Action Descriptions as Semantic Priors for 3D Human Pose Estimation

Andrei Mihalea[1], Mihai Nan[2], and Irina Mocanu[3]

***Abstract*— Human 3D pose estimation is an important problem in computer vision, having applications in a wide variety of fields, such as healthcare, sports science, human-computer interaction, entertainment and retail. However, when solely using 2D sequences of human joints as input data, this becomes an ill-posed problem, caused by the lack of a unique solution, since there are infinitely many 3D points which can be projected to the same 2D point. Current methods rely on temporal or anatomical priors for added information which can help the lifting process. In this work, we propose implementing an additional loss objective for the problem of 3D human pose estimation as a way of embedding semantic information into the lifting process by aligning projected skeleton features with CLIP text embeddings and show that such information can help improve skeleton lifting metrics, especially when applied on out-of-distribution data, without fine-tuning. Our method adds a small MLP projection head to an already existing 3D pose estimation network and a semantic loss between projected features and embeddings of action descriptions, while, at inference, it drops both components, making it adequate for application on novel data, without any prior knowledge of actions. Our code is publicly available at `https://github.com/HRIA-HAR/SARPose`.**

I. Introduction

3D human pose estimation is an important task in computer vision and requires accurate predictions of human joints in 3D space. Human 3D skeleton prediction is essential in problems involving human action recognition and a high quality estimate is crucial for differentiating between similar actions, where finer details become important. Regarding input data, there are two main classes of methods: those relying on single or multiple consecutive RGB frames, or sequences of 2D keypoints. The latter represents the main focus of this work and comes with additional challenges compared to situations where image input is available. These challenges appear due to information loss of 2D projections, which leads to ambiguities regarding depth, scaling, and self-occlusions, which are difficult to solve without prior information or assumptions regarding temporal or anatomical structures.

Existing work tackles the inherent ambiguity issues of the 2D-to-3D lifting problem by employing spatial and temporal consistency. For example, Temporal Convolutional Networks (TCNs) [1] and transformers [2] leverage the temporal dependencies between human motions across different time frames, while methods such as Graph Convolutional Networks (GCNs) [3] model the hierarchical structures of the human body. While the aforementioned constraints tackle low-level features like bone length and temporal smoothness, our proposal introduces a higher-level global prior by aligning skeleton features with the concept of an action, narrowing the search space of possible 3D projections that explain an input sequence of 2D points.

In this work, we propose a method that incorporates semantic information into 2D-to-3D lifting models, by aligning embeddings of action descriptions with motion feature representations extracted from the input sequence. The reasoning behind introducing this type of information starts from the assumption that embedding semantic knowledge about which actions are performed during a sequence of human movements can enforce better latent representations for the skeleton features. More specifically, our semantic alignment loss aims to regularize the backbone latent space using high-level action descriptions rather than relying solely on raw coordinates. Our method's computational overhead is insignificant compared to the baseline model, as it only adds an MLP head and a new loss component during training. The MLP head is discarded during evaluation and inference.

The main contributions of this work can be summarized as follows:

- Introducing a lightweight semantic alignment method that regularizes the feature space of 2D-to-3D lifting models, using a frozen CLIP [4] text encoder.
- Showing that semantic priors can improve out-of-distribution generalization for a dataset consisting of athletic motions unseen during training.
- Providing an analysis of the model's sensitivity to different types of action description prompts and performing ablation studies with irrelevant action descriptions, showing a degradation in performance for such cases.

The rest of the paper is organized as follows: Section II reviews related work. Section III outlines the proposed method. Section IV details the evaluations and the ablation study. Finally, Section V concludes the paper.

II. Related work

Human Action Recognition (HAR) can be performed from different types of input: RGB frames, skeletal representations, inertial sensors, or multimodal combinations. Models based on convolutional neural networks, such as I3D [5], have shown that temporal information can be learned directly from video, improving performance on datasets such as UCF101 [6] or Kinetics [7].

