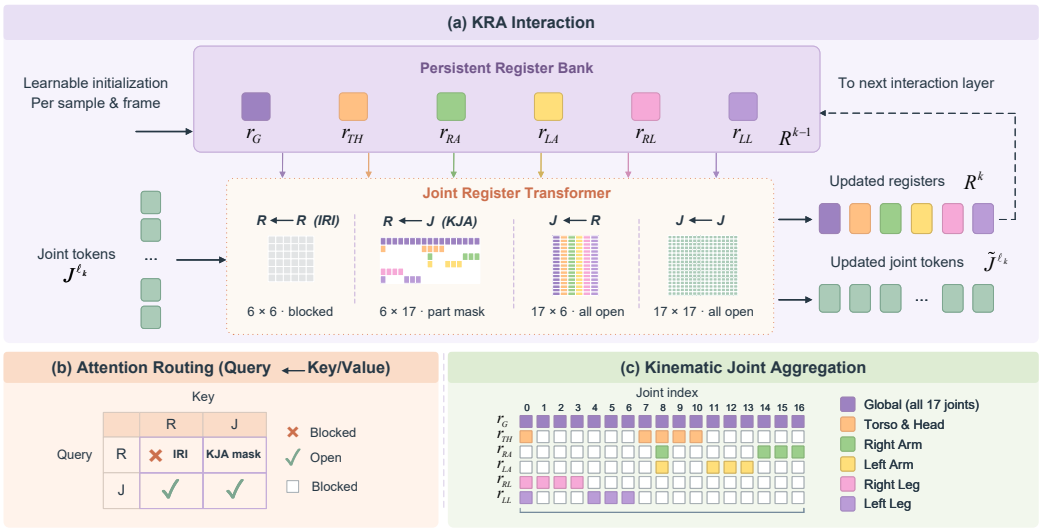

Figure 4: Architecture of KRA. **(a)** Registers interact with joint representations, and the updated register states are passed to the next interaction layer. **(b)** IRI blocks $R \leftarrow R$ attention, while KJA restricts $R \leftarrow J$ attention. Joint queries can attend to all registers and joints. **(c)** Joint groups used by KJA. Colored cells indicate the joints that each register can attend to. R and J denote registers and joints, respectively. Arrows follow the Query ← Key/Value convention.

where the register indices $r \in \{G, TH, RA, LA, RL, LL\}$ correspond to the whole body, torso/head, right arm, left arm, right leg, and left leg, respectively. Each initial register embedding is denoted by $\mathbf{r}_r \in \mathbb{R}^C$. For the first interaction, $\mathbf{R}_{\text{init}}$ is replicated along the batch and time dimensions to obtain the initial register states $\mathbf{R}^0 \in \mathbb{R}^{B\times T\times 6\times C}$. Interactions between registers and joints are performed independently within each frame.

3.2.1 Kinematic Joint Aggregation

KJA restricts each register to aggregate information from its designated kinematic group, encouraging part-specific structural representations and reducing redundant aggregation across registers that could arise from unrestricted access to all joints. Let $\mathcal{S}_r$ denote the set of joints visible to register r. Unlike the disjoint body parts used in GPEA, the groups in KJA can share connecting joints. As shown in Fig. 4, both arm groups include the connecting joint in the torso (index 8), and both leg groups include the pelvis joint (index 0), preserving the structural connections between the limbs and torso.

These groups are encoded as an attention bias from register queries to joint keys, $\mathbf{M}_{RJ} \in \{0, -\infty\}^{6\times J}$:

$$[\mathbf{M}_{RJ}]_{r,j} = \begin{cases} 0, & j \in \mathcal{S}_r, \\ -\infty, & j \notin \mathcal{S}_r. \end{cases} \tag{9}$$

Thus, the Global Register can aggregate information from all joints, while each Part Register can attend only to joint representations within its corresponding kinematic chain.

