Human Motion Prediction Using Spatial Semantics of Objects: The PADSO Approach

Michael Vanuzzo * **Ferdinando Pompanin** * **Mattia Guidolin** * **Monica Reggiani** *

* *Department of Management and Engineering, University of Padua, Padua, Italy (e-mail: michael.vanuzzo@phd.unipd.it, ferdinando.pompanin@studenti.unipd.it, mattia.guidolin@unipd.it, monica.reggiani@unipd.it)*

Abstract: Human Motion Prediction (HMP) is a key capability for a wide range of applications, including collaborative robotics, rehabilitation, and human-machine interaction. Despite advances in deep learning for HMP, most existing methods overlook the environmental context. In this paper, we present Prediction of Actions through Data about a Single Object (PADSO), a novel HMP model that enriches input human poses with spatial semantic information by incorporating environmental object data, thereby enabling joint processing of human pose and object pose predictions. Experimental evaluation on the GRasping Actions with Bodies (GRAB) dataset shows that PADSO significantly improves prediction accuracy over both a non-spatially aware version of the same architecture and the Zero-Velocity baseline. These results highlight the critical role of spatial context in enhancing the robustness of HMP. Furthermore, the efficient design of PADSO enables real-time inference, a critical feature for deployment in real-world scenarios.

Keywords: Human Motion Prediction, Spatial Semantics, Context Awareness, Real-Time

1. INTRODUCTION

Human Motion Prediction (HMP) is a critical research area with a wide range of applications in fields such as collaborative robotics, rehabilitation, and human-machine interaction (Lyu et al. (2022); Deng and Sun (2024)). Reliable prediction of future human poses enables intelligent systems to anticipate human actions, thereby improving safety, operational efficiency, and user experience. Despite its importance, full-body HMP remains a highly challenging task, primarily due to the complexity of human biomechanics, characterized by numerous degrees of freedom, and the inherent unpredictability of human actions. Moreover, the presence of obstacles in cluttered environments can further hinder the motion capture system's ability to correctly estimate human poses (Guidolin et al. (2024)).

Recent advances in Deep Learning (DL) have greatly improved HMP, as end-to-end learning approaches can directly capture complex motion dynamics from large-scale datasets. This capability enables models to extract meaningful spatio-temporal patterns, including gaits, transitions, and gestures, leading to significant improvements in prediction performance over traditional techniques (Brand and Hertzmann (2000); Taylor et al. (2006)). However, despite these advances, most existing methods still exploit only past human pose information, often neglecting the broader spatial semantic context. Critical environmental cues, such as the objects being manipulated by the user or the obstacles in the environment, are often overlooked. These non-spatially aware models can hinder prediction reliability, especially in structured settings such as collaborative robotics, where task-level spatial semantics and ob-

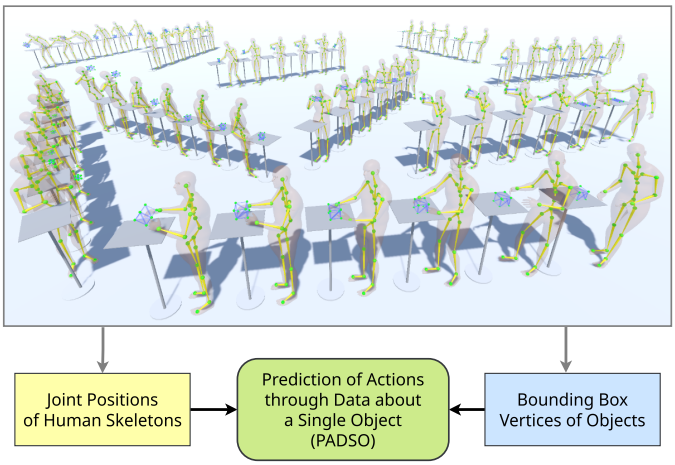

Fig. 1. The proposed approach combines past human pose data with spatial semantic information by incorporating environmental object data. In the pose sequences, both the positions of the skeletal joints and the vertices defining the bounding boxes of objects are highlighted. These are jointly exploited by the proposed PADSO model.