*This research was supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

[1,2,3]The authors are with the Faculty of Automatic Control and Computers, Computer Science Department, National University of Science and Technology Politehnica Bucharest, 313 Splaiul Independentei, Bucharest, Romania. Emails: {andrei.mihalea, mihai.nan, irina.mocanu}@upb.ro

Recent methods use models that combine information about objects, spatial relationships, and background with the temporal dynamics, such as UniformerV2 [8] and Video Swin Transformer V2 [9], which introduce hierarchical mechanisms that simultaneously capture local features (fine movements) and global context (interactions and scene), achieving higher performances on Kinetics-700 [10] and Something-Something V2 [11].

Skeleton-based methods have become very popular due to their robustness to background, lighting, and appearance variations. ST-GCN [12] was one of the works that introduced spatio-temporal convolutional networks on graphs for modeling the relationships between joints and their evolution over time.

Obtaining the human skeleton can be done through human pose estimation methods, either in 2D or 3D. In the case of 2D estimation, models such as HRNet [13] or OpenPose [14] predict joint positions via heatmaps to maintain robustness against lighting variations and occlusions. For 3D estimation, one possibility is represented by MotionAGFormer [15] that uses a Transformer-based architecture with Adaptive Graph Attention. In this case, joints are modeled as nodes in a graph, and their connections are dynamically learned.

Currently, Vision-Language Models are integrated into HAR by reformulating this problem as an alignment between video representations and textual descriptions of classes. Models such as ActionCLIP [16] and X-CLIP [17] perform action recognition as a matching between video embeddings and textual descriptions of actions. Additionally, subsequent frameworks have leveraged CLIP's [4] cross-modal representations for skeleton-based action recognition [18], [19]. These works demonstrate that despite being trained exclusively on text-image pairs, CLIP's language priors can effectively enhance and guide the feature space of skeleton-based models.

III. Proposed Method

Our proposed method relies on a state-of-the-art network for 2D to 3D human skeleton lifting - MotionAGFormer [15]. This network was selected because it offers a balanced tradeoff between computational performance and prediction quality, with results matching state-of-the-art methods for 3D human pose estimation. We have adapted the MotionAGFormer by adding a semantic loss for aligning skeletal representations with embeddings obtained after applying a CLIP [4] text encoder over action descriptions. As 2D to 3D pose lifting is an underconstrained problem with an infinite number of solutions, CLIP can provide a continuous manifold that acts as a semantic anchor by forcing the skeleton features to align with its latent space during training. Since CLIP was trained on millions of text-image pairs, we regularize the backbone to project into its latent space, making it inherit the pre-trained structural prior.

In order to introduce the semantic loss into the training pipeline, first, we divided the MotionAGFormer network for 3D human pose estimation into its two components: a feature representation section and the regression head for predicting the 3D joints for a sequence of 2D inputs. Additionally, a new projection block has been added after the feature representation section, between the last MotionAGFormer block and the 3D pose regression head. Attaching the projection head here and applying a temporal averaging over the sequence length allows the projection head to capture the most abstract features of the sequence. This projection block is a lightweight multi-layer perceptron (MLP) that maps the 512-dimensional skeletal features to the CLIP text embedding space using two linear layers separated by a ReLU activation. This design enables non-linear mapping with minimal computational overhead.

Despite CLIP being trained on text-image pairs, which is a different modality than the one presented in this particular 3D human pose estimation task, where input data consists of sequences of 2D human joints, the projection block takes the place of an adapter with the purpose of translating information extracted from the skeleton sequences into a space that can be aligned to text embeddings obtained by applying the CLIP text encoder over action descriptions.

During training, for each action in the dataset, a set of text prompts is defined as short descriptions of the action performed for the input sequence of 2D human joints. Since the set of descriptions is pre-defined, the CLIP text embedding model is only applied once before the training starts and the representations of each action description are stored. At each step, a cosine similarity metric is computed between the latent representations obtained by the projection head and the embedding corresponding to the current action sequence description. The overall framework can be visualized in Fig. 1. As seen in the figure, the input sequence is passed to the MotionAGFormer block, which will be trained alongside the MLP projection and the regression heads. The only frozen part of the framework is the CLIP text encoder, which won't allow gradient flow during the training process.

