This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LangPose: Language-Aligned Motion for Robust 3D Human Pose Estimation

Longyun Liao
McMaster University
liaol13@mcmaster.ca

Rong Zheng
McMaster University
rzheng@mcmaster.ca

Abstract

2D-to-3D human pose lifting is an ill-posed problem due to depth ambiguity and occlusion. Existing methods relying on spatial and temporal consistency alone are insufficient to resolve these problems especially in the presence of significant occlusions or high dynamic actions. Semantic information, however, offers a complementary signal that can help disambiguate such cases. To this end, we propose LangPose, a framework that leverages action knowledge by aligning motion embeddings with text embeddings of fine-grained action labels. LangPose operates in two stages: pretraining and fine-tuning. In the pretraining stage, the model simultaneously learns to recognize actions and reconstruct 3D poses from masked and noisy 2D poses. During the fine-tuning stage, the model is further refined using real-world 3D human pose estimation datasets without action labels. Additionally, our framework incorporates masked body parts and masked time windows in motion modeling, encouraging the model to leverage semantic information when spatial and temporal consistency is unreliable. Experiments demonstrate the effectiveness of LangPose, achieving SOTA level performance in 3D pose estimation on public datasets, including Human3.6M and MPI-INF-3DHP. Specifically, LangPose achieves an MPJPE of 36.7mm on Human3.6M with detected 2D poses as input and 15.5mm on MPI-INF-3DHP with ground-truth 2D poses as input.

1. Introduction

3D human pose estimation (HPE) is crucial for better understanding human motion and serves as a cornerstone for numerous downstream tasks such as human action recognition, human-computer interaction, and virtual reality. 2D-to-3D pose lifting, which estimates 3D human poses from 2D keypoints extracted from a single image or video frame, is a mainstream approach to 3D human pose estimation due to the maturity and computational efficiency of modern 2D pose detectors. However, it is an ill-posed problem as depth ambiguity and occlusion often result in multiple possible 3D poses corresponding to a single 2D pose.

To cope with the inherent ambiguity, prior work has relied on spatial and temporal cues to enforce consistency in 3D pose predictions. Various architectures, including TCN-based [4, 35], GCN-based [2, 6, 49], and transformer-based models [36, 55, 62], have been designed to capture this information. Pretraining strategies such as random frame masking, joint masking, and noise addition on large-scale dataset have been utilized to enhance the learning of spatial and temporal cues [62].

However, spatial and temporal consistency diminishes in the presence of significant occlusions or high dynamic actions such as those found in the real world (*e.g.*, in sports). Semantic information refers to high-level contextual knowledge about human actions, intentions, or interactions. Such information, often available in fine-grained textual descriptions of human activities, can inform or constrain pose estimation and help disambiguate pose configurations, particularly in cases of occlusion and depth ambiguity. Semantic information can serve as a complementary cue. For example, when only the upper body is visible, semantic priors associated with the action "sitting down" indicate that the hips are flexed and the knees are bent to roughly 90^o, allowing the model to plausibly reconstruct the occluded lower-limb joints. Similarly, in highly dynamic scenes, where temporal cues become less reliable, semantic priors linked to the ongoing action, for example, the expected limb trajectories and contact events in "dribbling a ball", can guide the estimation of a full pose sequence even if it is partially occluded over time.

One possible approach to incorporating action cues into 3D human pose estimation is through multi-task learning (MTL), where pose estimation and action recognition are jointly optimized. However, MTL typically represents action categories with one-hot vectors, thereby discarding the rich semantic relationships conveyed in fine-grained textual descriptions. Furthermore, such models do not generalize well to pose sequences from unseen action classes. Therefore, we argue for *explicitly* aligning the latency feature space of pose sequences with action descriptions and propose LangPose, a framework that learns

a shared pose–language embedding via cross-modal contrastive training. Datasets designed for human motion generation, such as BABEL [40] and KIT [18], offer richer and more fine-grained action descriptions. However, they introduce new challenges: (1) severe class imbalance—word like 'walk' and 'transition' occur with high frequency, while many other descriptions are underrepresented in BABEL [40]; and (2) the non-uniqueness of textual action descriptions, as the same motion can correspond to multiple valid descriptions, especially in the presence of concurrent or compound actions, and vice versa.

To address the above challenges, LangPose leverages a two-stage training process comprising a pretraining phase followed by fine-tuning. In the pretraining stage, the model is trained on a synthetic dataset [40] containing motion sequences paired with fine-grained action labels, jointly optimizing for semantic alignment and 3D pose reconstruction. The semantic knowledge acquired during pretraining is generalizable across different 3D pose estimation datasets, in contrast to previous methods [26, 27, 59] that require retraining an action classifier for each target dataset. To tackle the issues of imbalanced data distribution in the pretraining dataset and the inherent uncertainty in human action recognition tasks, we replace the cross-entropy loss in infoNCE with a combination of KL-divergence and focal loss [22] to supervise multi-modal alignment. Additionally, inspired by MotionBERT [62], we implement varying-size temporal window masking and body part-level masking during pretraining. These strategies further encourage the model to reconstruct poses given semantic information rather than relying solely on spatial and temporal consistency. By the end of the pretraining stage, the model is equipped to recognize actions. In the fine-tuning stage, the model is further refined using real-world 3D human pose estimation datasets without action descriptions.