⋆ This study was funded by Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004) within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership, CUP: C93C22005280001.

ject affordances provide valuable insights for interpreting and predicting human motion.

To overcome these limitations, we propose a novel HMP approach that explicitly integrates spatial semantic information into the pose prediction process, as shown in Fig. 1. Specifically, we introduce Prediction of Actions through Data about a Single Object (PADSO), a DL model designed to jointly process past motion sequences and spatial semantic information by including data about the objects in the environment. Our methodology is based on the observation that human motion is profoundly influenced by the scene in which it unfolds and by the proximity of relevant objects, which can serve as constraints or targets for interaction. For example, a person moving toward a tool is likely to manipulate it, while regions occupied by obstacles such as tables or machinery are impassable and thus constrain movement trajectories. By making the model spatially aware, we aim to improve the plausibility and accuracy of motion predictions. Furthermore, we specifically focus on real-world industrial scenarios as our primary use case, emphasizing the importance of real-time HMP, another aspect often overlooked in the current literature. Indeed, our goal is not only to improve prediction accuracy, but also to support time-critical applications where anticipating human motion has a direct impact on both safety and efficiency.

The main contributions of this paper are as follows:

- We propose PADSO, a DL model capable of incorporating spatial semantic information, specifically the object with which the human is interacting.
- We show that including spatial semantic object data in the model's inputs leads to significant improvements in motion prediction accuracy.
- We target real-world industrial settings, emphasizing real-time prediction capabilities to support practical applications that require online spatially aware HMP.

2. RELATED WORKS

Recent approaches to HMP use DL architectures designed to identify complex patterns in past motion sequences, thereby enabling accurate inference of subsequent poses (Lyu et al. (2022); Deng and Sun (2024)). Notable progress has been made, primarily through models based on architectures such as Recurrent Neural Networks (RNNs) (Yu et al. (2022); Pavllo et al. (2018); Wang et al. (2021); Guo et al. (2022)) and Graph Convolutional Neural Networks (GCNNs) (Ma et al. (2022); Li et al. (2020, 2021); Yang et al. (2024)), each introducing unique strategies to improve predictive performance. RNNs are particularly effective for sequence-to-sequence applications due to their ability to model temporal dependencies and capture the evolution of motion over time. Conversely, GCNNs excel at modeling the spatial relationships inherent in the human skeletal structure by explicitly representing the connectivity between body joints.

In parallel, recent research has explored the integration of additional data modalities to improve the accuracy of human motion or trajectory prediction. For instance, Cao et al. (2020) use images of the scene alongside sequences of 2D human poses as input, motivated by the notion that human movement is typically goal-directed. In another approach, Lou et al. (2024) extend their HMP model by incorporating gaze direction and local saliency point clouds, thereby increasing the complexity of the contextual information used. This work highlights the importance of multimodal inputs in predicting human behavior in complex 3D environments. However, the point cloud data required as input presents practical challenges in industrial settings where such information may be difficult to acquire. Similarly, Lee (2024) represents both human poses and surrounding objects as point clouds processed by PointNet (Qi et al. (2017)). Their method uses a cross-attention mechanism to capture mereotopological relationships between hands and nearby objects. However, this approach shares the aforementioned limitations regarding the accessibility and reliability of point cloud data in operational environments. Finally, Corona et al. (2020) integrate contextual information in the form of object bounding boxes. This information is processed by a dedicated RNN-based branch, which operates in parallel with another branch that implements an HMP model using only the sequence of past human poses. The reliance on less complex input representations compared to full-scene point clouds potentially enhances the system's deployability in real-world scenarios. However, the suitability of this approach for real-time applications remains unaddressed in their evaluation.