The loss functions during training are well established for the 3D human pose estimation problem and include the mean per-joint error (MPJPE), normalized MPJPE and the velocity loss. In addition to these, the semantic alignment loss between projected sequence features and CLIP text embeddings is introduced and is computed as shown in Eq. 1. Here, N is the number of sequences, $\mathbf{f}$ represents the projected features from the MLP, while $\mathbf{t}$ is the CLIP text embedding for the action description.

$$\mathcal{L}_{\text{sem}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\mathbf{f}_i^{\top}\mathbf{t}_i}{\|\mathbf{f}_i\|\|\mathbf{t}_i\|}\right) \tag{1}$$

The final loss formulation is described in Eq. 2, where $\mathcal{L}_{pose}$ represents the MPJPE 3D pose loss between predictions and ground-truth joint data, $\mathcal{L}_{vel}$ is the velocity error, which enforces temporal smoothness of movements across frames, and $\mathcal{L}_{scale}$ is a normalized MPJPE loss, where predictions are multiplied by a scaling factor that makes the prediction as close as possible to the target. Moreover, λ_{vel}, λ_{scale} and λ_{sem} are the weights associated with each respective loss component. For a fair comparison

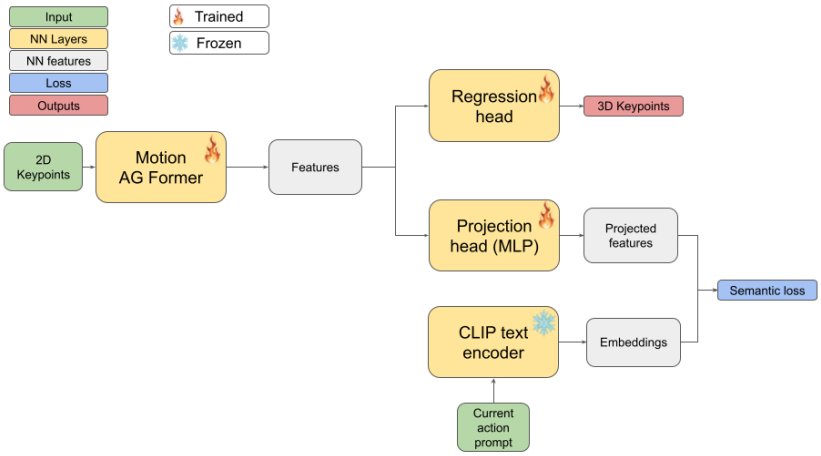


Fig. 1: Framework for training a 3D pose estimation network with semantic loss for action description alignment.

with the baseline, λ_{vel}, λ_{scale} are the same as in the original MotionAGFormer framework [15] and λ_{sem} has a value of 1 which keeps the loss scale relatively similar to the other loss components.

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + \lambda_{\text{vel}} \cdot \mathcal{L}_{\text{vel}} + \lambda_{\text{scale}} \cdot \mathcal{L}_{\text{scale}} + \lambda_{\text{sem}} \cdot \mathcal{L}_{\text{sem}} \quad (2)$$

There are three different types of text descriptions we have generated for the actions in Human3.6M, which differ in terms of how detailed the action description is and can be summarized as follows:

- **Action template**: a template like "a person {action}" is used for each action in the dataset, after its name is parsed.
- **Short descriptions**: similar in length to the template option, but involves a short description of the action instead of the action name specifically.
- **Elaborated descriptions**: represents a more detailed action description, giving more information about the characteristics of the action.

Some samples that exemplify the text descriptions can be seen in Tab. I. Except for the case of action template usage, for the short and elaborated description, multiple prompts have been generated as description for each action, using a hybrid approach between human and LLM generation. During training, we have implemented two strategies for obtaining the target embedding: first, the embeddings of all prompts corresponding to the same action description are averaged before computing the semantic alignment loss, and second, a random pre-computed embedding for the corresponding action is selected at each step.