3.2.2 Inter-Register Isolation

KJA restricts which joints each register can attend to but leaves attention between registers unrestricted. To preserve their distinct aggregation roles, we introduce IRI to block register-to-register attention, including self-attention. Each register retains its previous state through the residual connection and aggregates information only from the joints specified by KJA, allowing indirect communication through joint representations. For simplicity, we omit the batch and time indices below. At the k-th interaction, the current register states $\mathbf{R}^{k-1} \in \mathbb{R}^{6\times C}$ and joint representations $\mathbf{J}^{\ell_k} \in \mathbb{R}^{J\times C}$ are concatenated along the token dimension as $\mathbf{Z}^k = \text{Concat}(\mathbf{R}^{k-1}, \mathbf{J}^{\ell_k}) \in \mathbb{R}^{(6+J)\times C}$.

IRI defines the attention bias $\mathbf{M}_{RR} \in \{0, -\infty\}^{6\times 6}$ for attention between registers, where $r, s \in \{G, TH, RA, LA, RL, LL\}$: $[\mathbf{M}_{RR}]_{r,s} = -\infty$. Together with $\mathbf{M}_{RJ}$ defined by KJA, the full

attention bias is

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{RR} & \mathbf{M}_{RJ} \\ \mathbf{0}_{J\times 6} & \mathbf{0}_{J\times J} \end{bmatrix} \in \{0, -\infty\}^{(6+J)\times(6+J)}. \tag{10}$$

The zero blocks in $\mathbf{M}$ allow joint queries to attend to all registers and joints. Using this attention bias, a pre-norm Transformer block updates the concatenated tokens:

$$\mathbf{Z}'^k = \mathbf{Z}^k + \mathrm{MHA}\big(\mathrm{LN}(\mathbf{Z}^k), \mathbf{M}\big), \qquad \mathbf{Z}''^k = \mathbf{Z}'^k + \mathrm{MLP}\big(\mathrm{LN}(\mathbf{Z}'^k)\big). \tag{11}$$

The output is split according to the original token order as $(\mathbf{R}^k, \widetilde{\mathbf{J}}^{\ell_k}) = \mathrm{Split}(\mathbf{Z}''^k)$. The joint representations $\widetilde{\mathbf{J}}^{\ell_k}$ are passed to the next backbone block, or to the regression head if this is the final block. At the k-th KRA interaction, joint representations and register states are updated simultaneously, with joint queries attending to the register states $\mathbf{R}^{k-1}$. The updated register states $\mathbf{R}^k$ remain unchanged between KRA interactions and serve as input to the next interaction.

4 Experiments

4.1 Datasets and Evaluation Metrics

Human3.6M. Human3.6M (Ionescu et al., 2013) contains approximately 3.6 million frames of multi-view video from 11 subjects performing 15 actions. Following the standard setting (Pavllo et al., 2019; Li et al., 2022b; Zhang et al., 2022; Peng et al., 2024), we use subjects S1, S5, S6, S7, and S8 for training and S9 and S11 for testing, with a 17-joint skeleton. We set the input sequence length to $T = 243$ and use 2D poses detected by SH (Newell et al., 2016) in the main experiments. For ablation studies, we compare different 2D pose detectors and use ground-truth 2D joints to assess the upper-bound performance. We report MPJPE under Protocol 1 and P-MPJPE under Protocol 2, both in millimeters (mm), with lower values indicating better performance.

MPI-INF-3DHP. MPI-INF-3DHP (Mehta et al., 2017a) contains over 1.3 million frames captured in indoor, green-screen, and outdoor settings. Following common practice (Li et al., 2022b; Zhang et al., 2022; Peng et al., 2024), we use the official 2D joint annotations, set the input sequence length to $T = 81$, and evaluate on valid test frames. We report 3D Percentage of Correct Keypoints (PCK), Area Under the PCK Curve (AUC), and MPJPE, with higher PCK/AUC and lower MPJPE indicating better performance.