Despite significant progress in state-of-the-art HMP models, a significant gap remains in translating these advances into industry-ready solutions. Many current approaches focus primarily on increasing model sophistication, often overlooking essential practical constraints such as inference latency. This limitation hinders deployment in scenarios where the ability to achieve real-time predictions is critical. To address these challenges, our research prioritizes the design of a computationally efficient model optimized for online inference, thereby enhancing its suitability for real-world applications.

3. METHODOLOGY

The objective of HMP is to predict a sequence of future human poses. Let us represent a human pose as $x_t = (r_1, r_2, \ldots, r_J) \in \mathbb{R}^{J \times D}$, where each $r_j \in \mathbb{R}^D$ specifies the state of the j-th joint in a D-dimensional space. In this study, the state of each joint is characterized by its 3D position relative to a reference coordinate system centered on the subject's pelvis. This pose representation ensures consistency across different instances of a given posture, regardless of the absolute position at which the pose is performed. The set of J joints is chosen to capture the most relevant features of human motion, ensuring that the representation retains sufficient information for accurate predictive modeling.

Typically, the input to an HMP model consists of sequential data corresponding to past human motions. Formally, given a sequence of N past poses $\mathcal{X}_{0:N-1} = [x_0, x_1, \ldots, x_{N-1}]$, the task is to predict the subsequent M future poses $\mathcal{X}_{N:N+M-1}$. The time interval between consecutive poses is constant and defined by the chosen frame rate.

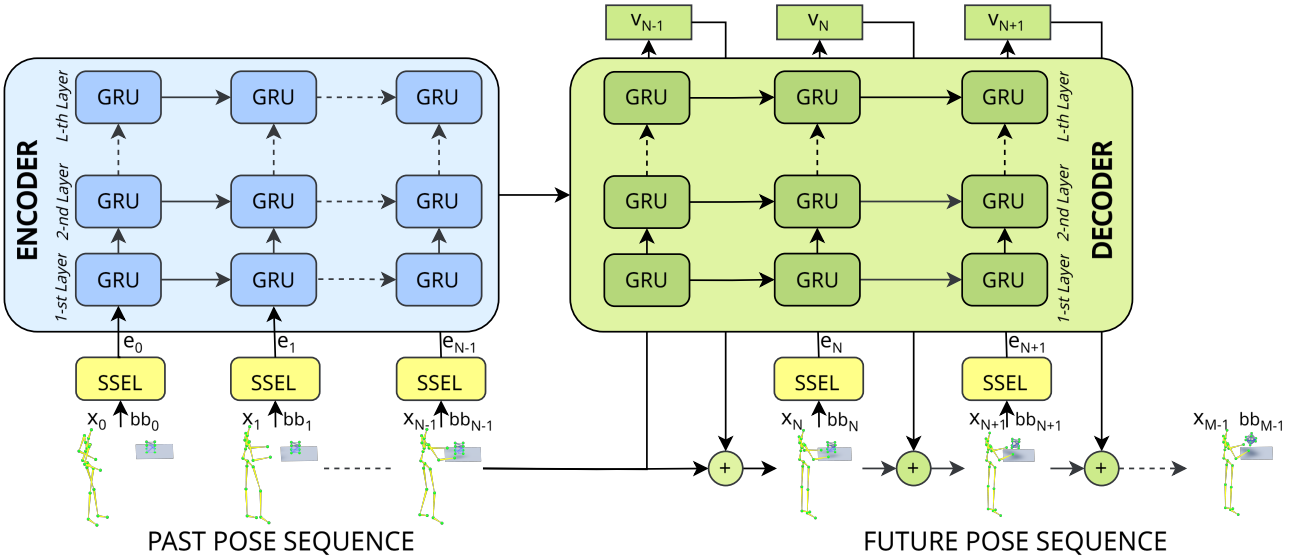

Fig. 2. Architecture of the PADSO model. The past human poses x_t are enriched with the bounding box position of the nearby object bb_t, and each frame is processed by the Spatial Semantic Enrichment Layer (SSEL) to generate an embedding $e_t \in \mathbb{R}^{G \times H}$. These embeddings are used in an encoder-decoder architecture based on GRUs, which iteratively predicts the variations v_t relative to the last pose, producing the sequence of future human poses.

