Physical Plausibility-aware Trajectory Prediction via Locomotion Embodiment

Supplementary Material

Abstract

This supplementary material covers details and additional results that could not be included in the main manuscript due to page limitations. First, Sec. A describes the implementation details and experimental setup. Section B presents additional results with EmLoco loss, including results on another real-world dataset, additional analyses, and more qualitative evaluation. Section C provides further evaluations of the LocoVal filter, including results with different HTP network baselines, detailed analyses, and visualizations. Lastly, Sec. D discusses failure cases and possible future directions.

A. Implementation Details

A.1. 3D Pose Conversion

We adopt the Skinned Multi-Person Linear model (SMPL) [7] as the pose format because the humanoid for the locomotion generator [15] is designed for the SMPL format. However, the HTP datasets used in our experiments (*i.e.*, JTA [3] and JRDB [12]) do not employ the SMPL format. Since the LocoVal function is trained through this humanoid in Stage 1 (explained in Sec. 4.1 in the main paper), it is necessary to align the pose formats of these HTP datasets with SMPL.

To align the pose formats of the HTP datasets with SMPL, we utilize Pose to SMPL [6] for both the JTA and JRDB datasets. Figure 7 illustrates an example of converting a 3D pose from the JTA format into the SMPL format. In Fig. 7, the poses before and after the alignment process are superimposed: the skeletons in magenta, yellow, and black are the poses before alignment, while the skeletons in red, green, and blue are the poses after alignment. As shown in Fig. 7(a), this alignment process preserves the original pose, enabling consistent format conversion without significant deformation.

However, we observed that joints are sometimes swapped after applying Pose to SMPL [6]. For instance, Fig. 7(b) shows an incorrect alignment where the left and right hips are swapped.

To address this issue, these incorrectly aligned joints are automatically swapped by finding these joints based on the inconsistent configuration of the left and right joints between the parent and child nodes. The aligned poses will be provided in our codebase; further details can be found there. The alignment process is also applied to the test data on the JTA and JRDB datasets.

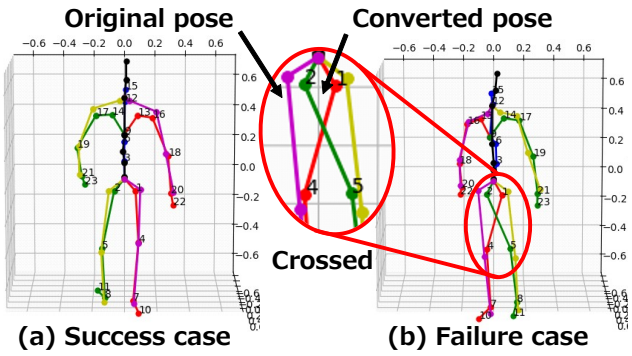


Figure 7. Examples of (a) a success case and (b) a failure case of conversion from the 3D poses in datasets [3, 12] to SMPL [7]. The 3D poses before and after the alignment process are superimposed: the skeletons in **magenta**, **yellow**, and **black** are the pose before alignment, while the skeletons in **red**, **green**, and **blue** are the pose after alignment. The numbers next to the poses indicate the joint IDs in the SMPL [7].

A.2. Experimental Setup

Locomotion Generator. The policy network of the locomotion generator is trained by Advantage Actor-Critic (A2C) [13] using Proximal Policy Optimization (PPO) [19]. The learning rate is 2×10^{-5} and Adam [4] is used as the optimizer. $1,600$ agents are trained in parallel for $5,000$ episodes. Other settings closely follow PACER [15]; *e.g.*, the reward structure of PACER [15] is used as is.

Locomotion Value (LocoVal) Function. Our LocoVal function is trained with 160 agents in parallel for $25,000$ episodes, and optimized by AdamW [9] with the learning rate of 1×10^{-3} and cosine annealing [8]. To diversify the pose-trajectory pairs, trajectories generated by PACER [15] are used in addition to trajectories from the trajectory datasets [3, 12]. Future trajectories are extracted at 2.5 fps. The humanoid's initial state $\boldsymbol{h}_0$ is sampled from AMASS [10].

To stabilize the training, we apply coordinate transformations to the inputs of the LocoVal function by constraining the input space. Specifically, we translate the humanoid's initial position to the origin and align the orientation of the humanoid and the future trajectory at the current timestep by yaw rotation. During this transformation, the relative angles between the future trajectory $\boldsymbol{\tau}_{\mathrm{f}}$, the initial pose $\boldsymbol{j}_{\mathbf{0}}$, and the initial root velocity $\boldsymbol{v}_{\mathrm{root},\mathbf{0}}$ are preserved.

HTP Network. Our HTP network and the baseline, Social-Transmotion (Social-Trans) [17], are trained with all the available annotations on the JTA [3] and JRDB [12] datasets for 30 and 100 epochs, respectively. The learning rate is 1×10^{-4}. After the training, the trained HTP networks with the best performance on the validation set are selected for evaluation on the test set.

Following Social-Trans [17], we applied random masking to train the HTP networks. Specifically, the inputs to the HTP networks are randomly masked in modality level, pose keypoint level, frame level, and location level. This modality-level masking allows the HTP networks to take the arbitrary combination of available modalities. Similarly, the frame-level masking allows the HTP networks to work with an arbitrary number of input past frames. In the experiments with momentary observations, we mask all modalities except for the most recent two frames.

In terms of our Embodied Locomotion (EmLoco) loss, the loss function shown in Eq. (2) is implemented as the Mean Squared Error (MSE) with the maximum plausibility score output from a sigmoid function *i.e.*, 1. The weight α of the EmLoco loss in Eq. (3) is set to 100 in the results of all experiments shown in the main manuscript and this supplementary material except for Table 7 in Sec. B.2.