In summary, the contributions of this work are three folds:

- To the best of our knowledge, this is the first work to perform multi-modal pretraining that directly aligns the text embedding of descriptive action labels with skeleton motion sequences for a 3D human pose estimation task.
- We propose a new pre-training and fine-tuning strategy for motion representation learning. By integrating multi-modal representation learning of motion with descriptive action labels and 3D human pose estimation, the model is trained to embed generalizable action cues into 3D pose estimation during the pretraining stage.
- Experiments show that LangPose achieves competitive results on public benchmarks, surpassing most methods on MPI-INF-3DHP and attaining SOTA-level performance on Human3.6M, with notable gains on challenging actions and lower inference cost than diffusion-based models.

2. Related work

2.1. 3D Human pose estimation

3D human pose estimation can be categorized into two main approaches: the first involves estimating 3D human poses directly from images [19, 33, 45], and the second involves lifting 2D poses to 3D poses [25, 29, 36, 55]. This work falls into the second category. Within this category, various architectures have been proposed to capture spatial and temporal information in motion dynamics, including TCN-based [4, 35], GCN-based [2, 6, 49], and transformer-based [24, 36, 55, 62] models.

Since action labels provide comprehensive summaries of motion, we adopt spatial and temporal transformers [62] as our backbone for the pose encoder, which effectively learns and fuses motion representations across both temporal and spatial dimensions.

2.2. Skeleton-based human action recognition

Another task related to understanding human motion is human action recognition. Unlike 3D human pose estimation, which focuses on precise pose reconstruction, human action recognition aims to comprehend human motion at a higher level. Despite the different objectives, both tasks require an understanding of the spatial relationships between the movements of different body joints and the temporal dynamics within motion. Consequently, recent works have designed various architectures to capture the spatio-temporal information of motion dynamics for human action recognition including transformer-based [1, 10, 62], GCN-based [3, 5, 53, 60], LSTM-based [23, 61], and CNN-based [52, 56].

2.3. Multi-task learning

We use multi-task learning when the knowledge gained from one task can benefit other tasks [57]. This is particularly relevant for human pose estimation and human action recognition. The work in [26] was the first to highlight the interconnections between these tasks and attempted to estimate 2D/3D human poses alongside human action recognition. In their approach, both tasks share a common multi-task CNN, with human action recognition having an additional head to classify actions based on estimated poses and visual features. However, their method does not fully utilize the rich semantic information of actions, as the action classes are encoded as one-hot vectors. Similarly, [27] aims to estimate 2D and 3D human poses and corresponding actions in real-time, but also encodes action classes as one-hot vectors.

2.4. Multi-modal representation learning

Multi-modal representation learning methods, such as CLIP [41] and ALIGN [16], have proven highly effective

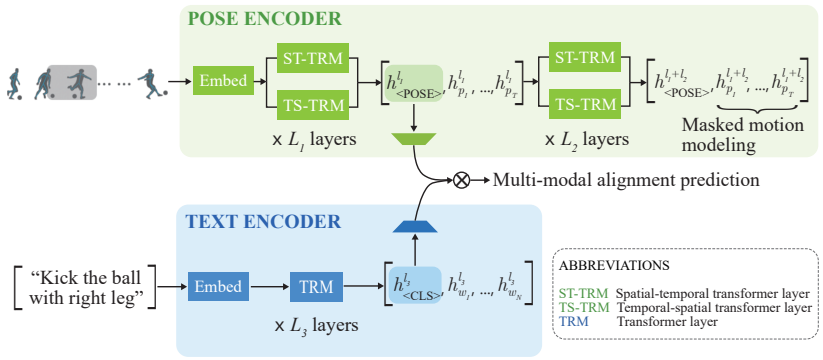


Figure 1. The overall architecture of LangPose.

for tasks like image captioning, text-image retrieval, and zero-shot image classification. These methods work by aligning text embeddings from a text encoder with feature embeddings from an image encoder. This alignment allows the image encoder to capture and utilize rich semantic information from images, significantly enhancing its performance on downstream tasks.

Pioneer works point out that similar techniques can be applied broadly across the motion–language spectrum, including text-to-motion generation [37, 38, 48], motion captioning [11], text–motion retrieval [32, 39], and text-aided discriminative tasks such as action recognition [51] and pose estimation [8, 12, 59]. Building on these task-specific advances, a recent trend is to pursue general-purpose motion–language models that unify generation, captioning, retrieval, and discriminative perception within a single pretrained backbone (e.g., shared encoders with lightweight task heads) [17, 20, 50]. In parallel, substantial effort has gone into curating large-scale motion–language corpora and whole-body resources to enable such unified modeling [14, 18, 21, 40, 58].