While past pose sequences provide the core data for HMP, incorporating additional contextual information, denoted as spatial semantics, can significantly improve the accuracy and robustness of the predictions. Spatial semantics refers not only to physical contextual attributes, such as the positions and dimensions of nearby objects, but also to their intrinsic properties, which could inform the model about potential interactions or constraints these elements may impose on human motion.

3.1 PADSO Architecture

The novel HMP model proposed in this work integrates historical human pose data and spatial semantic context. As illustrated in Fig. 2, the PADSO architecture adopts an encoder-decoder framework. The main advancement of the model is the Spatial Semantic Enrichment Layer (SSEL), which enriches the past human pose sequences by incorporating spatial semantic information.

Spatial Semantic Enrichment Layer The SSEL is implemented using a graph convolutional layer proposed by Kipf and Welling (2017) and is designed to integrate the information about the joint states, denoted as x_t, with the spatial data of the nearby object, represented by its bounding box bb_t. The input processed by the SSEL consists of a graph with $G = J + B$ nodes, where J nodes represent the human pose and B nodes correspond to the vertices defining the object's bounding box. Each node in the graph is represented by a feature vector composed of the concatenation of its 3D spatial coordinates and a one-hot encoding containing the semantic information indicating whether the node corresponds to a human joint or an interacting object.

For graph connectivity, nodes corresponding to human joints are connected according to the parent-child hierarchy defined by the Skinned Multi-Person Linear Model (SMPL) representation (Loper et al. (2015)). Bounding box vertices are interconnected to describe the object's geometry. In addition, each bounding box vertex is linked to the hand joints of the human pose. This graph topology allows for an effective exchange of information during the convolution process in the SSEL. The impact is particularly significant for the upper body joints, which are most directly affected by the interaction with the object. The graph is processed by the SSEL, which outputs an embedding $e_t \in \mathbb{R}^{G \times H}$, where H denotes the embedding size chosen for the graph convolution layer.

Human Motion Encoder-Decoder PADSO uses an RNN-based encoder-decoder architecture with Gated Recurrent Units (GRUs) to generate future human poses, following the approach initially proposed by Martinez et al. (2017).

The encoder takes as input the sequence of observed poses $\mathcal{X}_{0:N-1}$, enriched with the spatial semantic information processed by the SSEL. The output of the encoder is a hidden state that initializes the decoder. The decoder then iteratively predicts future poses using its hidden state and the previous pose. During the first iteration, the last observed pose in the input sequence, corresponding to the $(N-1)$-th time step, is used to start the decoding phase.

The outputs generated by the decoder are used in two ways: to extract the predicted future frames, and to provide the input to the SSEL module for the subsequent prediction step. To increase the stability and accuracy of the predictions, residual connections are introduced. In this approach, the output of the decoder is interpreted as a motion variation, called $v_{t-1} = [v_{x,t-1}, v_{bb,t-1}]$, relative to the previous pose. The predicted human pose, x_t, and the object pose, bb_t, are then computed by adding this variation to their respective previous poses, x_{t-1} and bb_{t-1}. This process is expressed mathematically as:

$$x_t = x_{t-1} + v_{x,t-1}, \quad bb_t = bb_{t-1} + v_{bb,t-1} \tag{1}$$

3.2 Loss Function

The training loss function $\mathcal{L}$ consists of two terms representing the prediction errors for human joint positions and object bounding box coordinates. These components are combined using a weighted sum, where the two weights, $\alpha \in (0, 1]$ and $1 - \alpha$, determine their respective contributions to the total loss. This formulation ensures accurate predictions of both human joints and object positions, with their relative importance controlled by α.