4.2 Implementation Details

We apply PSRM as a plug-and-play extension to three representative pose lifting backbones: MotionAGFormer (Mehraban et al., 2024), MixSTE (Zhang et al., 2022), and MotionBERT (Zhu et al., 2023). Unless otherwise specified, all ablation studies use MotionAGFormer-B. This backbone is also used for evaluation with ground-truth 2D joints on Human3.6M and for experiments on MPI-INF-3DHP. YOLOv8 (Jocher et al., 2023) is used to detect and crop human regions, while visual features are extracted in advance using a frozen ViT (Dosovitskiy et al., 2020) initialized with HMR 2.0 weights (Goel et al., 2023). All experiments are conducted on a single NVIDIA A100 GPU (80GB) using AdamW with an initial learning rate of 5×10^{-4}, weight decay of 0.01, and a learning rate decay factor of 0.99 after each epoch. Training lasts 80/90 epochs with batch sizes of 12/16 on Human3.6M/MPI-INF-3DHP, respectively. Horizontal flipping is applied during training and testing (Zhu et al., 2023; Mehraban et al., 2024; Huang et al., 2025). We use published MotionAGFormer results (Mehraban et al., 2024) for the main comparisons and retrain the ablation baseline under the same settings as our method.

4.3 Comparison with State-of-the-Art Methods

Results on Human3.6M. We evaluate PSRM on Human3.6M using detected 2D poses and an input sequence of $T = 243$ frames. As shown in Table 1, PSRM achieves state-of-the-art (SOTA) performance under both Protocol 1 (MPJPE) and Protocol 2 (P-MPJPE). Specifically, when integrated with MixSTE (Zhang et al., 2022), PSRM w. MixSTE reduces MPJPE from 40.9 mm to 33.7 mm (a 7.2 mm or 17.6% error reduction) and P-MPJPE to 27.6 mm. It outperforms SAMA-L (Lu et al., 2025) (36.5 mm) by 2.8 mm despite using fewer input frames ($T = 243$ vs. $T = 351$). Beyond

Table 1: Quantitative comparisons on Human3.6M using detected 2D poses. T: Number of input frames. P1: MPJPE (mm). P2: P-MPJPE (mm). Per-action results are reported under P1. The best and second-best P1/P2 scores are in bold and underlined, respectively.

Method	T	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	P1↓	P2↓
2D Pose-based Methods																		
VideoPose3D	243	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	36.5
PoseFormer	81	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	34.6
MHFormer	351	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	34.4
P-STMO	243	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	34.4
MixSTE	243	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	32.6
PoseFormerV2	243	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	45.2	35.6
MotionBERT	243	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	32.9
KTPFormer	243	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	31.9
MotionAGFormer-B	243	36.4	38.4	36.8	32.9	40.9	48.5	36.6	34.6	51.7	52.8	41.0	36.4	36.5	26.7	27.0	38.4	32.6
PoseMamba-X	243	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.1	31.5
PoseMagic	243	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.5	–
SAMA-L	351	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	36.5	31.0
Wu et al.	243	37.6	39.3	35.3	37.8	41.4	48.7	38.9	37.7	49.7	52.5	40.7	39.2	38.9	26.7	27.0	39.4	–
Visually Assisted Pose Lifting Methods																		
CA-PF-HRNet	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	41.6	33.9
APP w. MotionAGFormer-B	243	33.7	37.2	35.0	32.2	37.2	40.7	34.6	34.3	45.1	51.4	37.0	34.6	36.8	27.0	26.8	36.2	29.5
PSRM w. MotionBERT(ours)	243	34.1	36.0	34.0	31.3	35.0	40.1	33.0	33.6	42.5	47.5	36.1	32.0	35.8	27.3	27.7	35.1	28.4
PSRM w. MotionAGFormer-B(ours)	243	34.7	35.0	33.4	31.2	34.8	38.7	32.5	32.9	40.6	44.1	35.3	32.5	34.2	26.2	27.2	<u>34.2</u>	<u>27.8</u>
PSRM w. MixSTE(ours)	243	33.9	34.0	32.3	30.8	34.1	37.8	32.1	31.8	40.3	44.0	35.1	32.5	33.3	26.6	26.9	**33.7**	**27.6**

Table 2: Comparison on MPI-INF-3DHP. PCK/AUC in %; MPJPE in mm.