Among the studies above, the most relevant to our setting are two lines of work that reconstruct 3D motion from different inputs: multi-modal 3D human pose estimation from 2D to 3D, and text-conditioned motion generation from text to 3D. Our approach differs in two ways. First, prior multi-modal 3D HPE with textual cues usually centers on images rather than videos [8, 12], so the semantics of motion dynamics remain underexplored for the 3D HPE task. Second, our method complements text-conditioned generation without performing generation: rather than synthesizing motion from language, we leverage motion semantics captured by text–motion alignment as auxiliary supervision to regularize 2D-to-3D pose estimation, preserving camera-aware, metric accuracy while adding semantic guidance, especially under occlusion. In this way, we complete the picture—from learning semantics to using semantics for reconstruction.

The most closely related approach that aligns motion and text for human pose estimation is ActionPrompt [59]. Their method consists of two main components: an action-related text-prompt block and an action-specific pose-prompt block. The former enables the pose encoder to predict action labels from pose sequences using supervision from human pose estimation datasets [15, 44], while the latter refines the predicted poses based on the identified action. However, the semantic representations learned by the action-related text-prompt block are not transferable across datasets and must be retrained for each new domain. Moreover, since action labels in pose estimation datasets are typically coarse, this approach relies heavily on the action-specific pose prompts to recover fine-grained semantic details. This two-stage pipeline further assumes accurate classification of motion into discrete action labels, which is prone to failure in the presence of severe noise. In contrast, LangPose performs end-to-end alignment between motion and semantic information while reconstructing 3D pose sequences. The learned semantics are generalizable across datasets and eliminate the need for storing action-specific pose prompts during inference.

3. Methodology

3.1. Network architecture

The overall objective of LangPose is to enable 3D human pose estimation and text-motion alignment. The network consists of two encoders to extract features from text and pose data. Embedding pooling layers are then applied to generate global representations of the text and pose. These representations are used for text-motion alignment, and a regression head subsequently projects the pose features to estimate the 3D pose. The network architecture is illustrated in Fig. 1.

Text and pose encoders. The LangPose network consists of two parallel BERT-style models [9] operating over pose and text domains. The text stream utilizes three lay-

ers of transformer blocks, following the architecture of the original BERT [9]. The pose stream employs five layers of Spatial-Temporal transformer blocks, following MotionBERT [62].

Given a sequence of 2D pose skeletons $\mathbf{x} \in \mathbb{R}^{T\times J\times C_{in}}$, we first project them into a higher dimensional space through one MLP layer, resulting in features $\{p_1, p_2, ..., p_T\}$ where $p_i \in \mathbb{R}^{J\times C_f}$, C_{in} is the input dimension of 2D pose sequences, C_f is the feature dimension of pose embeddings. We then concatenate a learnable pose class token to these features, forming the sequence $\{p_0, p_1, p_2, ..., p_T\}$ where p_0 is the class token <POSE>.

For the text input, we have $\{w_0, w_1, w_2, ..., w_N\}$, where w_0 is <CLS> and w_N is <SEP>, namely tokens for classification and sentence separation. After processing the text inputs through the text encoder, our model generates encoded text features $\{h^{l_3}_{w_0}, h^{l_3}_{w_1}, h^{l_3}_{w_2}, ..., h^{l_3}_{w_N}\}$. Similarly, after processing the pose sequences through the pose encoder, our model produces encoded pose features $\{h^{l_1+l_2}_{p_0}, h^{l_1+l_2}_{p_1}, h^{l_1+l_2}_{p_2}, ..., h^{l_1+l_2}_{p_T}\}$.

Next, we introduce the structures to derive the global text representation and global pose representation, as follows.

Text embedding and pose embedding pooling layers for late fusion. The purpose of these layers is to project the learned features of the pose class token <POSE> and the text class token <CLS> to the same feature space. These layers consist of multiple MLPs (Multi-Layer Perceptrons).

Instead of average pooling all pose features, we use the pose feature in the pose class token, as it represents a weighted summation of all pose features through attention. Since pose features include an additional joint dimension, we perform a learnable weighted summation over the joint dimension of the pose class token <POSE>.

We extract the global pose representation from the middle layer of the pose encoder $h^{l_1}_{<\text{POSE}>}$. Shallow layers are chosen because they capture more action-specific information, which can then be propagated to deeper layers for 3D human pose estimation. After this stage, we denote the text global representation as h_W and the pose global representation as h_P for simplicity. These global representations are used in the task of multi-modal alignment prediction.

Pose estimation regression head. This structure is designed to project pose features $\{h^{l_1+l_2}_{p_1}, h^{l_1+l_2}_{p_2}, ..., h^{l_1+l_2}_{p_T}\}$ to predict 3D pose sequence $\mathbf{X} \in \mathbb{R}^{T\times J\times 3}$.

3.2. Pretraining tasks

We introduce two new pretraining strategies in addition to random joint and frame-level masked motion modeling: to promote semantic reasoning under weak spatial and temporal consistency and to improve the ability to classify actions using partial body poses.

Motion modeling with masked body parts. We partition the human body into six parts: upper body, lower body, head, arms, legs, and hips, as shown in Fig. 2. One part is then randomly chosen to be masked.

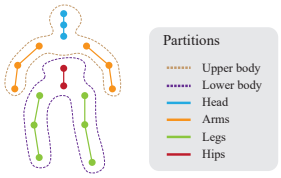


Figure 2. The human body is partitioned into six parts for motion modeling with masked body parts.

