RD-Diff: RLTransformer-based Diffusion Model with Diversity-Inducing Modulator for Human Motion Prediction

Haosong Zhang[1,2], Mei Chee Leong[1], Liyuan Li[1], and Weisi Lin[2]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2] Nanyang Technological University, Singapore
haosong001@e.ntu.edu.sg

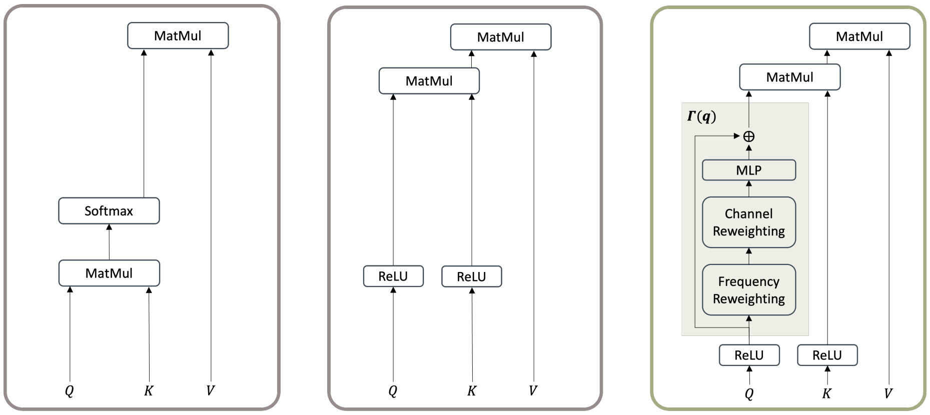
1 Architecture Details

1.1 Network Architecture

Table 1: Implementation parameters of RD-Diff. The values of J are 17 and 15 for Human3.6M and HumanEva-I, respectively. The values of l_1 are 8 and 4 for Human3.6M and HumanEva-I, respectively. The values of O are 25 and 15 for Human3.6M and HumanEva-I, respectively. Besides, $l_2 = 4$, $N = 20$, $K = 50$, $C = 3$, $D = 256$.

Component	Block	Layer	Input Size	Output Size
$\mathcal{M}_\psi$	$\mathcal{S}_\psi$	Embedding Layer	$N \times J \times C$	$N \times D$
		$l_2\times$ RLTransformer	$N \times D$	$N \times D$
		Gumbel-Softmax	$N \times D$	D
		MLP_1	D	D
		MLP_2	D	D
	Modulation	Reparameterization	D	D
		Repeat	D	$N \times D$
		Eq. (5) in main paper	$N \times D$	$N \times D$
$\mathcal{D}_\theta$	DCT	-	$(O + F) \times J \times C$, $N \times D$	$N \times J \times C$, $N \times D$
	$\boldsymbol{\epsilon}_\theta$	Embedding Layer	$N \times J \times C$, $N \times D$	$N \times D$, $N \times D$
		$l_1\times$ RLTransformer with FiLM	$N \times D$, $N \times D$	$N \times D$
		Projection	$N \times D$	$(O + F) \times J \times C$
	iDCT	-	$(O + F) \times J \times C$	$(O + F) \times J \times C$

As depicted by Fig. 1 in main paper and Tab. 1, our RD-Diff model operates on an input $\mathbf{x} \in \mathbb{R}^{O\times J\times C}$. Here, O represents the number of observed poses, J is the number of joints in a pose, and C is the dimension size of each joint. To enhance the quality of generated pose sequences, we replicate the last pose of $\mathbf{x}$ for F times, appending them to $\mathbf{x} \in \mathbb{R}^{(O+F)\times J\times C}$ [2]. The values of F are set to 100 and 60 for the Human3.6M and HumanEva-I datasets, respectively. Following [2], a Discrete Cosine Transform (DCT) operator is applied to induce temporal smoothness by transforming the temporal information along the $O+F$ dimension into frequency space. We retain only low-frequency coefficients and discard high-frequency ones, resulting in data of size $[N \times J \times C]$, where $N = 20$



(a) Softmax Transformer **(b)** Linear Transformer **(c)** RLTransformer

Fig. 1: Attention architectures of linear transformer, softmax transformer and RL-Transformer.

