Weakly-supervised Action Transition Learning for Stochastic Human Motion Prediction
— Supplementary Material —

Wei Mao[1], Miaomiao Liu[1], Mathieu Salzmann[2,3]
[1]Australian National University; [2]CVLab, EPFL; [3]ClearSpace, Switzerland
{wei.mao, miaomiao.liu}@anu.edu.au, mathieu.salzmann@epfl.ch

1. Weakly-supervised Transition Learning

We illustrate our weakly-supervised transition learning in Fig. 1 below. Given past N frames $\mathbf{X}_{1:N}$, our model will predict future $T + T_0 + P$ frames. The first T_0 frames are the transition between the past and future motion where a weak supervision signal ($\mathcal{L}_{\text{smooth}}$) is applied during training. The last P frames are supervised by the repeats of last pose in future motion included in the reconstruction loss ($\mathcal{L}_{\text{rec}}$). This encourages the model to predict static poses after predicting the future motion.

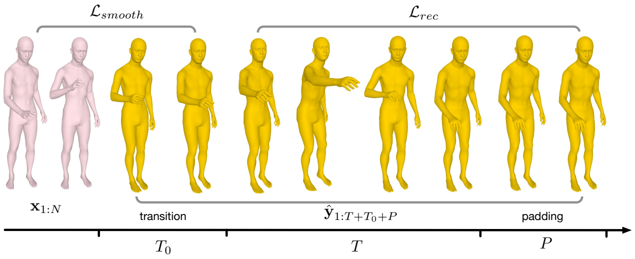

Figure 1. An illustration of our weakly supervision transition learning.

2. Results on HumanAct12

HumanAct12 (HAct) [2] is a subset of the PHSPD dataset [12]. It consists of 12 subjects performing 12 actions. We use 2 subjects (P11, P12) for testing and the remaining 10 subjects (P1-P10) for training. The minimum and maximum length of the motions are 35 and 290 frames, respectively. We remove the motions that are too short, i.e., less than 35 frames, leading to 727 training sequences and 197 testing ones. Our model is trained to observe 10 frames.

The results on HAct [2] are shown in Tables 1, 2 and 3.

	Method	Acc↑	$FID_{\text{tr}}\downarrow$	$FID_{\text{te}}\downarrow$	$Div_{\text{w}}\uparrow$	$Div\uparrow$
HAct	Act2Mot [2]	$24.5^{\pm 0.1}$	$245.35^{\pm 7.13}$	$298.06^{\pm 10.80}$	$0.31^{\pm 0.00}$	$0.60^{\pm 0.01}$
	DLow [11]	$22.7^{\pm 0.2}$	$254.72^{\pm 8.48}$	$143.71^{\pm 3.07}$	$0.35^{\pm 0.00}$	$0.53^{\pm 0.00}$
	ACTOR [7]	$44.4^{\pm 0.2}$	$248.81^{\pm 3.77}$	$381.56^{\pm 6.66}$	$0.84^{\pm 0.00}$	$0.95^{\pm 0.00}$
	Ours (RNN)	$\mathbf{59.0}^{\pm 0.1}$	$\mathbf{129.95}^{\pm 0.39}$	$164.38^{\pm 2.27}$	$\mathbf{0.74}^{\pm 0.00}$	$\mathbf{0.96}^{\pm 0.00}$
	Ours (Tran.)	$56.8^{\pm 0.2}$	$141.85^{\pm 2.51}$	$\mathbf{139.82}^{\pm 1.80}$	$0.67^{\pm 0.00}$	$0.88^{\pm 0.00}$

Table 1. **Quantitative results on HumanAct12 [2].** We adapt Action2Motion [2], ACTOR [7] and DLow [11] to our task.