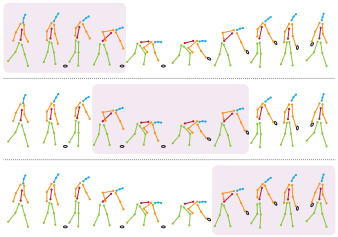


Figure 3. Masking different segments of a human motion has distinct effects. Top: masking the initial part of the motion. Middle: masking the middle segment. Bottom: masking the latter part of the motion.

This technique enables the pose encoder to learn human kinematics and anatomy more comprehensively than joint-level masking by capturing coordination not only among individual joints but also across multiple body parts. When masked body part motion modeling is combined with multi-modal alignment prediction, the pose encoder further learns the spatial relationships of human motion within the context of corresponding actions. This masking technique also simulates real-world scenarios where parts of the human body may be occluded in a camera's field of view.

Motion modeling with masked time windows. In addition to masked body parts, we propose another masking technique that masks the motion in the temporal domain with a varying window size between $[T_1, T_2]$. The starting frame of this masked time window also varies and is randomly chosen.

The method can be interpreted as follows: masking different segments of motion in the time domain serves distinct purposes. Masking the initial part of the motion helps infer the cause of the action. Masking the middle segment allows us to model the transition between different motion states. Masking the latter part aids in predicting the motion's intention and reconstructing human motion based on that predicted intention. Figure 3 provides a visualization of these interpretations.

Multi-modal alignment prediction. The dual-stream encoders are optimized together by contrasting the global

representation of text embeddings and pose embeddings at sample j:

$$s^j_{p,w} = \frac{exp(h^j_P \cdot h^j_{W+}/\tau)}{\sum_{i=0}^{K} exp(h^j_P \cdot h^j_{W_i}/\tau)}, \quad (1)$$

where h^j_{W+} is the corresponding text description of the action, $h^j_{W_i}$ for $i \in [0, 1, ..., K]$ includes one positive sample h^j_{W+} and K negative samples, and τ is the temperature.

Since motion-and-language datasets [18, 40] are highly imbalanced and human action recognition tasks are inherently non-deterministic, namely, motion and action labels do not have a one-to-one correspondence, we replace the cross-entropy in the infoNCE loss with a combination of focal loss [22] and KL divergence. The modified loss function is as follows:

$$\mathcal{L}_{con} = \sum_{j=1}^{M} (1 - s^j_{p,w})^{\gamma} y_j log(\frac{y_j}{s^j_{p,w}}), \quad (2)$$

where M is the number of overall samples and y_j is the ground-truth similarity score.

Objective function. During pretraining, we adopt several masking strategies: joint-level and frame-level masking 50% of the time to capture spatial consistency among neighboring joints and temporal consistency across adjacent frames; random body part masking 25% of the time; and random time window masking 25% of the time. Additionally, we introduce noises into the 2D pose sequences.

Alongside the multi-modal alignment loss, we perform 3D pose reconstruction from the corrupted 2D poses. The discrepancy between the predicted 3D pose sequence $\mathbf{X}$ and the ground-truth 3D pose sequence $\hat{\mathbf{X}}$ is penalized using the following loss functions:

$$\mathcal{L}_{3D} = \sum_{k=1}^{M} \sum_{t=1}^{T} \sum_{j=1}^{J} ||\hat{\mathbf{X}}^{\mathbf{k}}_{\mathbf{t,j}} - \mathbf{X}^{\mathbf{k}}_{\mathbf{t,j}}||_2, \quad (3)$$

$$\mathcal{L}_{v} = \sum_{k=1}^{M} \sum_{t=1}^{T} \sum_{j=1}^{J} ||\hat{\mathbf{V}}^{\mathbf{k}}_{\mathbf{t,j}} - \mathbf{V}^{\mathbf{k}}_{\mathbf{t,j}}||_2, \quad (4)$$

Where $\hat{\mathbf{V}}_{\mathbf{t}} = \hat{\mathbf{X}}_{\mathbf{t}} - \hat{\mathbf{X}}_{\mathbf{t-1}}$ and $\mathbf{V}_{\mathbf{t}} = \mathbf{X}_{\mathbf{t}} - \mathbf{X}_{\mathbf{t-1}}$.

The final pretraining loss is computed by combining the multi-modal alignment loss and the reconstruction loss functions as follows:

$$\mathcal{L} = \lambda_{con}\mathcal{L}_{con} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_v\mathcal{L}_v. \quad (5)$$

3.3. Fine-tuning

During pretraining, the parameters of both the pose encoder and text encoder are updated, but during fine-tuning, only the pose encoder is fine-tuned on the target dataset. Specifically, 2D poses are detected from videos and then lifted to 3D poses. The fine-tuning process is supervised by the reconstruction losses as specified in Eq. (3) and Eq. (4).