Method	T	AUC↑	PCK↑	MPJPE↓
D3DP	243	78.2	97.7	29.7
PoseFormerV2	81	78.8	97.9	27.8
HSTFormer	81	71.5	97.3	41.4
CA-PF-HRNet-32	1	75.4	98.0	32.7
MotionAGFormer-B	81	84.2	98.3	18.2
KTPFormer	27	84.4	98.9	19.2
KTPFormer	81	**85.9**	98.9	16.7
Wu et al.	243	78.3	98.1	29.3
PSRM w. MotionAGFormer-B(ours)	81	**85.9**	**99.4**	**15.6**

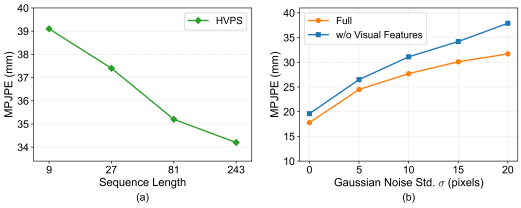

Figure 5: **(a)** Effect of input sequence length on Human3.6M. **(b)** Robustness to 2D pose noise with and without RGB features. σ is the standard deviation of zero-mean Gaussian noise in pixels; GT denotes ground-truth 2D poses.

MixSTE, PSRM also yields consistent improvements with MotionBERT (Zhu et al., 2023) and MotionAGFormer-B (Mehraban et al., 2024), reducing MPJPE from 39.2 mm to 35.1 mm and from 38.4 mm to 34.2 mm, respectively, with gains observed across almost all 15 action categories.

We also compare PSRM against existing visually assisted pose lifting methods under the identical temporal setting ($T = 243$). Equipped with the MotionAGFormer-B backbone, PSRM w. MotionAGFormer-B achieves 34.2 mm in MPJPE and 27.8 mm in P-MPJPE, outperforming APP (Zhang et al., 2024) by 2.0 mm (5.5%) and 1.7 mm (5.8%) under Protocol 1 and Protocol 2, respectively. This empirical consistency underscores the effectiveness of PSRM as a plug-and-play extension that combines visual context with hierarchical fusion and structured register refinement.

Results on MPI-INF-3DHP. To assess generalizability beyond Human3.6M, we evaluate PSRM on MPI-INF-3DHP, which encompasses indoor and outdoor environments. Utilizing ground-truth 2D poses and RGB features under $T = 81$, PSRM w. MotionAGFormer-B achieves 99.4% PCK, 85.9% AUC, and 15.6 mm MPJPE (Table 2). Compared to the standalone MotionAGFormer-B backbone (18.2 mm MPJPE, 98.3% PCK, 84.2% AUC), PSRM brings a substantial 14.3% relative MPJPE reduction, alongside consistent gains of 1.1% in PCK and 1.7% in AUC. Crucially, it outperforms competitive approaches such as KTPFormer (Peng et al., 2024) (16.7 mm MPJPE, 98.9% PCK) under the same sequence length. These consistent improvements across all metrics indicate that PSRM generalizes reliably to a second benchmark with more complex and diverse capture environments.

4.4 Ablation Study

Input Sequence Length. We examine the effect of input sequence length on 3D pose estimation. As shown in Fig. 5(a), increasing T from 9 to 27, 81, and 243 progressively reduces MPJPE from 39.1 mm to 37.4 mm, 35.2 mm, and 34.2 mm, respectively. The consistent improvement with longer sequences shows that PSRM benefits from increased temporal context and is powerful in modeling long-range dependencies. However, the gain becomes smaller beyond $T = 81$, suggesting

diminishing returns from additional temporal context. We therefore use $T = 243$ as the default input sequence length on Human3.6M unless otherwise specified.

Effect of Input Modalities. As shown in Table 3, we evaluate the individual and combined effects of 2D pose coordinates and visual context. Using 2D poses alone yields an MPJPE of 42.5 mm, while visual features alone achieve 40.0 mm. Combining modalities through the proposed fusion strategy reduces the error to 34.6 mm. These results demonstrate the complementary value of pose and visual information for 3D pose estimation.