A.3. 3D Pose Filtering

While JTA dataset [3] contains the ground truth 3D poses of locomotion in a simulated environment, in-the-wild 3D poses in JRDB dataset [12] include incorrect estimation and non-locomotion poses (*e.g.*, sitting or lying). Since the proposed method does not account for such incorrect 3D poses or poses other than locomotion, these poses are filtered out from the dataset. We applied three types of filtering as follows:

- **Rule-based Filtering**: We assume that the z-coordinate of the head should be higher than those of the knees and pelvis, and the pelvis should be higher than the ankles and lower than the shoulders when the pedestrian is walking. Based on these assumptions regarding the z-coordinate relationships of the joints, our filter removes poses that do not meet these assumptions.
- **Consistency-based Filtering**: For each pedestrian's pose sequence, this filter removes poses with z-scores greater than 2 based on the $L2$ distances from the moving average of each joint in the pose sequence.
- **Action-based Filtering**: Based on the pose-based labels (*e.g.*, walking and standing) provided in JTA-Act [2], this filter removes poses labeled with non-locomotion-related actions such as sitting or lying.

By using these filters, 29.7/17.0/20.7% of the 3D poses in the training/validation/test splits are filtered out. The visualization of Principal Component Analysis (PCA) to the 3D poses treated as vectors is shown in Fig. 8. One can see that **blue points** that are not filtered out are distributed densely. In contrast, **red points** that are filtered out are distributed more sparsely over a broader area, indicating poses that deviate significantly from typical locomotion. Examples of pose samples that are not filtered out and filtered out are shown on the left and right of Fig. 8.

Note that this 3D pose filtering affects only the results of our HTP network and Social-Trans, which use the 3D pose as input, while other traditional HTP methods do not require the 3D poses.

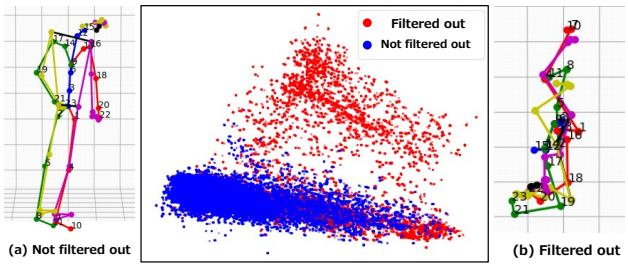


Figure 8. PCA visualization of 3D pose filtering on the JRDB dataset [12]. **the red points** represent the filtered-out data and **The blue points** represent the data that is not filtered out. (a) and (b) are examples of poses that are not filtered out and filtered out, respectively.

Table 7. Ablation studies on different EmLoco loss weights (α in Eq.(3) of the main paper). Results on the JTA dataset with 9 frames of past observations are shown.

Method	$\mathcal{L}_{\mathrm{T}}$	$\mathcal{L}_{\mathrm{E}}$	**JTA Dataset [3] 9 frames** ADE ↓	FDE ↓
Social-Trans [17] ($\alpha = 0$)	✓		1.11	2.26
Ours ($\alpha = 0.1$)	✓	✓	0.98	1.98
Ours ($\alpha = 1.0$)	✓	✓	0.96	1.95
Ours ($\alpha = 10.0$)	✓	✓	0.99	1.97
Ours ($\alpha = 25.0$)	✓	✓	0.96	1.96
Ours ($\alpha = 100.0$)	✓	✓	0.97	1.91
Ours ($\alpha = 250.0$)	✓	✓	1.03	2.07
Ours ($\alpha = 1,000.0$)	✓	✓	1.06	2.21
Ours ($\alpha = 10,000.0$)	✓	✓	1.49	2.77
w/o MSE Loss (same as $\alpha \approx \infty$)		✓	3.52	8.20

B. Additional Results on HTP with EmLoco Loss

B.1. Evaluation on the AMASS Dataset

In our framework, as a real-world motion capture dataset with accurate human poses, AMASS [10] is used as an additional training resource only in stage 1, not in the HTP network training. This raises concerns about unbalanced training resources compared to a baseline that does not use AMASS. From this perspective, to ensure fairness and provide performance on another real-world dataset, we evaluate HTP networks trained only with AMASS.

Experimental setup. Following PACER [15], we split ∼200 locomotion sequences in the AMASS dataset into training, validation, and test sets. This split is shared for training the locomotion generator, LocoVal function, and HTP network, *i.e.*, no additional data is introduced for training in the physics simulator. We use global translation and joint positions calculated via forward kinematics from the SMPL parameters in AMASS as trajectory and pose data.

Most training conditions are the same as in the experiments on JTA and JRDB described in the main manuscript. However, since many of the sequences in AMASS are short, we adjust the trajectory length accordingly. Specifically, while JTA and JRDB use 9 past frames and 12 future frames at 2.5 fps, AMASS operates at 30 fps with 12 past frames and 30 future frames. In addition, the HTP network is trained for 150 epochs, and α for the EmLoco loss weight relative to the GT loss is set to 20.

Result. Table 8 compares HTP networks [17] trained only with AMASS. These results show that also in AMASS, which captures real-world human locomotion, the proposed EmLoco loss effectively reduces prediction errors ('Ours w/o filter'). The predictions made by our method improve both the mean ADE / FDE ($4.8\%/2.0\%$) among 20 heads and minADE / minFDE ($4.8\%/3.3\%$), indicating an overall enhancement across multiple predicted trajectories.

Furthermore, applying the LocoVal filter to these predictions ('Ours w/ filter') further improves ADE / FDE while maintaining minADE / minFDE. Given that the ADE / FDE of the rejected trajectories is significantly large, this confirms that the filter successfully eliminates implausible predictions.