While training requires knowledge about which action is performed during the sequence, at inference time, the network only takes a sequence of 2D human joints, making it easily adaptable to data where the action is not annotated, without adding any overhead compared to the baseline method. For comparison, the original MotionAGFormer model has 6.6G FLOPs and 4.8M parameters, while the MLP has 0.53M FLOPs and 0.53M parameters.

IV. Experiments

As in [15], the training was performed on the Human3.6M dataset [20] for all experimental configurations. Evaluation is done on the test set of Human3.6M, but also on another dataset, AthletePose3D [21], which contains sequences of actions depicting sport activities, including track and field, figure skating and running. Since we focus on out-of-distribution performance, this dataset serves as our primary evaluation benchmark for comparing existing methods with those proposed in this work.

A. Experimental setup

For evaluation, we used the two main protocols found in literature for the 3D pose estimation problem, namely the mean per-joint error (MPJPE) and MPJPE after Procrustes alignment (P-MPJPE), which performs a full geometric alignment between ground-truth and predicted skeletons, by applying the necessary translation, rotation and scaling between the two. Both metrics are computed for each action in the datasets and then the per-action mean is reported.

We trained all our models following the same procedure used for MotionAGFormer [15], optimization utilizes Adam [22] over 60 epochs with a batch size of 16 on two GTX 1080Ti GPUs.

B. Main results

In Tab. II, we compared the baseline MotionAGFormer model with the different variations of models trained with the semantic alignment loss. When adding the semantic alignment to the 3D pose prediction network, we observe a decrease in both MPJPE and P-MPJPE for the cases where action templates and short action descriptions were given to the CLIP text encoder used for the alignment loss. To strengthen the statistical significance of our results, we repeated each experiment five times with different seeds and reported the mean and standard deviation (in parentheses). Additionally, we reported experiments using both strategies for computing the target embedding: random between all

TABLE I: Prompt formulations for different action descriptions.

Action	Action template	Short description	Elaborated description
Eating	A person eating	A person bringing food to mouth	A person bringing their hand up to their mouth in a repetitive motion, leaning slightly forward while seated
Sitting down	A person sitting down	Transitioning from standing to sitting	A dynamic pose capturing the moment of lower body flexion as a person prepares to sit

embeddings of the current action description (r) and mean of the action description's embeddings (m).

While in-distribution performance drops slightly, our model shows strong out-of-distribution generalization. On the AthletePose3D dataset, which consists of sport actions, the semantic loss model significantly improves performance, reducing MPJPE by an average of 63 mm compared to the non-semantic baseline. This shows that our proposed semantic alignment loss is encouraging the backbone used for motion feature extraction to learn semantically relevant features and acts as a regularizer, preventing the model from learning specific patterns that only apply to in-distribution data.

The optimal embedding aggregation strategy shifts based on prompt length. For concise text, *short description (r)* achieves the highest out-of-distribution MPJPE (219.79 mm), outperforming the static *action template*. Quantitative analysis over CLIP's latent space shows that short descriptions present a lower intra-class diversity ($\sigma^2_{\text{intra}} = 0.1283$ vs. 0.1571 for elaborated), while maintaining a higher inter-class separability, denoted by a lower cosine similarity value between action description prototypes ($\mathcal{S}_{\text{inter}} = 0.8594$ vs. 0.8858 for elaborated). On the other hand, when averaging embeddings used for short prompts (*short description (m)*), prediction performance degrades.

This trend reverses for detailed descriptions, where *elaborated description (m)* outperforms *elaborated description (r)* . Long descriptions increase the semantic crowding and semantic noise compared to short descriptions. This embedding-space compression leads to collapsed class boundaries, which explains the degraded performance when using longer prompts under random selection. Here, averaging slightly improves the out-of-distribution performance, acting as a stabilizing filter. Therefore, textual diversity only improves performance when the aggregation method is adequate for the selected prompt length.