	HAct	
Method	$ADE_{\text{w}}\downarrow$	$ADE\downarrow$
Act2Mot [2]	$1.09^{\pm 0.00}$	$1.38^{\pm 0.01}$
DLow [11]	$1.15^{\pm 0.01}$	$1.39^{\pm 0.01}$
ACTOR [7]	$1.21^{\pm 0.02}$	$1.54^{\pm 0.01}$
Ours (RNN)	$1.06^{\pm 0.01}$	$\mathbf{1.23}^{\pm 0.01}$
Ours (Tran.)	$\mathbf{1.02}^{\pm 0.01}$	$1.26^{\pm 0.01}$

Table 2. **Results of prediction accuracy with the ground truth action label.**

		Prediction Step				
	Metrics	1_{st}	2_{nd}	3_{rd}	4_{th}	5_{th}
HAct	Acc↑	$59.0^{\pm 0.1}$	$60.9^{\pm 1.0}$	$60.9^{\pm 0.7}$	$60.8^{\pm 0.8}$	$60.3^{\pm 0.5}$
	$FID_{\text{tr}}\downarrow$	$129.95^{\pm 0.39}$	$148.49^{\pm 9.44}$	$159.06^{\pm 10.27}$	$158.88^{\pm 8.93}$	$166.06^{\pm 9.41}$
	$FID_{\text{te}}\downarrow$	$164.38^{\pm 2.27}$	$240.23^{\pm 7.57}$	$259.46^{\pm 7.52}$	$258.19^{\pm 13.83}$	$262.20^{\pm 19.21}$
	$Div_{\text{w}}\uparrow$	$0.74^{\pm 0.00}$	$0.81^{\pm 0.01}$	$0.80^{\pm 0.01}$	$0.81^{\pm 0.01}$	$0.81^{\pm 0.00}$
	$Div\uparrow$	$0.96^{\pm 0.00}$	$1.05^{\pm 0.01}$	$1.04^{\pm 0.02}$	$1.06^{\pm 0.01}$	$1.06^{\pm 0.01}$

Table 3. **Results on prediction with action label sequences.**

3. Implementation Details

In this section, we describe the implementation details for each dataset and model. In Tables 4 and 5, we show the detailed network design and hyperparameters used in our experiments for the RNN-based and Transformer-based models, respectively.

Dataset	Feature Size		λ_{rec}	λ_{smooth}	LR	P
	GRU	MLP				
GRAB	128	[300, 200, 128]	100.0	100.0	0.002	50
NTU	128	[300, 200, 128]	100.0	20.0	0.002	50
BABEL	256	[512, 256, 256]	50.0	10.0	0.001	50
HAct	128	[300, 200, 128]	100.0	100.0	0.002	50

Table 4. **Implementation details of our RNN-based model.** We show the network design, loss weights (λ_{rec} and λ_{smooth}), learning rate (LR) and number of padding frames (P). Note that the posterior and prior modules share the same network architecture.

Dataset	Transformer d_model	 nhead	 dim_feedforward	 num_layers	λ_{rec}	λ_{smooth}	LR	P
GRAB	128	4	256	6	1000.0	100.0	0.0005	20
NTU	128	4	512	8	100.0	20.0	0.0005	20
BABEL	128	4	256	6	100.0	10.0	0.0001	20
HAct	128	4	256	6	1000.0	100.0	0.0005	50

Table 5. **Implementation details of our Transformer-based model.**

Training details. We mainly train our RNN-based models on an NVIDIA TITAN-V GPU. Training for 500 epochs takes 2-12 hours. For Transformer-based models, since they consume much more GPU memory, we train them on an NVIDIA RTX3090 GPU, which takes 5-60 hours.