represents the number of remaining coefficients. Continuing with the DIM $\mathcal{M}_\psi$, composed of observation encoder $\mathcal{S}_\psi$ and modulation operations, we learn a noise-modulated condition $\hat{\mathbf{z}}_\mathbf{k}(\mathbf{t})$. In $\mathcal{S}_\psi$, the Embedding Layer initially maps hidden features into $\mathbf{D}$, followed by l_2 layers of RLTransformer extracting a latent subspace $\mathbb{V}$ with dimensions $N \times D$. Subsequently, a random weight vector $\mathbf{w} \in \mathbb{R}^N$ is sampled using the Gumbel-Softmax technique. Multiplying $\mathbf{w}$ by $\mathbf{V}$ yields a point in $\mathbb{V}$ of shape $[D]$. Two parallel MLP layers, MLP_1 and MLP_2, then project the sampled point into mean b and variance A for a Gaussian distribution, respectively. In modulation operations, a latent noise variable z is drawn from Gaussian distributions using the reparameterization trick and repeated along the frequency dimension N times. The noise-modulated condition $\hat{\mathbf{z}}_\mathbf{k}(\mathbf{t})$ is obtained by modulating x with decreasing noise z through Eq. (5) in the main paper. In the RLTransformer-based diffusion model, $\hat{\mathbf{z}}_\mathbf{k}(\mathbf{t})$ and observation x are fed into the pretrained $\boldsymbol{\epsilon}_\theta$ to produce future motions in the frequency space of shape $[(O+F) \times J \times C]$, where FiLM [3] serves as the conditioning method. Subsequently, frequency features are projected back into the pose space using the iDCT function, resulting in pose sequences of shape $[(O+F) \times J \times C]$. Finally, a slice operator extracts only the future F frames, yielding shape $[F \times J \times C]$ results, representing the future pose sequence predicted by our network. In addition, the reweighting technique in "FreeU" boosts performance. For example, it enhances APD by 2.8% and ADE by 1.2% for Human3.6M dataset.

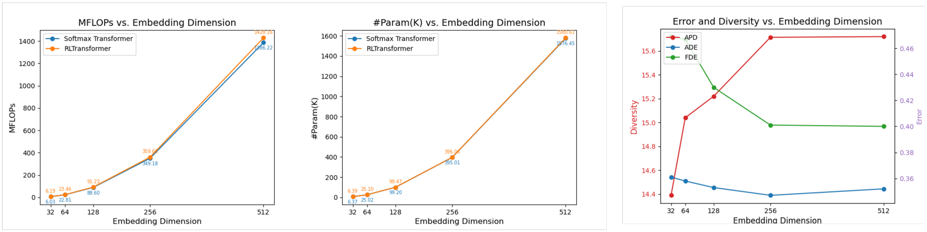
1.2 Transformer Architecture Comparison

For comparison, the detailed attention architectures in linear transformer, softmax transformer, and RLTransformer are shown in Fig. 1, which correspond to attention weights heatmaps in main paper Fig. 4. In softmax transformer [4], the similarity is computed as $\mathrm{Sim}\,(Q,K)=\exp(QK^T/\sqrt{d})$, where weights are normalized based on exponential terms for each input sequence, results in high computation complexity. In contrast, linear attention employs similarity measurement as $\mathrm{Sim}\,(Q,K)\,=\,\phi(Q)\phi(K)^T$, whose attention distribution tends to be excessively smooth, causing its output to approach the average of all features and lacking emphasis on more informative regions. This is attributed to the inadequacy of simple approximations, such as utilizing ReLU activation [1], which results in a notable decline in performance. Designed on top of a linear transformer, our RLTransformer manages to decouple the softmax operation with proper approximation. It's worth noting that we do not initialise the frequency re-weighting vector ($F \in \boldsymbol{R}^N$) directly, but via $F = \mathrm{norm}(\hat{q} \times V)$, which makes our model fit different numbers of frequency tokens. Compared to the softmax transformer, superior performance is achieved while maintaining comparable computational costs (FLOPs and number of parameters, see main paper Sec. 4.3). In addition, RLTransformer (L), where regulation function Γ is applied to both Q and K, has 0.094 GFLOPs and 0.105m parameters.

2 Additional Ablation Studies

Unless otherwise stated, all ablation studies are conducted on RD-Diff (B) and Human3.6M dataset.