The loss function is formally defined as:

$$\mathcal{L} = \frac{1}{M} \sum_{t=N}^{N+M-1} \alpha \left\| x_{t,\mathrm{pr}} - x_{t,\mathrm{gt}} \right\|_2^2 + (1-\alpha) \left\| bb_{t,\mathrm{pr}} - bb_{t,\mathrm{gt}} \right\|_2^2 \tag{2}$$

where $x_{t,\mathrm{pr}}$ and $x_{t,\mathrm{gt}}$ represent the predicted and ground truth positions of the human joints relative to frame t, while $bb_{t,\mathrm{pr}}$ and $bb_{t,\mathrm{gt}}$ denote the predicted and ground truth vertices of the object's bounding box.

3.3 Hyperparameter Optimization

The training procedure is configured using hyperparameters determined through an exhaustive grid search. The optimal configuration includes an encoder-decoder architecture of 4 layers each and an embedding dimension of $E = 4$ for the nodes within the SSEL. The model is trained using the Adam optimizer with a learning rate of 10^{-3} and no weight decay. Finally, the weighting factor α in the loss function is set to 0.8 to prioritize accuracy in predicting future human poses. This choice is based on a targeted test for the value of α, specifically comparing the inclusion ($\alpha < 1$) versus exclusion ($\alpha = 1$) of the bounding box component in the loss function. The results showed that including the bounding box term positively influences the model's final performance.

4. EVALUATION PROTOCOL

This section describes the dataset employed in the experiments, the preprocessing steps taken to prepare the data, and the metrics used to evaluate the proposed spatially aware HMP model.

4.1 Dataset

HMP models that incorporate spatial semantics remain underexplored, in part due to the lack of sufficiently large datasets containing detailed information about objects in the environment where human motion occurs. The GRasping Actions with Bodies (GRAB) dataset acquired by Taheri et al. (2020) is one of the few datasets that provide comprehensive data to enable spatially aware HMP. The dataset is specifically designed to study human-object interactions and contains motion sequences of 10 subjects interacting with 51 different objects. Each sequence is recorded at a frame rate of 120 fps and captures the interaction between a subject and an object, typically placed on a table. The dataset provides detailed 3D meshes, along with the position and orientation of each object, and a label describing the type of item. This combination of attributes provides comprehensive information regarding the spatial semantics of the scene. Finally, human poses are represented using the SMPL model (Loper et al. (2015)), which is increasingly being adopted as the standard for body representation. This widespread adoption makes the GRAB dataset preferable to other comparable datasets, such as the KIT Whole-Body Human Motion Database by Mandery et al. (2015) and the KIT Bimanual Manipulation Dataset by Dreher et al. (2019).

4.2 Data Preprocessing

The GRAB dataset is divided into 80 % training, 10 % validation, and 10 % test sets, with the division reflecting a balanced representation of object types and interactions across the subsets. Motion sequences are downsampled to 25 fps, following standard practices in HMP (Lyu et al. (2022)). Sequences of 75 frames, corresponding to 3 s of motion data, are extracted. For each sequence, the first 50 frames serve as input and represent past human motion, while the last 25 frames serve as ground truth to evaluate the model's accuracy in its prediction. Thus, the model is trained to predict future motion up to a time horizon of 1 s. To preprocess the spatial context, axis-aligned bounding boxes are computed for all objects using the meshes provided in GRAB. This results in a compact representation of their positions and occupied volumes. Using forward kinematics, the axis-angle representation of the joints in SMPL is converted to Cartesian coordinates. A coordinate system transformation then expresses the joint positions and the object bounding boxes relative to the reference frame located at the person's pelvis.

4.3 Metrics

We evaluate prediction accuracy using Mean Per Joint Position Error (MPJPE), a standard metric from the literature (Deng and Sun (2024); Vanuzzo et al. (2025)). It measures the Euclidean distance between the predicted and ground truth joint positions, providing a straightforward metric of spatial accuracy. MPJPE is defined as:

$$\mathrm{MPJPE_t} = \frac{1}{K \cdot J} \sum_{k,j} |\widehat{p}_{t,k,j} - p_{t,k,j}|_2 \tag{3}$$

where $\widehat{p}_{t,k,j}$ and $p_{t,k,j}$ denote the predicted and ground truth positions of joint j in the k-th sequence at frame t.