		Method	Acc↑	$FID_{\text{tr}}\downarrow$	$FID_{\text{te}}\downarrow$	$Div_{\text{w}}\uparrow$	$Div\uparrow$
GRAB	RNN	MLP	**93.5**$^{\pm0.6}$	**33.18**$^{\pm1.27}$	43.86$^{\pm1.08}$	**1.16**$^{\pm0.01}$	**1.39**$^{\pm0.01}$
		linear func.	92.6$^{\pm0.6}$	44.59$^{\pm1.39}$	**38.03**$^{\pm1.49}$	1.10$^{\pm0.01}$	1.37$^{\pm0.01}$
	Tran.	MLP	65.1$^{\pm1.2}$	167.14$^{\pm3.23}$	55.41$^{\pm2.08}$	0.02$^{\pm0.00}$	0.00$^{\pm0.00}$
		linear func.	**85.5**$^{\pm1.2}$	**48.58**$^{3.05}$	**25.72**$^{\pm2.16}$	**1.05**$^{\pm0.01}$	**1.08**$^{\pm0.01}$
NTU	RNN	MLP	**76.3**$^{\pm0.2}$	79.55$^{\pm1.32}$	113.61$^{\pm0.89}$	**1.36**$^{\pm0.00}$	**2.20**$^{\pm0.00}$
		linear func.	76.0$^{\pm0.2}$	**72.18**$^{\pm0.93}$	**111.01**$^{\pm1.28}$	1.25$^{\pm0.00}$	**2.20**$^{\pm0.00}$
	Tran.	MLP	61.2$^{\pm0.1}$	316.15$^{\pm6.04}$	237.20$^{\pm3.59}$	0.00$^{\pm0.00}$	0.01$^{\pm0.00}$
		linear func.	**71.3**$^{\pm0.2}$	**83.14**$^{\pm1.74}$	**114.62**$^{\pm0.93}$	**1.25**$^{\pm0.00}$	**2.19**$^{\pm0.01}$
BABEL	RNN	MLP	**54.4**$^{\pm0.4}$	**20.86**$^{\pm0.29}$	**21.46**$^{\pm0.35}$	**1.55**$^{\pm0.00}$	**1.78**$^{\pm0.00}$
		linear func.	49.6$^{\pm0.4}$	22.54$^{\pm0.27}$	22.39$^{\pm0.36}$	1.35$^{\pm0.00}$	1.74$^{\pm0.00}$
	Tran.	MLP	37.9$^{\pm0.5}$	44.05$^{\pm0.68}$	41.92$^{\pm0.71}$	0.00$^{\pm0.00}$	0.00$^{\pm0.00}$
		linear func.	**39.5**$^{\pm0.3}$	**20.02**$^{\pm0.24}$	**19.41**$^{\pm0.35}$	**1.39**$^{\pm0.00}$	**1.82**$^{\pm0.01}$
HAct	RNN	MLP	**61.0**$^{\pm0.0}$	140.34$^{\pm1.15}$	**142.59**$^{\pm2.36}$	**0.88**$^{\pm0.00}$	**1.06**$^{\pm0.00}$
		linear func.	59.0$^{\pm0.1}$	**129.95**$^{\pm0.39}$	164.38$^{\pm2.27}$	0.74$^{\pm0.00}$	0.96$^{\pm0.00}$
	Tran.	MLP	47.5$^{\pm0.3}$	217.63$^{\pm3.33}$	206.94$^{\pm2.60}$	0.01$^{\pm0.00}$	0.01$^{\pm0.00}$
		linear func.	**56.8**$^{\pm0.2}$	**141.85**$^{\pm2.51}$	**139.82**$^{\pm1.80}$	**0.67**$^{\pm0.00}$	**0.88**$^{\pm0.00}$

Table 6. **Ablation study** on the function to produce T_0. The model we choose is highlighted with a blue background.

4. Functions for T_0

As described in the main paper, we formulate the number of transition frames T_0 between pose sequences from two different motions as a function

$$T_0 = f(\mathbf{x}_N, \mathbf{y}'_1) \,, \tag{1}$$

where $\mathbf{x}_N$ and $\mathbf{y}'_1$ are the last pose of the historical sequences ($\mathbf{X}$) and the first pose of future sequences ($\mathbf{Y}'$), respectively.