2.1 Embedding Dimension



(a) MFLOPs

(b) #Param(K)

(c) Error and Diversity

Fig. 2: Ablation on different embedding dimensions.

Apart from the computational cost mentioned in the main paper Sec. 4.3, which adopts 128 as the embedding dimension D, we conduct more experiments using different embedding dimensions on one transformer layer. As shown in Fig. 2 (a) and (b), plots of the Softmax transformer and our RLTransformer are almost overlapped, showing our RLTransformer can achieve better performance than the softmax transformer at comparable computational cost. As shown in Fig. 2 (c), the best APD and FDE are obtained when $D = 512$, and the best ADE is obtained when $D = 256$. When $D = 256$, APD and FDE are also comparable with those when $D = 512$, we finally choose 256 as the default value of D, the most cost-effective choice.

2.2 One-stage v.s. Two-stage Training

Table 2: One-stage v.s. Two-stage Training.

Pipeline	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
One-stage	15.178	0.361	0.410	0.447	0.447
Two-stage	15.714	0.347	0.401	0.445	0.444

We experimented with a one-stage training pipeline, where we jointly train RLTransformer-based diffusion model $\mathcal{D}_\theta$ and Diversity-Inducing Modulator (DIM) $\mathcal{M}_\psi$. Results on Human3.6M are shown in Tab. 2. Inferior performance on the one-stage training pipeline shows that training the $\mathcal{D}_\theta$ with a noisy condition diminishes the performance as the network learns to associate constructing latent subspace with diffusion backbone learning, which entangles noisy conditions with specific motion sequences. Furthermore, the decomposition of establishing a motion prior (i.e., $\mathcal{D}_\theta$) and diversity-induced sampling process (i.e., $\mathcal{M}_\psi$) enhances control and flexibility.

2.3 RLTransformer

The regulated linear attention in RLTransformer is to balance the trade-offs between focus and generalization. Applying reweighting functions ensures that the model can effectively capture both local and global frequency features.

Table 3: Linear attention variations.

Method	APD↑	ADE↓	FDE↓
Linear Transformer [52]	14.951	0.362	0.495
Enhanced Linear Attn [14]	15.328	0.356	0.473
FLatten Transformer [37]	15.379	0.354	0.441
RLTransformer	15.714	0.347	0.401

In Tab. 3, RLTransformer outperforms other linear attention variations, showing the advantage over the context aggregation in main paper [14] via fixed-scale convolutions and the fixed mapping function in main paper [37] via adjusting element-wise power in capturing the dynamic nature of human motion.

We clarify that a balanced sharpness achieves the best model performance, as shown in Fig. 3, where intermediate entropy, rather than higher entropy, yields optimal results.

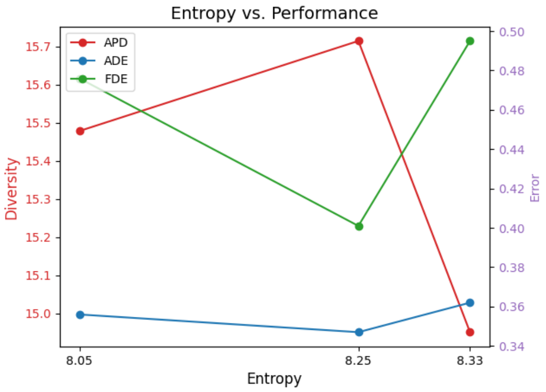


Fig. 3: Entropy v.s. performance. The entropy of softmax transformer, RLTransformer and linear transformer is 8.05, 8.25 and 8.33, respectively.

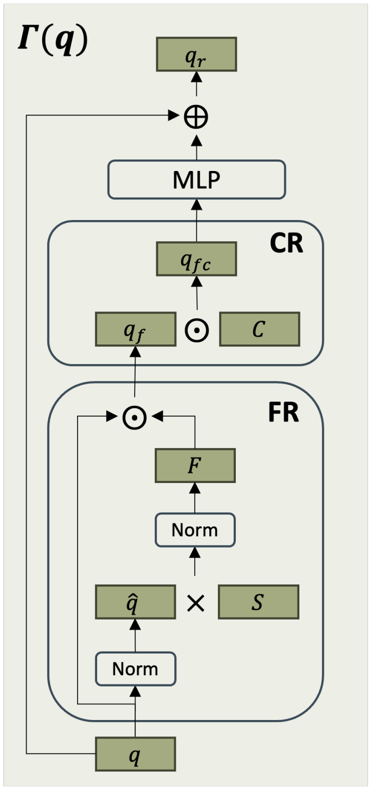


Fig. 4: Regulation Function Γ.