4. Experiments

4.1. Pretraining

The BABEL [40] dataset is a large dataset that includes language labels describing the actions performed in mocap sequences. It provides both sequence-level labels, which describe the overall action in the sequence, and frame-level labels, which detail fine-grained actions in every frame. Each frame in the sequence can have multiple corresponding fine-grained action labels. The motion sequences in the BABEL dataset are sourced from the large mocap dataset AMASS [28], which uses a common parameterization framework. Instead of obtaining videos and extracting 2D skeleton sequences from the mocap dataset, we project the 3D skeletons extracted from parametric models to 2D from both a side view and a front view, assuming an orthographic camera, as done in [62]. We align the body keypoint definitions with those of Human3.6M and convert the camera coordinates to pixel coordinates using the approach outlined in [7].

Each positive data pair for pretraining consists of a fine-grained frame-level action label and its corresponding motion sequence. For sequences labeled as "transition", we concatenate the motion sequence before and after the "transition", and use a template "transit from A to B" to construct new action labels. A and B are the corresponding action labels before and after the "transition". Negative data pairs are generated by randomly associating action labels with unmatched motion sequences.

The overall pretraining framework consists of three layers of pretrained BERT [9] as the text encoder and five layers of pretrained MotionBERT [62] as the pose encoder. The network is trained on a single NVIDIA H100 GPU with a batch size of 16 and a sequence length of 243 frames and for 300 epochs using an AdamW optimizer. The loss weights are set to $\lambda_{con} = 0.5$, $\lambda_{3D} = 1$, and $\lambda_v = 20$, and the model is trained for 300 epochs. Please refer to the Supplementary Material for more implementation details.

4.2. Fine-tuning for 3D human pose estimation

During fine-tuning for 3D human pose estimation, the text encoder is no longer required; only the pose encoder is fine-tuned on the target test dataset.

Datasets and evaluation metrics. We evaluated our methods on the public dataset Human3.6M [15] and MPI-INF-3DHP [31].

Human3.6M [15] contains 3.6 million video frames of human motion performed by professional actors in an indoor environment. Following previous works [36, 62], we use subjects 1, 5, 6, 7, and 8 for fine-tuning, and subjects

Table 1. Quantitative comparison of 3D human pose estimation on the Human3.6M dataset using detected 2D poses. The per-action performance is shown under P1, and the average metrics across P1, P2, and P3 are summarized. All methods use a temporal window of 243 frames. †indicates the use of temporal information, * represents diffusion-based methods, and ⋄ indicates pretraining-based methods. Results from other models are directly quoted from the respective papers.

Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Pur.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	P1 ↓	P2 ↓	P3 ↓
P-STMO [42] †	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	54.9
MixSTE [55] †	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	52.2
GLA-GCN [54] †	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	34.8	58.3
STCFormer [47] †	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	31.8	53.3
KTPFormer [36] †	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	51.8
TCPFormer [24] †	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.9	31.7	–
RePose [46] †	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**	**30.7**	47.7
DiffPose [13] †*	**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	–	47.1
D3DP [43] †*	37.3	39.5	35.6	37.8	41.3	48.2	39.1	37.6	49.9	52.8	41.2	39.2	39.4	27.2	27.1	39.5	31.6	50.3
ActionPrompt [59]	37.7	40.2	39.8	40.6	43.1	48.0	38.8	38.9	50.8	63.2	42.0	40.0	42.0	30.5	31.6	41.8	–	54.0
MotionBERT [62] †⋄	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	46.9	53.9	39.5	36.3	35.8	25.1	25.3	37.5	–	49.0
LangPose †⋄	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	31.3	**47.0**

Table 2. Quantitative comparison of 3D human pose estimation on the MPI-INF-3DHP dataset. All methods use a temporal window of 81 frames. †indicates the use of temporal information, * represents diffusion-based methods, and ⋄ indicates pretraining-based methods. Results from other models are directly quoted from the respective papers.

Method	PCK ↑	AUC ↑	P1 ↓
P-STMO [42] †	97.9	75.8	32.2
GLA-GCN [54] †	98.5	79.1	27.7
STCFormer [47] †	98.7	83.9	23.1
DiffPose [13] †*	98.0	75.9	29.1
D3DP [43] †*	97.7	78.2	29.7
KTPFormer [36] †*	98.9	85.9	16.7
MotionAGFormer-L [30] †	98.2	85.3	16.2
TCPFormer [24] †	**99.0**	**87.7**	**15.0**
RePose [46] †	98.3	86.7	15.5
LangPose (Ours) †⋄	98.9	87.0	15.5

9 and 11 for testing. We report results using the mean per joint position error (MPJPE), which measures the average Euclidean distance between predicted and ground-truth joint positions, typically in millimeters. We also report the Procrustes-aligned MPJPE (P-MPJPE), where MPJPE is calculated after applying Procrustes alignment to the predicted and ground-truth positions. Futhermore, we observe that three actions are consistently the most challenging to predict across state-of-the-art methods as shown in Tab. 1: *Sitting* (involving external objects), *Sitting Down* (prone to self-occlusions), and *Taking Photo* (less repeatable motion). We also find that these actions are the most heavily occluded, according to a rough estimation of their visibility[1]. We denote the average MPJPE over these three actions as H-MPJPE. For simplicity, we refer to MPJPE, P-MPJPE, and H-MPJPE as P1, P2, and P3, respectively.