We then define this as a simple linear function

$$T_0 = \lfloor k \|\mathbf{x}_N - \mathbf{y}'_1\|_2 \rfloor \,, \tag{2}$$

where $k \geq 0$ is calculated from the training data. More specifically, given pose sequences $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots]$ from the training set ($\mathcal{D}$), k is computed as

$$k = \mathop{\mathbb{E}}_{\mathbf{X} \in \mathcal{D}, i \neq j} \left[\frac{|i-j|}{\|\mathbf{x}_i - \mathbf{x}_j\|_2} \right] \,. \tag{3}$$

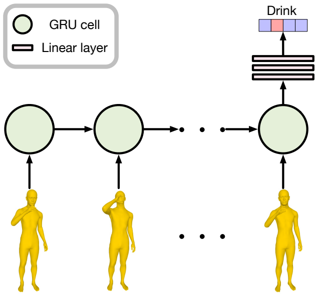

Figure 2. **Network structure** of the action recognition model.

To demonstrate the effectiveness of such a simple linear function, we learn a more complicated one, modeled by a multilayer perception (MLP), from training data. As shown in Table 6, for RNN-based models both functions perform on par with each other. By contrast, for Transformer-based models, the linear function yields better performance than the MLP one.

5. Action Recognition Model

Dataset	GRAB	NTU	BABEL	HAct
Acc	85.5	89.5	60.8	70.4
FID_{tr}	116.46	127.44	9.62	127.77

Table 7. **Results of action recognition model.** We report the action recognition accuracy (Acc) on the test set, the FID of test set to training data (FID_{tr}).

In Fig. 2, we provide the network structure of the action recognition model used for evaluation. It consists of a GRU layer to encode the temporal information and of a 3-layer MLP to produce the final classification results. Each layer of the MLP is followed by a ReLU [6] activation function, except for the last layer, which is followed by a sigmoid function. The model is then trained for 500 epochs using the ADAM [4] optimizer with an initial learning rate of 0.002.

To compute the Fréchet Inception Distance (FID) [3], we take the features generated by the penultimate layer of the pretrained MLP. We then collect three sets of features computed from the training data, the testing data and generated motions, respectively. Each set of future motions is summarized as a multivariate Gaussian by calculating the mean and covariance matrix of these features. The FID then measures the distance between pairs of Gaussian distributions. We report the action recognition results, as well as the FID between the test set and the training data in Table 7.

6. Additional Details

Smoothness prior. We use $L = 10$ frames for our smoothness prior, which means that we take the last 10