We design regulation function $\varGamma$ on top of linear transformer, forming RL-Transformer. Detailed structure of $\varGamma$ is shown in Fig. 4.

2.4 DCT Transform

Table 4: Ablation studies on DCT transform.

Method	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
w/o DCT	17.833	0.417	0.432	0.458	0.456
w/ DCT	15.714	0.347	0.401	0.445	0.444

We conduct comparison studies to assess the impact of DCT versus not using it, where "w/o DCT" denotes removing DCT from RD-Diff, and "w/ DCT" denotes RD-Diff. As shown in Tab. 4, DCT can reduce prediction error by 5% to 16%.

2.5 Diversity Loss Only v.s. DIM

We investigate the advantages of our DIM over the diversity loss directly included in the main predictor (i.e., Baseline model in main paper Tab. 2). As shown in Tab. 5, diversity loss causes a trade-off between diversity and accuracy, whereas DIM boosts both diversity and accuracy.

3 More Visualisation

Additional examples generated from the Human3.6M and HumanEva-I datasets are shown in Fig. 5 and Fig. 6 for visual evaluation. The "GT" row indicates the ground truth motion sequence, and the following "Generated Samples" rows indicate the generated motion sequence by our RD-Diff. The "Observation" frames are identical for the "GT" and "Generated Samples" rows. Various "Prediction" sequences show high fidelity and diversity of RD-Diff's prediction.

4 Limitation and Future Work

Limitation. Our proposed method offers a diverse and accurate solution for HMP as demonstrated in the range of evaluation metrics. However, current metrics, specifically APD in measuring the diversity, could not measure the reliability of predicted poses. To check for failure cases / poses, we have to visualize the predicted outputs and manually inspect them. A new metric tailored to assessing the reasonability of predicted actions in diverse results is necessary to enhance the evaluation process. **Future Work.** Our proposed pipeline demonstrates potential applicability to different backbones. Investigating its generalization capability across various backbone models is a promising avenue for future research.

Table 5: Diversity Loss Only v.s. DIM.

Method	Human3.6M					HumanEva-I				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
Baseline	6.325	0.371	0.484	0.513	0.549	6.300	0.217	0.229	0.346	0.349
Baseline + Diversity Loss	8.261	0.369	0.419	0.483	0.483	6.373	0.217	0.228	0.330	0.350
Baseline + DIM	15.657	0.351	0.408	0.449	0.457	6.515	0.203	0.222	0.323	0.326