Additional results regarding the effect of the LocoVal filter on other datasets are discussed in Sec C.

Table 8. Stochastic HTP results with 20 heads on AMASS [10].

Method	ADE	FDE	minADE	minFDE
Social-Trans [17]	0.187	0.457	0.168	0.419
Ours w/o filter	0.178	0.448	**0.160**	**0.405**
Ours w/ filter	**0.175**	**0.438**	**0.160**	**0.405**
Rejected by filter	0.514	1.480	0.458	1.359

B.2. Detailed Analysis on EmLoco Loss Weight

We investigated the significance of the EmLoco loss by varying the hyper-parameter α, as defined in Eq.(3), which controls the balance between the MSE loss and the EmLoco loss. Note that setting $\alpha = 0$ is equivalent to excluding the EmLoco loss. Moreover, we also evaluated the performance when only the EmLoco loss was used as in Table. 4 in the main manuscript. The experimental results are summarized in Table 7.

As shown in Table 7, the ADE and FDE remain relatively stable for α values ranging from 0.1 to 100. However, when α exceeds 100, performance begins to degrade. To better understand the balance between the loss components, we examined the scale of each loss function when the training losses converged. On the JTA dataset, the scale of the MSE loss becomes approximately equal to that of the EmLoco loss when $\alpha = 1,000$.

These findings indicate that α should be adjusted such that the EmLoco loss remains smaller than the MSE loss. This is because while the EmLoco loss is low around the ground truth future trajectory, no clear peak is observed in the EmLoco loss, unlike the MSE loss. The widely distributed low values of the EmLoco loss make it difficult to train the HTP network to improve ADE and FDE. Thus, the MSE loss primarily contributes to reducing ADE and FDE, while the EmLoco loss should be utilized as a support term to enhance plausibility.

Furthermore, the robustness of the EmLoco loss to changes in the hyperparameter α is evident from the small variations in ADE and FDE observed for α values between 0.1 and 100. This robustness allows us to easily integrate the EmLoco loss into the overall loss function.

B.3. Evaluation of Displacement Errors at Each Timestep

To further demonstrate the effectiveness of the EmLoco loss, Fig. 9 compares the displacement error at each future time step with the baseline. These results correspond to the result on the JTA dataset in Table 1 of the main paper.

This bar chart highlights two key observations: First, the proposed HTP network trained with the EmLoco loss consistently outperforms the baseline across all time steps. Second, the performance gain relative to the baseline increases toward future frames, with an improvement of 8.8% for the initial frame, growing to 15.7% for the final frame. As also shown in the χ^2 distance evaluation in Table 2 of the main paper, our HTP network acquires physics-based prior knowledge through the EmLoco loss, resulting in features such as velocity that are closer to real-world data. This enhanced ability to

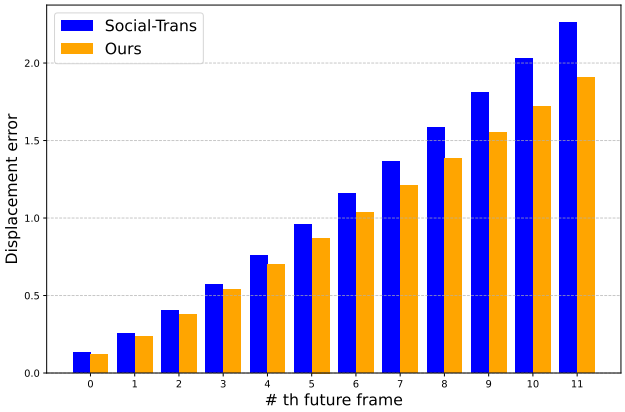


Figure 9. Evaluation of displacement errors for each future timestep.

Table 9. Stochastic HTP results with 20 heads on JTA. 'Deterimin-stic' indicates the HTP network with a single head. 'minMSE' and 'meanMSE' corresponds to the stochastic HTP network trained with the minimum MSE and the mean MSE among heads, respectively.

Method	ADE	FDE	minADE	minFDE
Social-Trans (Deterministic)	**1.11**	2.26	-	-
Social-Trans (minMSE)	2.14	4.26	0.71	**0.54**
Social-Trans (meanMSE)	1.24	**1.98**	0.93	1.97
Ours	1.80	3.56	**0.66**	**0.54**

capture such features can be the reason for the discrepancy between the baseline and our HTP network accumulating over time.

B.4. Comparison with MSE Loss Averaging in Stochastic HTP Network Training

In Sec. 1, we pointed out as follows: 'training with the MSE loss alone forces all predicted trajectories to align with a single ground truth. That is, minimizing the MSE essentially reduces the diversity of the predicted trajectories.' Here, we compare the performance of the baseline [17] trained by this MSE loss averaging with our proposed method.

As shown in Table 9, while averaging MSE among heads ('meanMSE') improves ADE / FDE, it leads to poorer minADE / minFDE. This result demonstrates that forcing the outputs of all heads closer to the single ground truth loses diversity. On the other hand, ours encourages plausibility by EmLoco loss while maintaining diversity by minMSE loss, resulting in better overall performance.