5. EXPERIMENTS

This section introduces the models used as references for comparing the performance of the proposed PADSO model and presents the experimental results obtained.

5.1 Comparison Models

To evaluate the impact of integrating spatial semantic data, we compare PADSO with a modified version that excludes object information. This non-spatially aware variant is referred to as subject trajectory only PADSO (sto-PADSO). In sto-PADSO, the number of input nodes G of the SSEL is limited to J, since each node exclusively represents the 3D coordinates of the joints without any one-hot encoding in the node features. To ensure a fair

Table 1. MPJPE for the GRAB dataset comparing PADSO, sto-PADSO, and the ZeroVel baseline at different time horizons.

	MPJPE [mm]					
Time [ms]	120	200	400	600	800	1000
Zero-Velocity	18.2	30.1	56.3	76.7	93.7	108.8
sto-PADSO	18.7	27.6	45.7	57.8	67.5	78.2
PADSO	**15.1**	**23.9**	**40.9**	**52.2**	**59.9**	**68.2**

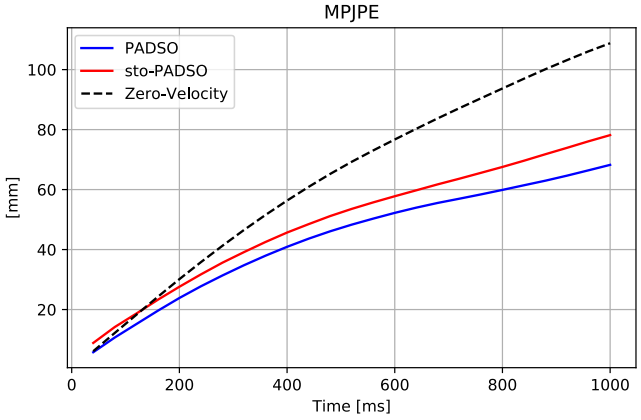

Fig. 3. Trend of the MPJPE for the GRAB dataset comparing PADSO, sto-PADSO, and the ZeroVel baseline.

comparison across models, a dedicated hyperparameter grid search is conducted to optimize sto-PADSO. The optimization resulted in an encoder-decoder architecture with 4 layers per module and an embedding dimension of $E = 16$ for the node representations within the SSEL. The model training process uses the Adam optimizer with a learning rate of 10^{-3} and no weight decay. Finally, the weighting factor α in the loss function is set to 1 to reflect the absence of bounding box data in this configuration.

We also include in our experimental comparisons the Zero-Velocity (ZeroVel) baseline, originally introduced by Martinez et al. (2017). The ZeroVel model generates future pose predictions by repeating the last observed pose, assuming that the human pose remains static. Despite its simplicity, the ZeroVel baseline showed remarkably strong performance when it was first introduced, outperforming more complex models (Martinez et al. (2017)). Its relevance continues today, as it remains a widely used baseline for evaluating the effectiveness of newly developed HMP models (Barquero et al. (2023)).

5.2 Experimental Results

Table 1 and Fig. 3 show the prediction errors obtained using PADSO, along with the results of sto-PADSO and the ZeroVel baseline. The values are given in terms of MPJPE over a prediction horizon of 1000 ms. The metric is calculated at uniformly distributed time stamps, specifically 200 ms, 400 ms, 600 ms, 800 ms, and 1000 ms. Additionally, a data point at 120 ms, corresponding to the third frame, is included to provide a more detailed analysis of the model's performance during the early stages of prediction.