Table 3: Ablation study of input modalities on Human3.6M.

2D Pose	Visual Features	MPJPE↓
✓	×	42.5
×	✓	40.0
✓	✓	**34.6**

Effect of Model Components. As shown in Table 4, adding either GPEA or KRA to the baseline reduces MPJPE from 35.7 mm to 34.6 mm, showing that both global-to-part visual adaptation and structured register modeling are individually effective. Combining the two modules further improves MPJPE to 34.2 mm, yielding a 1.5 mm (4.2%) reduction over the baseline and a further 0.4 mm gain over either module alone. This additional improvement indicates that adapting visual information to joint representations and maintaining hierarchical structural context provide complementary benefits.

Effect of KRA Configurations. Table 5 analyzes the key designs of KRA. Introducing register tokens alone reduces MPJPE from 35.7 mm to 35.4 mm and P-MPJPE from 30.0 mm to 28.6 mm, showing the benefit of explicit structural representations. Adding KJA or IRI further lowers MPJPE to 35.1 mm and 34.9 mm, respectively. Combining both achieves the best performance of 34.6 mm MPJPE and 28.0 mm P-MPJPE, demonstrating that constrained joint aggregation and inter-register isolation further strengthen structured register modeling.

Table 4: Ablation study of GPEA and KRA on Human3.6M.

Models	GPEA	KRA	MPJPE↓
Baseline	×	×	35.7
Baseline w. GPEA	✓	×	34.6
Baseline w. KRA	×	✓	34.6
Baseline w. PSRM	✓	✓	**34.2**

Table 5: Ablation study of different KRA configurations on Human3.6M.

Variant	MPJPE↓	P-MPJPE↓
Baseline	35.7	30.0
Registers Only	35.4	28.6
w/o IRI	35.1	28.3
w/o KJA	34.9	28.2
Full KRA	**34.6**	**28.0**

Robustness to 2D Pose Noise. To evaluate robustness to 2D localization errors, we independently perturb the normalized x- and y-coordinates of ground-truth 2D joints at test time with zero-mean Gaussian noise of standard deviation $2\sigma/W$, where $W = 1000$ is the original image width and $\sigma \in \{0, 5, 10, 15, 20\}$ is measured in original-image pixels. As shown in Fig. 5(b), compared with the variant without visual features (*w/o Visual Features*), the full model (*Full*) consistently achieves lower MPJPE as noise increases, with the gap widening at higher noise levels. These results show that RGB visual features provide complementary information that mitigates 2D localization errors and improves robustness to noisy 2D inputs.

5 Conclusion

We present PSRM, a flexible framework for 3D human pose estimation that combines global-to-part visual–pose adaptation with structured register refinement. Specifically, GPEA progressively adapts frame-level visual context to joint representations through global and part experts, while KRA maintains whole-body and part-level structural context across layers for joint refinement. Extensive experiments on Human3.6M and MPI-INF-3DHP demonstrate the effectiveness of PSRM, with ablation studies further validating the complementary contributions of GPEA and KRA. Moreover, the plug-and-play design of the two modules enables PSRM to be readily applied to different pose lifting backbones, consistently improving their performance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature searches and identifying relevant prior work. We did not use generative AI tools to develop the core research ideas or methods, design or conduct experiments, or generate experimental results. We reviewed the AI-assisted literature findings and checked the relevant claims and references against the original publications. We take responsibility for the final content of this work, including any content informed by AI-assisted literature research.

REFERENCES

Jialun Cai, Hong Liu, Runwei Ding, Wenhao Li, Jianbing Wu, and Miaoju Ban. Htnet: Human topology aware network for 3D human pose estimation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Hanyuan Chen, Jun-Yan He, Wangmeng Xiang, Zhi-Qi Cheng, Wei Liu, Hanbing Liu, Bin Luo, Yifeng Geng, and Xuansong Xie. HDFormer: High-order directed transformer for 3D human pose estimation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 581–589, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=2dnO3LLiJ1`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14737–14748. IEEE, 2023.