B.5. Qualitative Evaluation on the JRDB Dataset

Visualizations of the predictions by the baseline [17] and our method on the JRDB dataset [12] are shown in Fig. 10. Consistent with the results on the JTA dataset [3] (Fig. 6 of the main paper), our method predicts plausible trajectories

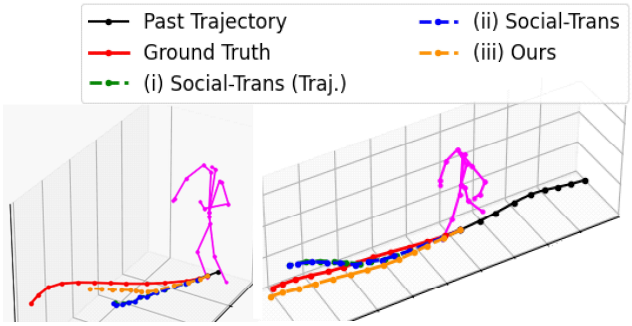


Figure 10. Visualizations of the prediction by the baseline [17] and our method on JRDB dataset [12]. Left: results with 2-frame momentary observations. Right: results with 9 frames of observations. 'Traj.' indicates that only trajectory is used as input. The scale of **the human pose** is doubled for a presentation purpose only.

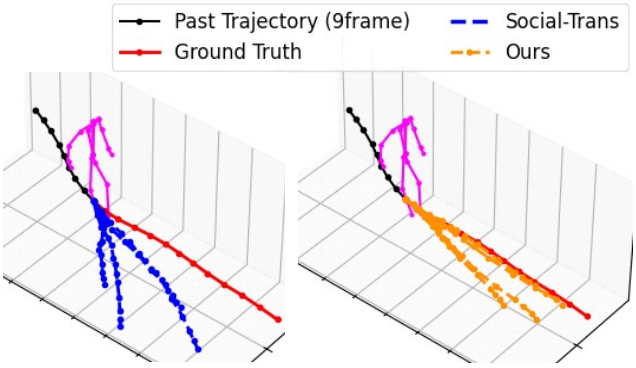


Figure 11. Comparison of stochastic HTP results by the baseline [17] and ours on the JRDB dataset [12] with 9 frames of observations. The scale of **the human pose** is doubled for a presentation purpose only.

across both observation lengths. Furthermore, the visualization of the stochastic HTP results in Fig. 11 demonstrates that, while the baseline predictions [17] deviate significantly from the ground truth for all trajectories, our HTP network successfully predicts plausible trajectories while maintaining reasonable diversity. Even with the imperfect 3D poses in the JRDB dataset, the proposed method makes a prediction close to the ground truth trajectory. These results support the effectiveness of the proposed method in real-world scenarios, not only in simulated environments such as JTA [3].

C. Additional Results on LocoVal Filter

C.1. Evaluation with Another HTP Network Incorporating Human Poses

While we employed the Social-Trans [17] as the baseline, our method can be applied to other HTP networks. To this end, as an additional baseline incorporating human poses, we evaluate HST. Table 10 evaluated on the full JRDB dataset

Table 10. LocoVal filter on HST [18] with 6 heads on the JRDB dataset [12]. The filtering threshold λ is set to 0.5.

Method	ADE	FDE	minADE	minFDE
Pretrained HST [18]	0.57	0.98	**0.28**	**0.45**
Pretrained HST (w/ filter)	**0.46**	**0.80**	**0.28**	0.46
Rejected by filter	0.95	1.64	0.60	1.01

Table 11. LocoVal filter on NSP [26] with 20 heads on SDD [16]. The filtering threshold λ is set to 0.55.

Method	ADE	FDE	minADE	minFDE
Pretrained NSP [26]	24.17	49.32	**6.52**	**10.59**
Pretrained NSP (w/ filter)	**24.13**	**49.24**	**6.52**	**10.59**
Rejected by filter	256.24	548.09	210.39	464.17

(in the main manuscript, a subset provided by the Social-Trans [17] was used) demonstrates that our LocoVal filter can also improve the ADE / FDE of HST while preserving the minADE / minFDE.

C.2. Evaluation with HTP Network Based on Neural Social Physics

There is a type of physics-aware HTP method, called Neural Social Physics (NSP) [26]. While our method evaluates the locomotion of individual persons, NSP [26] models interactions between people through the concept of 'social force'. Since these contributions are independent, the proposed method can incorporate NSP and enhance its performance. Table 11 presents the effect of the LocoVal filter with pretrained NSP on the Stanford Drone Dataset (SDD) [16]. Since the trajectory in the SDD dataset is in pixels, we converted the scale to meters for input to the LocoVal function [1] [1]. The significantly large ADE / FDE of the rejected samples suggests that even physics-aware existing HTP networks can produce implausible predictions that considerably deviate from the ground truth.

C.3. Detailed Analysis on Filtering Threshold

The LocoVal filter introduced in Sec. 4.3 of the main paper allows control over the strictness of the trajectory filtering by changing the threshold λ. While Tables 5 and 6 in the main paper present results for a single threshold, this section investigates how the filtering results vary with different threshold settings. In addition to the results on the JTA [3] and ETH / UCY datasets [5, 14] presented in the main manuscript, this section also provides the filtering results on the JRDB dataset [12]. Since the JRDB dataset lacks 3D poses for some pedestrians, we evaluate the performance both before and after filtering for pedestrians with 3D poses.

[1] https://github.com/crowdbotp/OpenTraj/tree/master/datasets/SDD

Table 12. Results of the LocoVal filter with various λ for stochastic HTP on the JTA [3] dataset with 9 and 2-frame observations. The results are presented as left/right values, where the left denotes evaluations with 5 heads and the right denotes 20 heads.