Both PADSO and sto-PADSO show superior performance compared to the ZeroVel baseline in terms of motion prediction accuracy. In particular, the PADSO model achieves a 27.35 % reduction in MPJPE at 400 ms and a 37.32 % reduction at 1000 ms when compared to the ZeroVel baseline. These results highlight the effectiveness of the proposed architecture for HMP, even in scenarios with a limited amount of training data. A direct comparison shows that PADSO reduces MPJPE by 10.50 % compared to sto-PADSO at 400 ms, and by 12.79 % at 1000 ms. As highlighted in Table 1, although the percentage improvement for longer-term predictions decreases due to the increase in absolute prediction error, it is evident that the absolute error reduction grows consistently over longer time horizons. This observation highlights the significant contribution of incorporating spatial semantics to improve the accuracy of HMP. By providing contextual information that inherently reduces the uncertainty surrounding potential future actions, the addition of spatial semantic data proves particularly beneficial in extending the prediction horizon.

An analysis of the inference times shows that PADSO is suitable for real-time applications. Specifically, the model achieves an average inference time of 14.21 ms, with a standard deviation of 0.25 ms, allowing the generation of 1 s of predicted poses at a frame rate of ∼70 fps. Using a smaller prediction window of 400 ms, corresponding to 10 frames, the average inference time decreases to 7.75 ms, with a standard deviation of 0.07 ms. This corresponds to a frame rate of ∼130 fps, further establishing the efficiency of the proposed model. All experiments are performed on a consumer-grade desktop PC equipped with a 13th Gen Intel Core i7-13700 CPU and an NVIDIA GeForce RTX 4070 Ti GPU with 12 GB of VRAM.

6. CONCLUSIONS

This work investigates the impact of integrating spatial semantic information into HMP models. Specifically, our proposed model, PADSO, incorporates environmental context by encoding the object with which the person is interacting as a bounding box. Experimental evaluations show that incorporating object data significantly improves the prediction accuracy in terms of MPJPE. Both PADSO and its non-spatially aware counterpart, sto-PADSO, outperform the ZeroVel baseline, demonstrating the effectiveness of the architecture for HMP. Furthermore, PADSO achieves higher accuracy over all time horizons, with improvements becoming more pronounced as the prediction horizon lengthens, effectively reducing long-term errors. These results highlight the importance of exploiting spatial semantics in HMP. Indeed, by integrating contextual information from human-object interactions, PADSO reduces uncertainty and improves long-term prediction reliability, especially in scenarios where the surrounding context strongly influences human motion. Another key feature of PADSO is its ability to perform real-time inference, which is critical for industrial applications where accurate and timely predictions ensure operational efficiency and safety.

The current limitations of the proposed architecture include its ability to handle only a single object. While this work provides a solid foundation for future development,

real-world industrial scenarios typically involve operators interacting with multiple objects of different types and purposes. Future work will therefore focus on extending PADSO to handle multiple interacting objects and incorporate environmental obstacles. Leveraging the graph-based structure of the SSEL, these extensions will aim to improve the model's ability to interpret complex scenarios. Incorporating more detailed labels for object types will also provide richer contextual input and further improve the model's understanding and decision-making capabilities. Another limitation is the general lack of large-scale datasets that include object information. To overcome this, we plan to either expand the number of datasets being used or acquire new, dedicated datasets. These changes will enable the inclusion of more diverse motion sequences, further enhancing the model's generalization, thereby increasing its applicability to a wide range of real-world contexts.

REFERENCES

Barquero, G., Escalera, S., and Palmero, C. (2023). BeLFusion: Latent diffusion for behavior-driven human motion prediction. In *Proc. of the IEEE/CVF Int. Conf. on Comput. Vis.*, 2317–2327.

Brand, M. and Hertzmann, A. (2000). Style machines. In *Proc. 27th Annu. Conf. on Comput. Graph. and Interactive Techn.*, 183–192.

Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., and Malik, J. (2020). Long-term human motion prediction with scene context. In *Proc. 16th Eur. Conf. on Comput. Vis. (ECCV)*, 387–404. Springer.

Corona, E., Pumarola, A., Alenya, G., and Moreno-Noguer, F. (2020). Context-aware human motion prediction. In *2020 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 6992–7001.