Nate Hagbi, Oriel Bergig, Jihad El-Sana, and Mark Billinghurst. Shape recognition and pose estimation for mobile augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1369–1379, 2010.

Yunlong Huang, Junshuo Liu, Ke Xian, and Robert Caiming Qiu. Posemamba: Monocular 3D human pose estimation with bidirectional global-local spatio-temporal state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3842–3850, 2025.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. `https://github.com/ultralytics/ultralytics`, 2023.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, et al. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32058–32069, 2026.

Sijin Li and Antoni B Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pp. 332–347. Springer, 2014.

Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3D human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2022a.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13147–13156, 2022b.

Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1159–1168. IEEE, 2018.

Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

Ye Lu, Jie Wang, Jianjun Gao, Rui Gong, Chen Cai, and Kim-Hui Yap. A structure-aware and motion-adaptive framework for 3D human pose estimation with mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7958–7968, 2025.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2659–2668. IEEE, 2017.

Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3D human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6905–6915. IEEE, 2024.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision*, pp. 506–516. IEEE, 2017a.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3D human pose estimation with a single rgb camera. *ACM Transactions on Graphics*, 36(4):1–14, 2017b.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1263–1272. IEEE, 2017.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7753–7762, 2019.

Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1123–1132, 2024.

Xiaoye Qian, Youbao Tang, Ning Zhang, Mei Han, Jing Xiao, Ming-Chun Huang, and Ruei-Sung Lin. Hstformer: Hierarchical spatial-temporal transformers for 3D human pose estimation. *arXiv preprint arXiv:2301.07322*, 2023.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3D human pose estimation. In *European Conference on Computer Vision*, pp. 461–478. Springer, 2022.

Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3D human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14761–14771, 2023.

Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2602–2611, 2017.

Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision*, pp. 536–553. Springer, 2018.

Mikael Svenstrup, Soren Tranberg, Hans Jorgen Andersen, and Thomas Bak. Pose estimation and adaptive robot behaviour for human-robot interaction. In *2009 IEEE International Conference on Robotics and Automation*, pp. 3571–3576. IEEE, 2009.

Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3D human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4790–4799, 2023.

Feng Wang, Jiahao Wang, Sucheng Ren, Guoyizhe Wei, Jieru Mei, Wei Shao, Yuyin Zhou, Alan Yuille, and Cihang Xie. Mamba-reg: Vision mamba also needs registers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14944–14953. IEEE, 2025.

Yi Wu, Jingtian Li, Shangfei Wang, Guoming Li, Meng Mao, and Linxiang Tan. Learning knowledge from textual descriptions for 3D human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10835–10843, 2026.

Jinglin Xu, Yijie Guo, and Yuxin Peng. Finepose: Fine-grained prompt-driven 3d human pose estimation via diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 561–570. IEEE, 2024.

Zepeng Yang, Junxuan Bai, Hao Li, Ju Dai, Junjun Pan, Yongfeng Yin, and Bin Li. Vimcan: Visual-inertial 3d human pose estimation with hybrid mamba-cross-attention network. *arXiv preprint arXiv:2605.07552*, 2026.

Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. Gla-gcn: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8818–8829, 2023.

Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13232–13242, 2022.

Jinyan Zhang, Mengyuan Liu, Hong Liu, Guoquan Wang, and Wenhao Li. App: Adaptive pose pooling for 3D human pose estimation from videos. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1672–1681, 2024.

Xinyi Zhang, Qiqi Bao, Qinpeng Cui, Wenming Yang, and Qingmin Liao. Pose magic: Efficient and temporally consistent human pose estimation with a hybrid mamba-gcn network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10248–10256, 2025.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3D human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3420–3430. IEEE, 2019.

Qitao Zhao, Ce Zheng, Mengyuan Liu, and Chen Chen. A single 2d pose with context is worth hundreds for 3D human pose estimation. *Advances in Neural Information Processing Systems*, 36: 27394–27413, 2023a.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8877–8886, 2023b.