Method		λ	9 frames		2 frames	
			ADE	FDE	ADE	FDE
Social-Trans (w/o $\mathcal{L}_E$)	w/o Filtering	-	1.86/2.14	3.51/4.26	2.23/2.46	4.55/5.07
	w/ Filtering	0.65	1.83/2.02	3.48/4.02	2.10/2.28	4.34/4.70
	w/ Filtering	0.70	1.81/1.97	3.49/3.93	2.06/2.21	4.27/4.56
	w/ Filtering	0.75	1.85/1.94	3.69/3.97	2.11/2.17	4.46/4.56
	Rejected	0.65	2.71/3.99	4.24/7.88	4.92/5.09	8.91/10.52
	Rejected	0.70	2.37/3.41	3.75/6.65	3.83/4.32	7.16/8.87
	Rejected	0.75	1.92/2.50	3.19/4.78	2.50/3.03	4.76/6.06
Ours (w/ $\mathcal{L}_E$)	w/o Filtering	-	1.68/1.80	3.34/3.56	1.94/2.12	3.96/4.47
	w/ Filtering	0.65	1.66/1.78	3.32/3.54	1.92/2.10	3.93/4.44
	w/ Filtering	0.70	1.65/1.76	3.31/3.52	1.90/2.08	3.88/4.41
	w/ Filtering	0.75	1.64/1.75	3.42/3.62	1.90/2.08	4.00/4.50
	Rejected	0.65	2.68/2.81	4.41/4.89	3.80/4.15	7.11/8.02
	Rejected	0.70	2.32/2.40	3.86/4.19	3.22/3.25	6.18/6.41
	Rejected	0.75	1.83/1.93	3.11/3.44	2.09/2.24	3.83/4.36

Tables 12, 13, and 14 presents the results on the JTA, JRDB, and ETH / UCY datasets.

Consistent with the experimental results in the main text, the LocoVal filter improved the average ADE / FDE across a wide range of settings, including different datasets, varying numbers of prediction heads, different numbers of input frames, and HTP networks trained with and without the EmLoco loss. Interestingly, however, a certain trade-off can be observed in these results. When the threshold λ is relaxed (See $\lambda = 0.65, 0.70, 0.75$ in the JTA, JRDB, and ETH / UCY datasets, respectively), it does not significantly impact the filtering performance. For example, Table 12 shows that the ADE with 9 frames of observation degrades from 1.81/1.97 (λ = 0.70) to 1.83/2.02 (λ = 0.65) with Social-Trans, and from 1.65/1.76 (λ = 0.70) to 1.66/1.78 (λ = 0.65) with our method. This performance degradation is quite small. However, it becomes capable of rejecting extremely incorrect results (*e.g.*, ADE and FDE of rejected samples on the ETH / UCY dataset are 14.48 and 32.87). Setting the high threshold (See $\lambda = 0.70, 0.75, 0.80$ in the JTA, JRDB, and ETH / UCY datasets, respectively) sometimes degrades the performance (*e.g.*, after the filtering with $\lambda = 0.80$ on the JRDB dataset with momentary observations, ADE / FDE of our HTP network become worse than those of without filtering). This is because high plausibility is not always equivalent to being close to the ground truth trajectory. While feasible locomotion has a certain level of plausibility, humans may perform implausible locomotion due to interactions with obstacles or others. While the threshold λ can be freely controlled, it is necessary to set it appropriately, considering this trade-off.

Table 13. Results of the LocoVal filter with various λ for stochastic HTP on the JRDB [12] dataset with 9 and 2-frame observations. The results are presented as left/right values, where the left denotes evaluations with 5 heads and the right denotes 20 heads.

Method		λ	9 frames		2 frames	
			ADE	FDE	ADE	FDE
Social-Trans (w/o $\mathcal{L}_E$)	w/o Filtering	-	0.71/0.76	1.40/1.58	0.63/0.68	1.28/1.46
	w/ Filtering	0.70	0.70/0.73	1.39/1.52	0.59/0.67	1.21/1.43
	w/ Filtering	0.75	0.69/0.71	1.38/1.48	0.58/0.66	1.18/1.42
	w/ Filtering	0.80	0.68/0.64	1.36/1.34	0.56/0.60	1.13/1.30
	Rejected	0.70	1.60/2.31	3.03/4.73	1.86/1.60	4.10/3.41
	Rejected	0.75	1.25/1.62	2.37/3.30	1.44/1.13	3.12/2.41
	Rejected	0.80	0.75/0.89	1.48/1.83	0.76/0.77	1.55/1.64
Ours (w/ $\mathcal{L}_E$)	w/o Filtering	-	0.61/0.71	1.26/1.50	0.57/0.67	1.17/1.46
	w/ Filtering	0.70	0.60/0.70	1.23/1.49	0.56/0.65	1.15/1.43
	w/ Filtering	0.75	0.59/0.69	1.21/1.46	0.57/0.64	1.16/1.40
	w/ Filtering	0.80	0.59/0.64	1.22/1.38	0.59/0.61	1.20/1.33
	Rejected	0.70	1.76/1.60	3.42/3.20	0.97/1.51	1.97/3.25
	Rejected	0.75	1.15/1.19	2.32/2.45	0.74/1.14	1.52/2.47
	Rejected	0.80	0.66/0.79	1.36/1.64	0.57/0.73	1.16/1.59

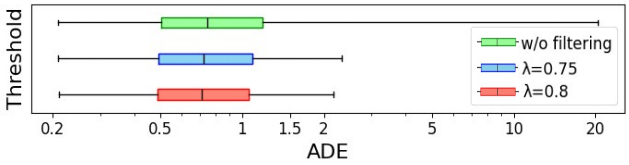


Figure 12. Box plot of ADE before/after LocoVal filtering on EqMotion [24] with 20 heads on ETH / UCY [5, 14]. λ is the threshold.

C.4. Trade-off between Retaining Plausible Samples and Suppressing Others

While our LocoVal filter can suppress implausible trajectories, trajectories close to the ground truth still can be judged as implausible. This is the potential trade-off of the LocoVal filter between retaining plausible samples and suppressing others. However, according to the results shown in the box plot in Fig. 12, our LocoVal filter preserves the minimum and median well while effectively rejecting trajectories with high ADE, achieving its purpose.