MPI-INF-3DHP [31] is another large-scale public dataset. This dataset contains recordings from 14 cameras capturing 8 actors performing 8 activities for the training set and 7 activities for evaluation. Following the setting of previous work [30, 36, 42], our evaluation metrics included the area under the curve (AUC), percentage of correct keypoints (PCK), and mean per-joint position error (MPJPE).

Performance comparison on the Human3.6M dataset. We evaluated LangPose on the Human3.6M dataset [15], and reported P1, P2, and P3 errors in millimeters. The 2D skeletons are extracted using the Stacked Hourglass network [34], and the pose encoder is fine-tuned on the Human3.6M training set. As shown in Tab. 1, LangPose achieves competitive performance compared to prior temporal models, with particularly large improvements on the three most challenging actions. This highlights the effectiveness of incorporating semantic information when spatial and temporal consistency alone is insufficient. Moreover, LangPose achieves performance comparable to diffusion-based methods, while being more efficient during inference, as diffusion models rely on iterative refinement.

Among all baselines, MotionBERT [62] is most closely related to our framework in terms of its pretraining and finetuning strategy. As shown in Tab. 1, our model outperforms MotionBERT by a large margin, especially on the hardest actions, demonstrating the benefits of integrating semantic information during pretraining. It is worth noting that the motion sequences from BABEL [40] originate from the AMASS dataset [28], which is already included in the pretraining data of MotionBERT. Although our pretraining includes additional action labels to facilitate semantic alignment, no extra motion sequences are used beyond what is available to MotionBERT.

Performance comparison on the MPI-INF-3DHP dataset. To evaluate the generalization capability of LangPose, we assessed its performance on the MPI-INF-3DHP dataset. Compared to Human3.6M, the test set of MPI-INF-3DHP features more diverse motion types

[1] See the Supplementary Material for details of the visibility estimation.

and greater variation in camera viewpoints. For instance, MPI-INF-3DHP includes sports-related poses that are absent from Human3.6M. Additionally, approximately 67% of the poses in Human3.6M fall under the "stand/pose" category, while only 25% correspond to "sit" and 8% to "crouch"—the latter being more complex and challenging to model [31].

Following prior works [30, 36], we fine-tuned the pose encoder using ground-truth 2D poses as input. As shown in Tab. 2, LangPose achieves SOTA level performance, obtaining a PCK of 98.9%, an AUC of 87.0%, and an MPJPE of 15.5 mm. These results surpass most existing methods, including both temporal-based and diffusion-based approaches, highlighting the strong generalization ability of Pose.L

Table 3. Ablation study of LangPose on MPI-INF-3DHP and Human3.6M datasets. Results are reported using P1 and P3 metrics (in mm). We refer to masked motion modeling as MMM and multi-modal alignment prediction as MAP.

Configuration	Modules enabled		MPI (P1)	H36M (P1)	H36M (P3)
	MMM	MAP			
Baseline	✗	✗	16.7	37.5	49.0
Baseline + MMM	✓	✗	16.2	36.8	47.6
Baseline + MAP	✗	✓	15.7	37.3	47.9
MMM + MAP (ours)	✓	✓	**15.5**	**36.7**	**47.0**

Table 4. Ablation study comparing the proposed modified InfoNCE loss with the original InfoNCE formulation on MPI-INF-3DHP (abbrev. MPI) and Human3.6M (abbrev. H36M). MMM denotes masked motion modeling. Results are reported using P1 and P3 metrics (mm).

Method (Loss)	MPI (P1)	H36M (P1)	H36M (P3)
Baseline (w/o InfoNCE)	16.7	37.5	49.0
Baseline + original InfoNCE + MMM	16.3	37.5	48.5
Baseline + MMM + **Modified InfoNCE (ours)**	**15.5**	**36.7**	**47.0**

Ablation study. To verify the effectiveness of our proposed framework, we conducted ablation studies on the MPI-INF-3DHP dataset (T = 81) using ground-truth 2D poses as inputs and on the Human3.6M dataset (T = 243) using detected 2D poses. Table 3 presents the results for each component added to our framework, reported as P1 for the Human3.6M dataset and as both P1 and P3 for the MPI-INF-3DHP dataset (in mm).

For the baseline, we used the pretrained MotionBERT model fine-tuned on the MPI-INF-3DHP and Human3.6M datasets, respectively. The pretrained MotionBERT weights were obtained from the official GitHub page. On MPI-INF-3DHP, pretraining with the proposed masked time windows and masked body parts motion modeling (MMM) reduced the error by 0.5mm compared to the baseline. Incorporating random frame and joint-level masking, random noise, and multi-modal alignment prediction further reduced the error by 1.0mm (MAP). Combining these strategies (MMM + MAP) resulted in a total error reduction of 1.2mm from the baseline. Similarly, on Human3.6M, pretraining with masked time windows and masked body parts motion modeling reduced the error by 0.7mm, and adding random frame and joint-level masking, noise, and multi-modal alignment reduced the error by an additional 0.2mm. The combination of these techniques led to a total error reduction of 0.8mm from the baseline. P3 is additionally reported to highlight the performance of the proposed modules on the most difficult actions in Human3.6M, where the joint contribution of masked motion modeling and semantic alignment proves especially beneficial.