Weixi Zhao, Weiqiang Wang, and Yunjie Tian. Graformer: Graph-oriented transformer for 3D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20438–20447, 2022.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11656–11665, 2021.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15085–15099, 2023.

A SUPPLEMENTARY MATERIAL

Effect of 2D Pose Estimators and Visual Encoders. Table 6 evaluates PSRM with different 2D pose estimators and visual encoders across three representative lifting backbones. PSRM remains effective with SH, CPN, and ViTPose inputs, demonstrating compatibility with different 2D pose sources. With ViT features, SH inputs achieve the lowest MPJPE across all three backbones, reaching 34.2, 33.7, and 35.1 mm on MotionAGFormer-B, MixSTE, and MotionBERT, respectively. For both SH and CPN inputs, replacing ViT with HRNet-w32 consistently increases MPJPE and P-MPJPE across all three backbones. These results show that PSRM remains effective across different 2D pose sources, visual encoders, and lifting backbones, demonstrating its robustness to variations in both input representations and backbone architectures.

Table 6: Results with different 2D pose estimators and visual encoders on Human3.6M. Errors are in mm. * denotes results reported in the corresponding papers. Best results for each backbone and metric are highlighted in bold.

2D Pose	Visual Encoder	MotionAGFormer-B		MixSTE		MotionBERT	
		MPJPE↓	P-MPJPE↓	MPJPE↓	P-MPJPE↓	MPJPE↓	P-MPJPE↓
Reported Pose-Only Baselines							
SH*	–	38.4	32.6	–	–	39.2	–
CPN*	–	–	–	40.9	32.6	–	–
PSRM (Pose + Vision)							
SH	ViT	**34.2**	27.8	**33.7**	27.6	**35.1**	28.4
SH	HRNet-w32	40.0	33.8	38.5	32.8	39.1	32.7
CPN	ViT	35.8	28.0	34.5	**27.2**	37.7	29.1
CPN	HRNet-w32	39.4	31.8	37.8	30.4	38.8	31.5
ViTPose	ViT	35.3	**27.6**	34.7	27.3	36.4	**27.9**

A.1 RESULTS WITH GROUND-TRUTH 2D POSES ON HUMAN3.6M

To eliminate the effect of 2D detection errors, we further evaluate PSRM using ground-truth 2D poses on Human3.6M. As shown in Table 7, PSRM w. MotionAGFormer-B achieves an MPJPE of 17.8 mm, outperforming the standalone MotionAGFormer-B baseline (19.4 mm) and the visually assisted pose lifting method APP (Zhang et al., 2024) (19.1 mm) by 1.6 mm and 1.3 mm, respectively. These results show that the gains of PSRM persist even with accurate 2D inputs, indicating that its improvements extend beyond robustness to 2D detection noise.

Table 7: Comparison on Human3.6M using ground-truth 2D poses. MPJPE in mm.

Method	T	MPJPE↓
MHFormer	351	30.5
StridedFormer	351	28.5
MixSTE	243	21.6
P-STMO	243	29.3
GLA-GCN	243	21.0
STCFormer-L	243	21.3
HDFormer	96	21.6
KTPFormer	243	18.1
MotionAGFormer-B	243	19.4
APP w. MotionAGFormer-B	243	19.1
PSRM w. MotionAGFormer-B	243	**17.8**

A.2 ADDITIONAL ABLATION STUDIES

Effect of Global-to-Part Adaptation. Table 8 isolates the individual and progressive contributions of the Global Expert and Part Experts in GPEA. Incorporating the Global Expert alone reduces the MPJPE from 35.7 mm to 35.0 mm, whereas using the Part Experts alone reaches 35.4 mm. This contrast indicates that global adaptation guided by whole-body pose context yields a larger individual performance gain than isolated part-level modeling.

Cascading the Part Experts after global adaptation further reduces the error to 34.6 mm, outperforming the global-only and part-only variants by 0.4 mm and 0.8 mm, respectively. These results confirm that fine-grained part refinement complements global adaptation, supporting the progressive global-to-part design of GPEA.