C.5. Detailed Analysis on Evaluated Plausibility Scores

The performance improvement by the LocoVal filter depends on the quality of the plausibility score evaluation performed by the LocoVal function, which is trained by our proposed method. To investigate this, we provide bar charts regarding the plausibility scores evaluated by our LocoVal function and the corresponding HTP performance in Figs. 13 and 14. Figure 13 presents the results with our HTP network and Fig. 14 represents the results of the baseline [17]. While high plausibility is not always equivalent to closeness to the ground truth trajectory as mentioned above, both figures demonstrate a consistent trend: trajectories with lower plausibility scores exhibit higher ADE values, while those with higher scores tend to have lower ADE values. This observation indicates that trajectories with higher plausibility scores are more likely to be closer to the ground truth trajectories. These results support the validity of our proposed LocoVal function, which has been learned to effectively evaluate plausibility as embodied locomotion.

Furthermore, comparisons of the plausibility scores between the trajectories predicted by our HTP network and those predicted by the baseline are shown in Fig. 13 and Fig. 14, respectively. These results reveal that the baseline, Social-Trans [17], generates a higher proportion of trajectories with lower plausibility scores. Comparing the number of trajectories with a plausibility score of 0.7 or lower, the baseline has $2,319$ trajectories, whereas the proposed method reduces this to $1,232$, achieving a 46.9% decrease. This indicates that the Social-Trans, which is trained solely on the MSE with respect to the ground truth, is unable to generate plausible trajectories. In contrast, by incorporating the EmLoco loss into the training objective, our model is capable of predicting more plausible trajectories.

C.6. Visualization of Plausibility Scores

We provide qualitative examples of plausibility score evaluation by our LocoVal function in Fig. 15. Here, to evaluate the validity of the plausibility score evaluation by the LocoVal function, we utilize predictions from Social-Trans [17], which predicts diverse trajectories ranging from plausible to implausible. In the left case, the two trajectories indicated in cyan are implausible due to their excessive movement from the current pose. As expected, these trajectories have lower plausibility scores and are farther from the ground truth trajectory. In the right case, predicted trajectories that deviate from the ground truth and involve sharper turns tend to have lower plausibility scores. Our LocoVal filter enables more plausible and accurate HTP by excluding such implausible trajectories at inference.

C.7. Visualization of LocoVal Filtering on ETH / UCY Datasets

Furthermore, we visualized the effect of the LocoVal filter with $\lambda = 0.8$ on the predictions of the pre-trained EqMotion [24] on the ETH / UCY dataset [5, 14], as shown in Fig. 16. This demonstrates that, despite performing zero-shot filtering using the LocoVal function that does not rely on pose information, the LocoVal filter effectively identifies and eliminates implausible trajectories (*e.g.*, too fast, involving sharp turns, or lack of smoothness). Again, the filtered results maintain trajectory diversity while constraining predictions to those plausible and closer to the ground truth, confirming the effectiveness of the trained LocoVal function at inference.

D. Limitations and Future Work

While the proposed method improves ADE and FDE for many samples, there are still some failure cases. One mode

Table 14. Results of zero-shot filtering with various λ by the LocoVal filter on the predictions of a pre-trained 20 heads trajectory predictor [24] on the ETH / UCY dataset [5, 14]. 'Mean' represents the average performance across the 5 subsets.

	Method	λ	ETH		HOTEL		UNIV		ZARA1		ZARA2		Mean	
			ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
Pretrained EqMotion [24]	w/o Filtering	-	2.18	4.63	0.64	1.31	1.30	2.81	0.82	1.84	0.65	1.47	1.12	2.41
	w/ Filtering	0.75	1.58	3.27	0.63	1.30	1.05	2.26	0.82	1.84	0.65	1.47	0.95	2.03
	w/ Filtering	0.80	1.41	2.88	0.61	1.26	0.93	2.04	0.80	1.80	0.64	1.45	0.88	1.89
	w/ Filtering	0.85	2.11	4.62	0.81	1.76	1.30	2.84	0.77	1.70	0.86	1.95	1.17	2.57
	Rejected	0.75	14.48	32.87	5.46	11.12	7.68	16.57	4.50	8.81	4.28	8.77	7.28	15.63
	Rejected	0.80	8.89	19.72	2.69	5.53	4.33	9.18	1.70	3.67	2.21	4.72	3.96	8.56
	Rejected	0.85	2.18	4.63	0.64	1.31	1.30	2.81	0.82	1.84	0.65	1.47	1.12	2.41

of such a case is shown in Fig. 17. Since the JTA dataset [3] is synthetic, some of the ground truth trajectories are implausible. Consequently, as illustrated in Fig. 17, there are cases where trajectories with lower plausibility scores are closer to the ground truth among multiple predicted trajectories while our LocoVal function reasonably scores each trajectory. As mentioned earlier, people would move along such implausible trajectories in real-world scenarios due to interactions with others. Therefore, future research may integrate techniques to score trajectories considering surrounding contexts [21].

Furthermore, as mentioned in Sec. 6 of the main manuscript, the JRDB dataset [12] contains incorrect 3D poses. Although our experiments demonstrate that our proposed method can improve HTP performance even with such imperfect poses, as shown in Fig. 18, both the Social-Trans [17] and our proposed method make significant errors when faced with incorrect poses. As discussed in the main paper, it is expected that more accurate 3D pose estimation or methods that can account for pose uncertainty [20, 22] can further enhance the effectiveness of our proposed method in real-world scenarios.