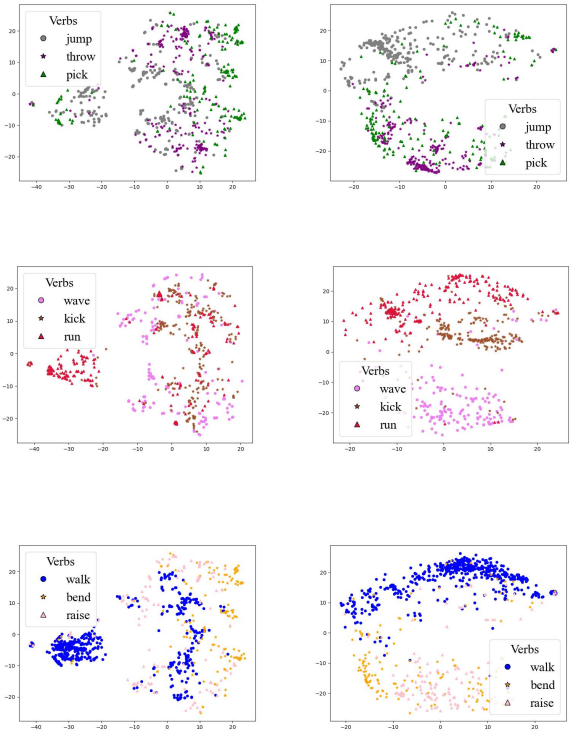
To further validate the effectiveness of our proposed loss design, we conduct an additional ablation study by comparing the modified InfoNCE loss with the original InfoNCE formulation. Specifically, in the modified InfoNCE loss, the cross-entropy term is replaced with a KL-divergence objective, and a focal loss component is incorporated to better handle hard examples. As shown in Tab. 4, this modification leads to consistent performance improvements across both benchmarks. Moreover, we observe that the original InfoNCE loss tends to suffer from overfitting during training, whereas our modified InfoNCE loss successfully mitigates this issue, providing a stronger optimization signal and yielding more robust representations for downstream 3D human pose estimation.

4.3. Qualitative results on global representation learning

To verify LangPose's ability to learn rich semantic information about actions, we visualize the global representations of pose embeddings in Fig. 4. Ideally, semantically similar actions should have closer embeddings, while different actions should be farther apart. For instance, 'throw' and 'pick' both emphasize upper body movements, and thus their representations are close. In constrast, 'jump', which emphasizes lower body movement, is distinctly separate from them. Similarly, 'run' and 'kick' are similar from leg movements and are positioned near each other, while 'wave' focused on the upper body, is more distant. Additionally, 'bend' and 'raise' are closely related, whereas 'walk' is farther from these actions.

As shown in Fig. 4, LangPose effectively groups semantically similar actions closer together while pushing semantically different actions farther apart. In contrast, MotionBERT's global representation lacks a clear pattern, with significant overlap between both similar and distinct actions.

Please note that there is some diversity within the same action due to variations in concurrent actions and different forms and styles of the actions. For example, "kicking while jumping" differs from "kicking while standing," and the single action label with the verb 'kick' include motion sequences for both "kicking a soccer ball" or "kicking with the right leg back and up".



(a) MotionBERT (b) LangPose

Figure 4. The t-SNE visualization of global representations of motion embeddings on the BABEL dataset: on the left (a) are embeddings obtained through random joint-level and frame-level masking, as proposed by MotionBERT [62], with average pooling of all pose features in the sequence; on the right (b) are embeddings obtained through LangPose.

5. Discussion

Learning sufficient action semantics during text embedding alignment with motion representation requires a large dataset. When the action classes in the pretraining dataset differ from those in the target dataset, the benefits of fusing action knowledge through pretraining may diminish[2]. Despite this limitation, beyond 3D human pose estimation, LangPose can be extended to other human motion-related tasks, such as motion prediction and completion. For motion prediction, the model can use historical motion to predict intentions, thereby guiding future movements. Similarly, for motion completion, the model can learn transitions between starting and ending poses, constrained by the corresponding actions. The framework's ability to learn rich semantic information about actions also has potential benefits for downstream tasks such as motion captioning from noisy inputs and generation from text descriptions. These applications are reserved for future work.

[2]See the Supplementary Material for more discussion.

6. Conclusion

In this work, we demonstrate how understanding human motion from an action perspective can enhance 3D human pose estimation. We propose a novel pretraining and fine-tuning strategy that embeds action cues during the pretraining stage. Experimental results confirm the effectiveness of the proposed masked motion modeling and multimodal alignment tasks. Additionally, the ability to incorporate semantic information into global representation learning shows potential for benefiting downstream tasks such as human motion captioning, generation, and prediction.