Another analysis we propose compares the delta of per joint, per sub-action prediction errors for two cases: first between the baseline model predictions and the ground-truth values, and second, between predictions of the best performing model with added semantic loss and the ground truth. The delta between the two errors can be seen in Fig. 2, where red denotes the model trained with semantic loss has a smaller error, while blue means baseline is better for the respective joint and sub-action. As observed, there are several joints where the improvement brought by introducing the semantic loss is significant, namely those belonging to both legs, especially for actions which involve different ice skating routines. Other keypoints such as the shoulders and upper torso also show improvements for these actions.

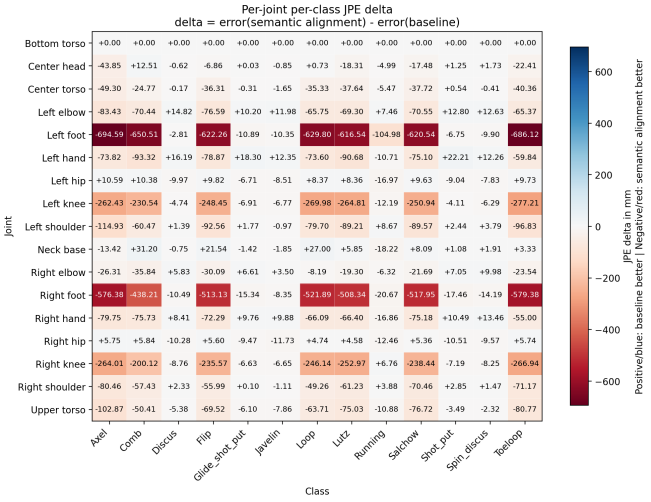


Fig. 2: Delta error comparison between baseline and a model trained using semantic loss.

Furthermore, we qualitatively analyze some model predictions for actions where the largest improvements have been achieved by using semantic loss during training. In Fig. 3, a frame from an ice skating sequence can be seen, alongside predictions made by the baseline MotionAGFormer (top), and the model trained with semantic alignment loss (bottom). Analyzing the leg joints, which showed the largest improvements according to Fig. 2, it can be observed that the angle between the upper and lower leg seems to be closer to the real value seen in the frame, for the model trained with semantic alignment loss.

C. Ablation study

For assessing whether inserting semantic information in the 3D pose estimation network brings benefits for the lifting problem, we have designed a set of ablation experiments, by varying the action descriptions given as alignment targets during training. The following setups have been used:

- **Constant prompts**: instead of computing the action description embedding, a generic text (i.e., "a human motion") is applied to all training clips.
- **Random embedding**: a randomized vector is generated for each sample during training.
- **Wrong prompts**: for each action, the text description used for alignment belongs to a random, different action from the dataset.

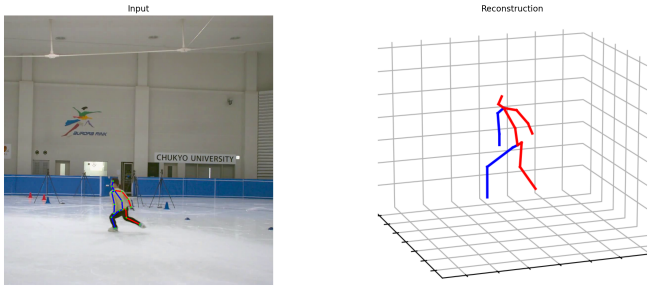
Tab. III shows the results of different ablation experiments we performed by varying the text descriptions, compared with the model trained with our proposed semantic loss. As seen in the table, all ablated models lead to severely degraded

TABLE II: Comparison of models trained with different action descriptions on Human3.6M and AthletePose3D.