Lastly, while training with SMPL humanoid [7] and Isaac Gym [11] enables realistic human motion generation [25] and accurate reconstruction of real-world human locomotion [23], the more accurate physics simulators are anticipated to more realistically simulate the human locomotion in the real world.

References

[1] Javad Amirian, Bingqing Zhang, Francisco Valente Castro, Juan Jose Baldelomar, Jean-Bernard Hayet, and Julien Pettre. Opentraj: Assessing prediction complexity in human trajectories datasets. In *ACCV*, 2020. 5

[2] Mahsa Ehsanpour, Fatemeh Sadat Saleh, Silvio Savarese, Ian D. Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *CVPR*, 2022. 2

[3] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *ECCV*, 2018. 1, 2, 4, 5, 7

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[5] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, 2007. 5, 6, 7, 9

[6] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. *arXiv:2304.00553*, 2023. 1

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ToG*, 34(6), 2015. 1, 7

[8] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[10] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *ICCV*, 2019. 1, 3

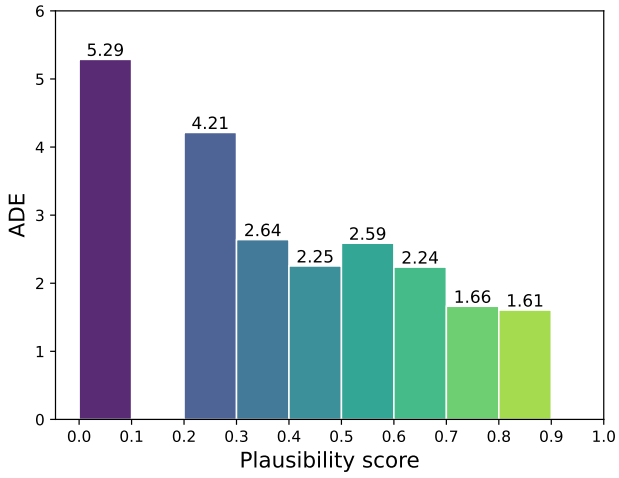
[11] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance GPU based physics simulation for robot learning. In *NeurIPS*, 2021. 7

[12] Roberto Martín-Martín, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *TPAMI*, 45(6), 2023. 1, 2, 4, 5, 6, 7

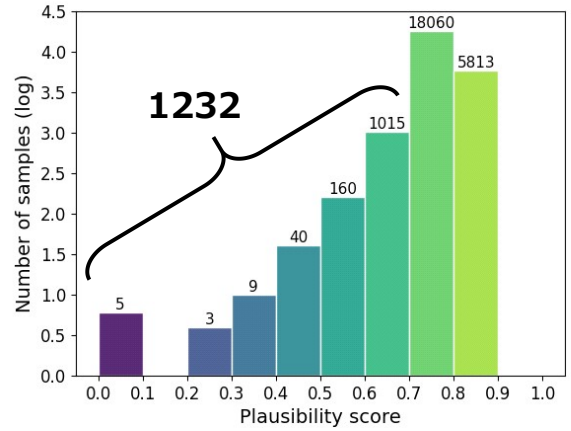
[13] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016. 1

[14] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *ECCV*, 2010. 5, 6, 7, 9

[15] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *CVPR*, 2023. 1, 3

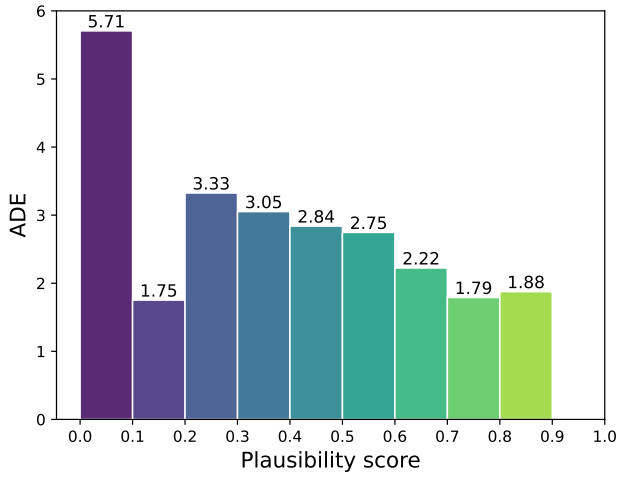


(a) ADE of our HTP network vs. plausibility scores.

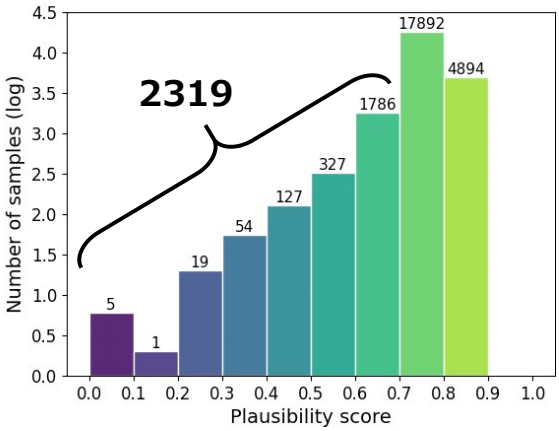


(b) Number of samples vs. plausibility scores.

Figure 13. Bar charts illustrating the plausibility scores evaluated by the LocoVal function and the corresponding bin-wise ADE and the number of samples. The numbers displayed above each bin indicate the ADE and the count of predicted trajectories with the corresponding plausibility score. The results show the predictions made by 5-head our HTP network with 9 frames observations on the JTA dataset.



(a) ADE of the basline [17] vs. plausibility scores.



(b) Number of samples vs. plausibility scores.