Effect of KRA Placement. We investigate how the number and placement of KRA interaction layers affect estimation accuracy on Human3.6M. As shown in Table 9, increasing the number of interactions from $[2, 5]$ to $[1, 3, 5]$ does not improve MPJPE, indicating that more interactions alone do not necessarily yield better performance. With the same number of interactions, $[1, 4, 7]$ outperforms $[1, 3, 5]$ on both metrics, suggesting that extending register–joint interactions to deeper layers benefits pose refinement. The configuration $[1, 3, 5, 7]$ achieves the lowest MPJPE and P-MPJPE among the tested settings. These results highlight the importance of both the number and placement of KRA interactions.

Table 8: Ablation study of Global and Part Experts in GPEA on Human3.6M.

Method	Global	Part	MPJPE↓
Baseline	×	×	35.7
Part Expert	×	✓	35.4
Global	✓	×	35.0
Global + Part Expert	✓	✓	**34.6**

Table 9: KRA interaction layers on Human3.6M. Layer indices are zero-based. Errors are in mm.

Interaction Layers	MPJPE↓	P-MPJPE↓
$[2, 5]$	35.7	30.0
$[1, 3, 5]$	35.8	29.0
$[1, 4, 7]$	35.2	28.3
$[1, 3, 5, 7]$	**34.6**	**28.0**

Effect of Fusion Strategies. Table 10 compares different fusion strategies under identical 2D pose and visual inputs. Concatenation stacks pose and visual features along the channel dimension and projects them to the backbone feature dimension, while Addition combines the projected features element-wise. Multilayer Perceptron applies a nonlinear mapping to the concatenated features, whereas Cross-Attention uses pose features as queries and visual features as keys and values. These alternatives achieve MPJPEs ranging from 35.0 to 35.7 mm, whereas the proposed GPEA achieves the lowest error of 34.6 mm, outperforming the next-best Cross-Attention by 0.4 mm. These results indicate that the proposed global-to-part adaptation enables more effective cross-modal integration than the evaluated alternatives.

Table 10: Ablation study of fusion strategies on Human3.6M.

Fusion Strategy	MPJPE↓
Concatenation	35.6
Addition	35.7
Multilayer Perceptron	35.3
Cross-Attention	35.0
GPEA (Ours)	**34.6**

A.3 Qualitative Results

We visualize the estimated 3D poses on Human3.6M and compare PSRM using the MotionAGFormer backbone with PoseMamba (Huang et al., 2025) in Fig. 6. The selected examples include challenging poses with bent arms and legs. As highlighted by the green circles and red arrows, our predictions are closer to the ground truth in these regions. In particular, our method better estimates the leg positions in the upper-right and lower-left examples, while the arm poses in the other two examples also show better agreement with the ground truth. These results demonstrate that our method can estimate more accurate 3D poses for the challenging examples shown.

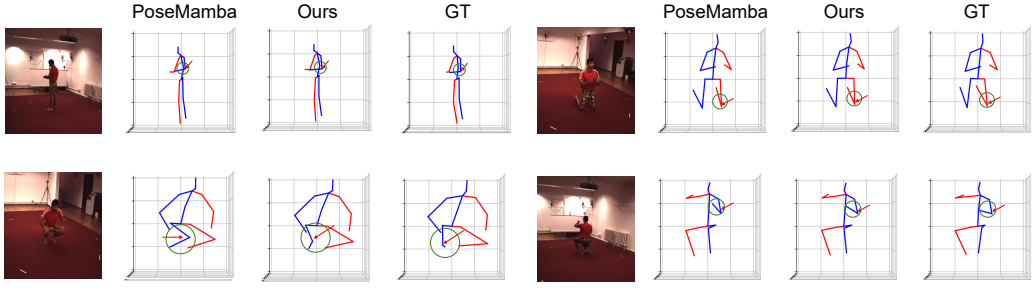

Figure 6: Qualitative comparison between PoseMamba and PSRM on Human3.6M. Green circles and red arrows highlight differences between the predictions and ground truth (GT).