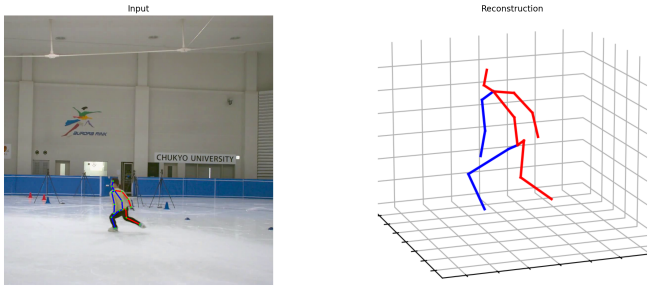
	Human3.6M		AthletePose3D	
	MPJPE ↓	P-MPJPE ↓	MPJPE ↓	P-MPJPE ↓
MotionAGFormer [15]	42.22 (0.22)	35.11 (0.14)	282.68 (18.13)	93.65 (1.98)
+ action template	43.65 (0.23)	35.98 (0.21)	220.70 (14.60)	**89.26 (2.65)**
+ short description (m)	43.86 (0.31)	35.97 (0.27)	233.70 (16.18)	95.44 (6.93)
+ short description (r)	43.48 (0.29)	36.04 (0.24)	**219.79 (5.41)**	91.34 (2.71)
+ elaborated description (m)	43.43 (0.28)	35.83 (0.40)	234.99 (15.03)	93.19 (3.32)
+ elaborated description (r)	43.15 (0.29)	35.55 (0.33)	237.87 (13.78)	94.15 (3.47)

TABLE III: Ablation study on H3.6M and AP3D datasets.

	H3.6M		AP3D	
	MPJPE↓	P-MPJPE↓	MPJPE↓	P-MPJPE↓
Action template	43.65 (0.23)	35.98 (0.21)	**220.70 (14.60)**	**89.26 (2.65)**
Random embeddings	54.90 (1.21)	44.78 (1.02)	298.23 (26.68)	93.77 (1.86)
Constant prompts	42.75 (0.36)	35.24 (0.41)	300.67 (19.45)	93.24 (2.17)
Wrong prompts	42.58 (0.29)	35.27 (0.30)	286.48 (23.08)	94.64 (4.79)



(a) Baseline MotionAGFormer.



(b) Semantic loss with template prompts.

Fig. 3: Comparison between predictions made by baseline MotionAGFormer and a model trained with semantic loss.

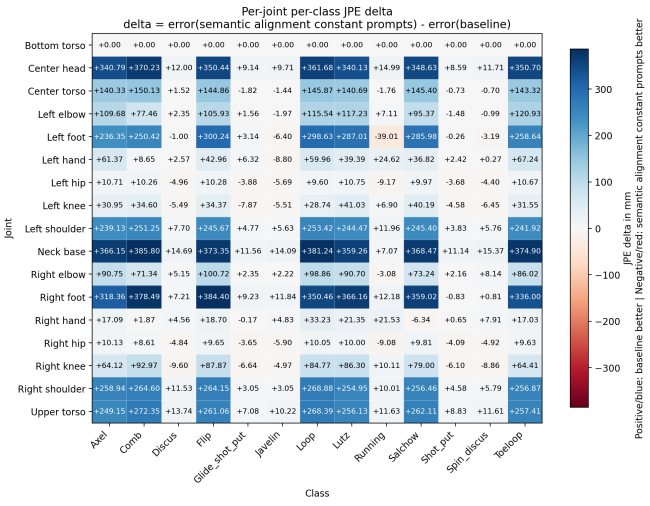


Fig. 4: Delta error comparison between baseline and a model trained using semantic loss with constant action description.

performance compared to the case where meaningful action descriptions are used for semantic alignment during training. Notably, the model's performance when using wrong prompts, despite lagging behind the model that uses relevant action descriptions, still outperforms the other two ablation configurations, which use either random embeddings or a constant prompt. This result might suggest that giving a discriminative task that has to differentiate between classes, even if they are the wrong ones, offers more information than optimizing to learn random noise or a constant embedding for each action.

Similar to what is shown in Fig. 2, we explore the influence of irrelevant semantic information over specific joints and actions. For this purpose, the model trained by leveraging semantic alignment with constant prompts has been selected, and we replicated the error delta plot for this configuration. Like previously, the delta error represents the difference in errors between the baseline model's predictions and ground truth and those of the model trained with semantic alignment. These results can be visualized in Fig. 4. The same color coding has been applied, where blue denotes better performance for the baseline and red for our model. As observed in the figure, when using a constant prompt for all the actions, most joint predictions have a higher error when compared to the model trained without semantic loss, unlike what we have previously observed in Fig. 2, where most joints and actions showed large improvements when evaluating a model that aligns relevant action descriptions to the latent representation space of the human motion. This degradation, compared even to the baseline can be explained by the dilution of relevant information flow during training, due to loss weighting, since the semantic alignment loss optimizes for irrelevant information, affecting other components of the network.