Figure 14. Bar charts illustrating the plausibility scores evaluated by the LocoVal function and the corresponding bin-wise ADE and the number of samples. The numbers displayed above each bin indicate the ADE and the count of predicted trajectories with the corresponding plausibility score. The results show the predictions made by 5-head Social-Trans [17] with 9 frames observations on the JTA dataset.

[16] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016. 5

[17] Saeed Saadatnejad, Yang Gao, Kaouther Messaoud, and Alexandre Alahi. Social-transmotion: Promptable human trajectory prediction. In *ICLR*, 2024. 2, 3, 4, 5, 6, 7, 8

[18] Tim Salzmann, Hao-Tien Lewis Chiang, Markus Ryll, Dorsa Sadigh, Carolina Parada, and Alex Bewley. Robots that can see: Leveraging human pose for trajectory prediction. *RA-L*, 8(11), 2023. 5

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017. 1

[20] Megh Shukla, Roshan Roy, Pankaj Singh, Shuaib Ahmed, and Alexandre Alahi. Vl4pose: Active learning through out-of-distribution detection for pose estimation. In *BMVC*, 2022. 7

[21] Jianhua Sun, Yuxuan Li, Liang Chai, and Cewu Lu. Stimulus verification is a universal and effective sampler in multi-modal human trajectory prediction. In *CVPR*, 2023. 7

[22] Hiromu Taketsugu and Norimichi Ukita. Active transfer learning for efficient video-specific human pose estimation. In *WACV*, 2024. 7

[23] Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidler, and Sameh Khamis. Learning human dynamics in autonomous driving scenarios. In *ICCV*,

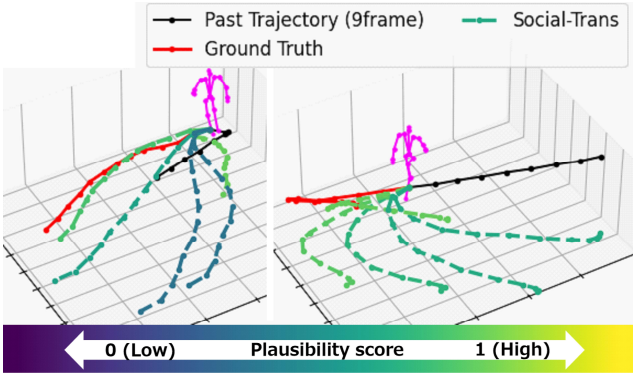


Figure 15. Plausibility score evaluation by our LocoVal function on the JTA dataset with 9 frames of observations. The scale of **the human pose** is doubled for a presentation purpose only.

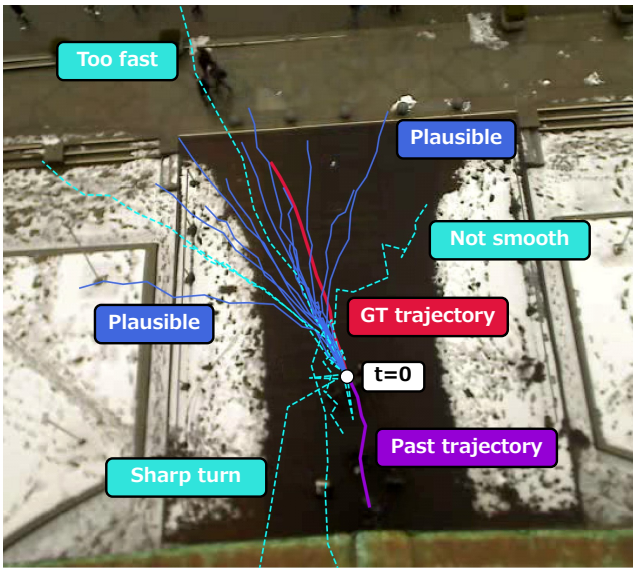


Figure 16. HTP result of pre-trained EqMotion [24] with our LocoVal filter on ETH / UCY dataset [5, 14]. **Blue lines** represent filtered trajectories and **light blue dashed line** represent rejected trajectories.

2023. 7

[24] Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *CVPR*, 2023. 6, 7, 9

[25] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 7

[26] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *ECCV*, 2022. 5

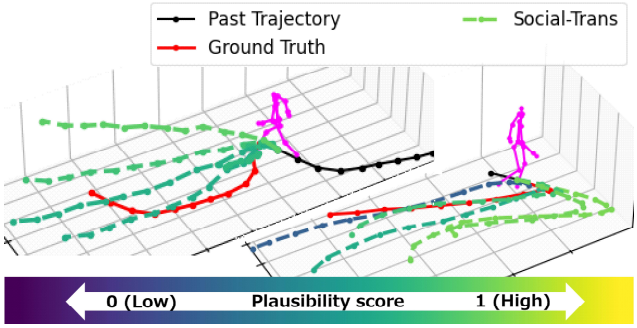


Figure 17. Failure cases and estimated plausibility scores on the JTA dataset. The scale of **the human pose** is doubled for a presentation purpose only.

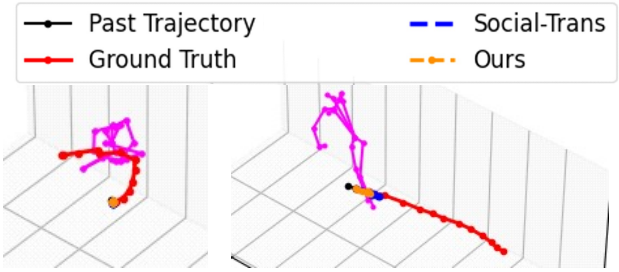


Figure 18. Failure cases in the JRDB dataset. The scale of **the human pose** is doubled for a presentation purpose only.