Deng, T. and Sun, Y. (2024). Recent advances in deterministic human motion prediction: A review. *Image Vis. Comput.*, 143, 104926.

Dreher, C.R., Wächter, M., and Asfour, T. (2019). Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robot. and Automat. Lett. (RA-L)*, 5(1), 187–194.

Guidolin, M., Vanuzzo, M., Michieletto, S., and Reggiani, M. (2024). Enhancing real-time body pose estimation in occluded environments through multimodal musculoskeletal modeling. *IEEE Robot. and Automat. Lett. (RA-L)*, 9(12), 10748–10755.

Guo, C., Liu, R., Che, C., Zhou, D., Zhang, Q., and Wei, X. (2022). Fusion learning-based recurrent neural network for human motion prediction. *Intell. Serv. Robot.*, 15(3), 245–257.

Kipf, T.N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Int. Conf. on Learn. Representations (ICLR)*.

Lee, S.U. (2024). Commonsense spatial knowledge-aware 3-D human motion and object interaction prediction. In *2024 IEEE Int. Conf. on Robot. and Automat. (ICRA)*, 3057–3063.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2021). Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(6), 3316–3333.

Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., and Tian, Q. (2020). Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction. In *2020 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 211–220.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M.J. (2015). SMPL: a skinned multi-person linear model. *ACM Trans. on Graph. (TOG)*, 34(6), 1–16.

Lou, Z., Cui, Q., Wang, H., Tang, X., and Zhou, H. (2024). Multimodal sense-informed forecasting of 3D human motions. In *2024 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 2144–2154.

Lyu, K., Chen, H., Liu, Z., Zhang, B., and Wang, R. (2022). 3D human motion prediction: A survey. *Neurocomputing*, 489, 345–365.

Ma, T., Nie, Y., Long, C., Zhang, Q., and Li, G. (2022). Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *2022 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 6427–6436.

Mandery, C., Terlemez, Ö., Do, M., Vahrenkamp, N., and Asfour, T. (2015). The KIT whole-body human motion database. In *2015 Int. Conf. on Adv. Robot. (ICAR)*, 329–336.

Martinez, J., Black, M.J., and Romero, J. (2017). On human motion prediction using recurrent neural networks. In *2017 IEEE Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 4674–4683.

Pavllo, D., Grangier, D., and Auli, M. (2018). QuaterNet: A quaternion-based recurrent model for human motion. In *Brit. Mach. Vis. Conf. (BMVC)*.

Qi, C.R., Su, H., Mo, K., and Guibas, L.J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In *2017 IEEE/CVF Conf. on Comput. Vis. and Pattern Recognit. (CVPR)*, 652–660.

Taheri, O., Ghorbani, N., Black, M.J., and Tzionas, D. (2020). GRAB: A dataset of whole-body human grasping of objects. In *Proc. 16th Eur. Conf. on Comput. Vis. (ECCV)*, 581–600.

Taylor, G.W., Hinton, G.E., and Roweis, S. (2006). Modeling human motion using binary latent variables. In *Advances in Neural Inf. Process. Syst.*, volume 19, 1345–1352.

Vanuzzo, M., Casarin, M., Guidolin, M., Michieletto, S., and Reggiani, M. (2025). Evaluating human motion prediction: A framework for accuracy and realism. *IEEE Trans. on Emerg. Topics in Comput. Intell.* Submitted.

Wang, H., Dong, J., Cheng, B., and Feng, J. (2021). PVRED: A position-velocity recurrent encoder-decoder for human motion prediction. *IEEE Trans. Image Process.*, 30, 6096–6106.

Yang, S., Li, H., Pun, C.M., Du, C., and Gao, H. (2024). Adaptive spatial-temporal graph-mixer for human motion prediction. *IEEE Signal Process. Lett.*, 31, 1244–1248.

Yu, Y., Tian, N., Hao, X., Ma, T., and Yang, C. (2022). Human motion prediction with gated recurrent unit model of multi-dimensional input. *Appl. Intell.*, 52(6), 6769–6781.