Backed by the results seen in Tab. III and Fig. 4, we can

attest that introducing semantic information about actions via the proposed alignment loss brings valuable benefits in out-of-distribution generalization, while using irrelevant information or random noise degrades performance, even when compared to the baseline.

V. Conclusions and Future Work

The problem of 2D to 3D lifting in human pose estimation is often ill-posed, since there is no unique solution, due to the fact that projecting 3D to 2D points is a many-to-one mapping, with information that is lost and cannot be recovered without additional knowledge. In this work, we have introduced semantic information as a way to embed prior knowledge about the action that is performed during a short sequence of movements. This information is given to the network without a significant overhead in terms of computational complexity and training time, since it only adds a small projection MLP over the extracted features from the input sequence, which is shown to be a strong regularizer for the latent representations of human motion sequences.

Additionally, we have analyzed different ways of describing actions and compared the performance of the 3D human pose estimation task before and after adding semantic knowledge about actions, through the semantic similarity loss. We have observed that lifting performance is enhanced for data that comes out of the original training distribution and the ablation experiments emphasized that semantic information is valuable, since replacing it with noise and irrelevant information degraded the overall performance.

Experimental results demonstrate that our semantic alignment loss significantly boosts out-of-distribution generalization, yielding a 63 mm MPJPE improvement on athletic data without fine-tuning. Current limitations include sensitivity to prompt variations and a reliance on action annotations during training.

There are a few directions we propose for further investigation in future work. First, decreasing the sensitivity to prompt changes by either adapting the text encoder and train it alongside the projection head, or leverage the power of multimodal large language models and transition from cross-modal embedding alignment to pose tokenization. A second exploratory direction involves investigating whether the semantic prior information can be beneficial to action classification tasks, especially for out-of-distribution data. Finally, other latent space regularization methods could be explored.

References

[1] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. 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.

[2] R. Liu, J. Shen, H. Wang, C. Chen, S.-c. Cheung, and V. Asari, "Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5064–5073, 2020.

[3] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. 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. 3425–3435, 2019.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[6] A. R. Z. Khurram Soomro and M. Shah in *The Wild, CRCV-TR-12-01*, 2012.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[8] K. Li, Y. Cao, Y. Zhong, L. Ji, Z. Lin, Y. Chen, C. Zhang, H. Hu, and J. Dai, "Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer," *arXiv preprint arXiv:2211.09552*, 2022.

[9] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Feng, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," *arXiv preprint arXiv:2111.09883*, 2022.

[10] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note about kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[11] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Endres, D. Batra, A. Torralba, *et al.*, "The "something something" video database for learning and evaluating visual common sense," *arXiv preprint arXiv:1706.04261*, 2017.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[13] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] S. Mehraban, V. Adeli, and B. 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. 6920–6930, 2024.

[16] M. Wang, J. Xing, Y. Liu, and Y. Wang, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[17] B. Ni, H. Peng, M. Chen, Y. Zhang, F. Meng, J. Fu, and S. Xiang, "Expanding language-image pretrained models for general video recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[18] L. Yuan, Z. He, Q. Wang, L. Xu, and X. Ma, "Skeletonclip: recognizing skeleton-based human actions with text prompts," in *2022 8th International Conference on Systems and Informatics (ICSAI)*, pp. 1–6, IEEE, 2022.

[19] A. Zhu, J. Zhu, J. Bailey, M. Gong, and Q. Ke, "Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13876–13885, 2025.

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

[21] C. Yeung, T. Suzuki, R. Tanaka, Z. Yin, and K. Fujii, "Athletepose3d: A benchmark dataset for 3d human pose estimation and kinematic validation in athletic movements," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5945–5956, 2025.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.