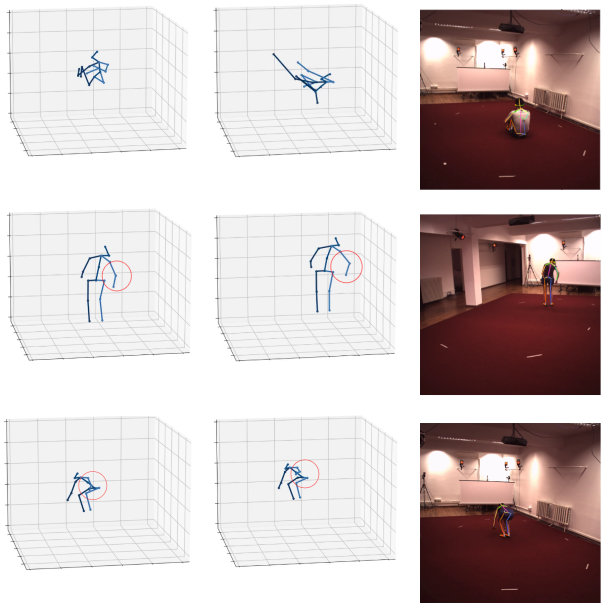


Figure 5. Qualitative results of LangPose compared to the baseline model MotionBERT. The left column shows the results from LangPose, the middle column shows the results from MotionBERT for the same frames, and the right column displays the corresponding original 2D pose inputs.

References

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: A spatio-temporal cross attention transformer for human action recognition. In *2023*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 3319–3328, 2023. 2

[2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2272–2281, 2019. 1, 2

[3] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 2

[4] Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10631–10638, 2020. 1, 2

[5] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022. 2

[6] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2262–2271, 2019. 1, 2

[7] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. Locally connected network for monocular 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1429–1442, 2022. 5

[8] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Poseembroider: Towards a 3d, visual, semantic-aware human pose representation. In *ECCV (71)*, pages 55–73, 2024. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 4, 5

[10] Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition, 2024. 2

[11] Qihang Fang, Chengcheng Tang, Bugra Tekin, Shugao Ma, and Yanchao Yang. Humocon: Concept discovery for human motion understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[12] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 3

[13] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13051, 2023. 6

[14] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 3

[15] 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, 2014. 3, 5, 6

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[17] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] Franziska Krebs, Andre Meixner, Isabel Patzer, and Tamim Asfour. The kit bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506, 2021. 2, 3, 5

[19] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Computer Vision – ACCV 2014*, pages 332–347, Cham, 2015. Springer International Publishing. 2

[20] Zhe Li, Weihao Yuan, Yisheng HE, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, , Yuan Dong, Zilong Dong, and Laurence T. Yang. Lamp: Language-motion pre-training for motion generation, retrieval, and captioning. *International Conference on Learning Representations (ICLR)*, 2025. 3

[21] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 2023. 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2, 5

[23] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3671–3680, 2017. 2

[24] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:5478–5486, 2025. 2, 6

[25] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5063–5072, 2020. 2

[26] Diogo C. Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5137–5146, 2018. 2

[27] Diogo C. Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2752–2764, 2021. 2

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 5, 6

[29] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[30] 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 (WACV)*, pages 6920–6930, 2024. 6, 7

[31] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 5, 6, 7

[32] Nicola Messina, Jan Sedmidubsky, Fabrizio Falchi, and Tomás Rebok. Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 2420–2425, New York, NY, USA, 2023. Association for Computing Machinery. 3

[33] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019. 2

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 6

[35] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7745–7754, 2019. 1, 2

[36] 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*, pages 1123–1132, 2024. 1, 2, 5, 6, 7

[37] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[38] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[39] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[40] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, 2021. 2, 3, 5, 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[42] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 461–478. Springer, 2022. 6

[43] 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 (ICCV)*, pages 14761–14771, 2023. 6

[44] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vision*, 87(1–2):4–27, 2010. 3

[45] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Computer Vision – ECCV 2018*, pages 536–553, Cham, 2018. Springer International Publishing. 2

[46] Ziming Sun, Yuan Liang, Zejun Ma, Tianle Zhang, Linchao Bao, Guiqing Li, and Shengfeng He. Repose: 3d human pose estimation via spatio-temporal depth relational consistency. In *ECCV (19)*, pages 309–325, 2024. 6

[47] 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*, pages 4790–4799, 2023. 6

[48] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022. 3

[49] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *Computer Vision – ECCV 2020*, pages 764–780, Cham, 2020. Springer International Publishing. 1, 2

[50] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *The*

Thirteenth International Conference on Learning Representations, 2025. 3

[51] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10276–10285, 2023. 3

[52] Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2866–2874, 2022. 2

[53] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. 2

[54] Bruce X.B. 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 (ICCV)*, pages 8818–8829, 2023. 6

[55] 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 (CVPR)*, pages 13232–13242, 2022. 1, 2, 6

[56] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[57] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34 (12):5586–5609, 2022. 2

[58] Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025. 3

[59] Hongwei Zheng, Han Li, Bowen Shi, Wenrui Dai, Botao Wang, Yu Sun, Min Guo, and Hongkai Xiong. Actionprompt: Action-guided 3d human pose estimation with text and pose prompting. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2657–2662, 2023. 2, 3, 6

[60] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefining topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[61] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016. 2

[62] 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*, 2023. 1, 2, 4, 5, 6, 8