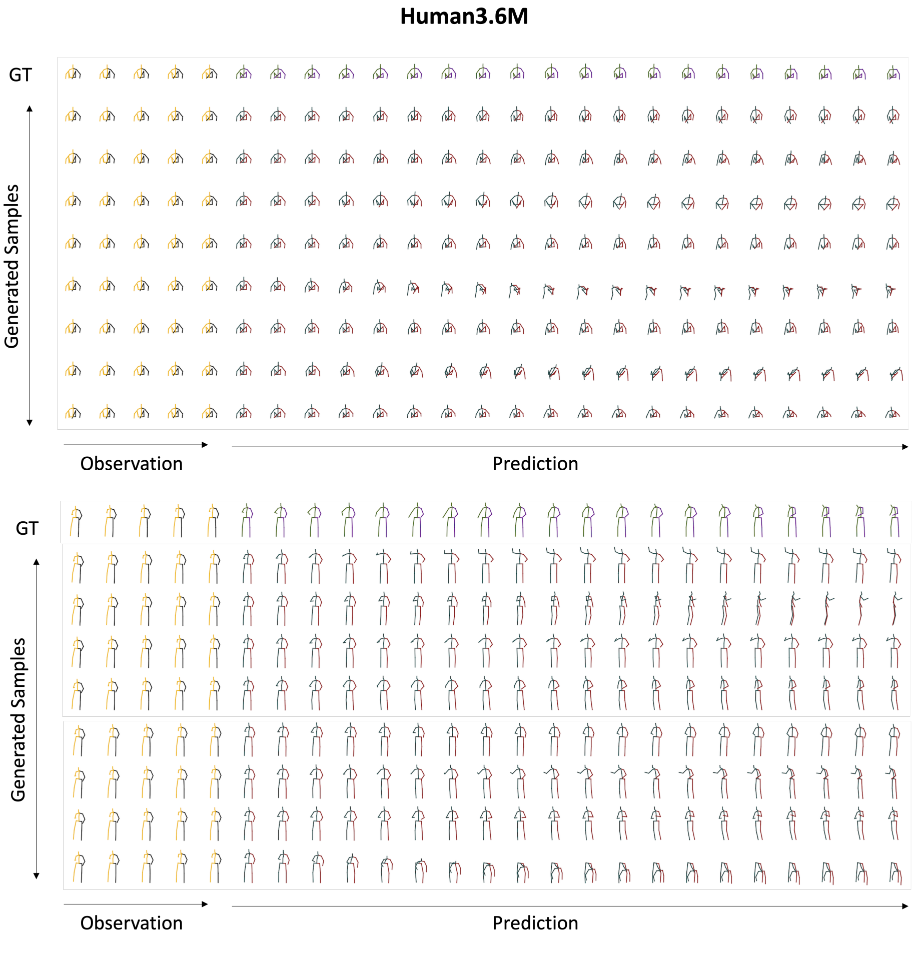


Fig. 5: Generated samples from the Human3.6M dataset. GT denotes the ground truth motion sequence.

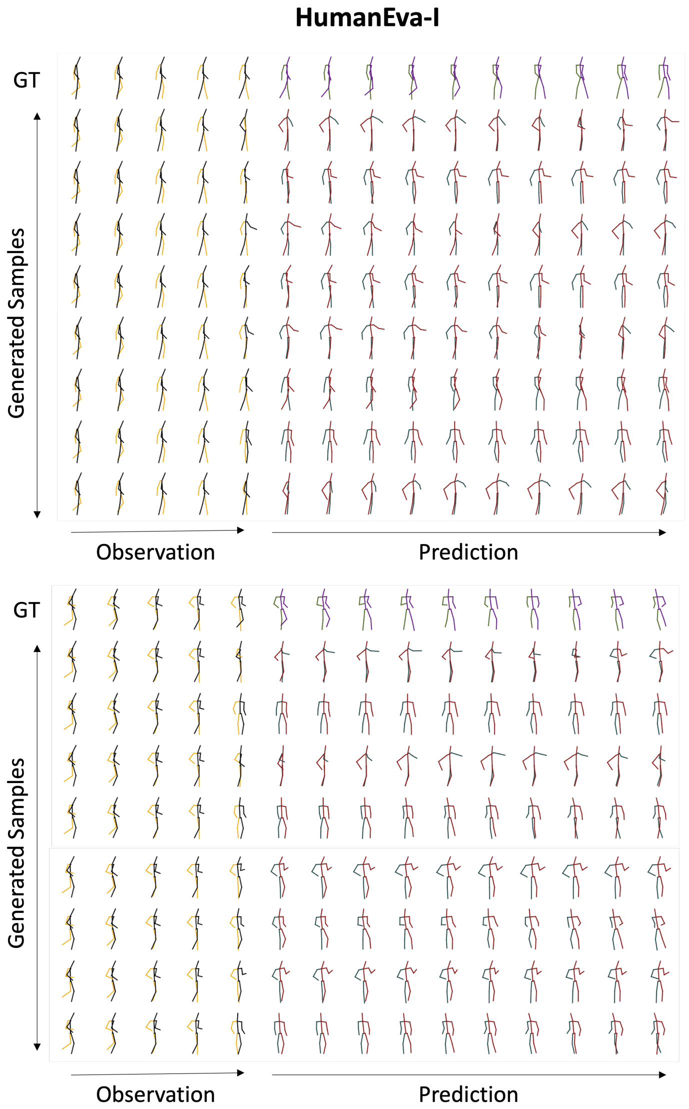


Fig. 6: Generated samples from the HumanEva-I dataset. GT denotes the ground truth motion sequence.

References

1. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. arXiv preprint arXiv:2205.14756 (2022) 3
2. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9544–9555 (October 2023) 1
3. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer (2017) 2
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017) 2