		P	Acc↑	FID_{tr} ↓	FID_{te} ↓	Div_{w} ↑	Div ↑
GRAB	RNN	10	$91.0^{\pm0.4}$	$54.61^{\pm1.32}$	$\mathbf{29.86}^{\pm0.69}$	$\mathbf{1.18}^{\pm0.01}$	$1.34^{\pm0.01}$
		20	$90.6^{\pm0.6}$	$\mathbf{37.79}^{\pm1.05}$	$30.80^{\pm0.90}$	$1.11^{\pm0.01}$	$1.34^{\pm0.01}$
		50	$\mathbf{92.6}^{\pm0.6}$	$44.59^{\pm1.39}$	$38.03^{\pm1.49}$	$1.10^{\pm0.01}$	$\mathbf{1.37}^{\pm0.01}$
	Tran.	10	$77.4^{\pm1.0}$	$75.65^{\pm2.08}$	$\mathbf{17.12}^{\pm0.93}$	$\mathbf{1.16}^{\pm0.01}$	$1.12^{\pm0.01}$
		20	$\mathbf{85.5}^{\pm1.2}$	$48.58^{\pm3.05}$	$25.72^{\pm2.16}$	$1.05^{\pm0.01}$	$1.08^{\pm0.01}$
		50	$85.3^{\pm0.6}$	$\mathbf{41.71}^{\pm2.07}$	$30.91^{\pm1.30}$	$1.12^{\pm0.01}$	$\mathbf{1.22}^{\pm0.01}$
NTU	RNN	10	$75.6^{\pm0.1}$	$76.53^{\pm1.22}$	$126.15^{\pm1.28}$	$\mathbf{1.27}^{\pm0.00}$	$2.05^{\pm0.01}$
		20	$75.9^{\pm0.1}$	$75.64^{\pm1.04}$	$120.30^{\pm0.93}$	$1.23^{\pm0.00}$	$2.12^{\pm0.01}$
		50	$\mathbf{76.0}^{\pm0.2}$	$\mathbf{72.18}^{\pm0.93}$	$\mathbf{111.01}^{\pm1.28}$	$1.25^{\pm0.00}$	$\mathbf{2.20}^{\pm0.01}$
	Tran.	10	$71.1^{\pm0.1}$	$88.53^{\pm1.25}$	$127.21^{\pm1.48}$	$\mathbf{1.27}^{\pm0.00}$	$1.88^{\pm0.01}$
		20	$\mathbf{71.3}^{\pm0.2}$	$\mathbf{83.14}^{\pm1.74}$	$\mathbf{114.62}^{\pm0.93}$	$1.25^{\pm0.00}$	$\mathbf{2.19}^{\pm0.01}$
		50	$57.3^{\pm0.1}$	$478.00^{\pm6.77}$	$328.87^{\pm3.98}$	$0.00^{\pm0.00}$	$0.00^{\pm0.00}$
BABEL	RNN	10	$48.8^{\pm0.2}$	$33.40^{\pm0.26}$	$32.55^{\pm0.35}$	$\mathbf{1.48}^{\pm0.00}$	$\mathbf{1.71}^{\pm0.00}$
		20	$\mathbf{49.8}^{\pm0.3}$	$27.05^{\pm0.25}$	$26.56^{\pm0.34}$	$1.42^{\pm0.00}$	$\mathbf{1.71}^{\pm0.01}$
		50	$49.6^{\pm0.4}$	$\mathbf{22.54}^{\pm0.27}$	$\mathbf{22.39}^{\pm0.36}$	$1.35^{\pm0.00}$	$1.74^{\pm0.00}$
	Tran.	10	$38.6^{\pm0.3}$	$22.78^{\pm0.31}$	$22.03^{\pm0.42}$	$\mathbf{1.45}^{\pm0.01}$	$1.77^{\pm0.01}$
		20	$\mathbf{39.5}^{\pm0.3}$	$20.02^{\pm0.24}$	$19.41^{\pm0.35}$	$1.39^{\pm0.00}$	$\mathbf{1.82}^{\pm0.01}$
		50	$37.1^{\pm0.3}$	$\mathbf{19.65}^{\pm0.27}$	$\mathbf{18.93}^{\pm0.37}$	$1.20^{\pm0.00}$	$1.73^{\pm0.01}$
HAct	RNN	10	$50.3^{\pm0.1}$	$159.19^{\pm2.58}$	$164.58^{\pm2.97}$	$0.78^{\pm0.00}$	$\mathbf{1.06}^{\pm0.00}$
		20	$58.1^{\pm0.2}$	$\mathbf{121.70}^{\pm1.31}$	$186.60^{\pm3.23}$	$\mathbf{0.79}^{\pm0.00}$	$0.97^{\pm0.00}$
		50	$\mathbf{59.0}^{\pm0.1}$	$129.95^{\pm0.39}$	$\mathbf{164.38}^{\pm2.27}$	$0.74^{\pm0.00}$	$0.96^{\pm0.01}$
	Tran.	10	$47.6^{\pm0.2}$	$214.35^{\pm4.38}$	$151.66^{\pm2.62}$	$0.66^{\pm0.00}$	$0.79^{\pm0.00}$
		20	$50.0^{\pm0.1}$	$186.28^{\pm3.73}$	$140.95^{\pm2.43}$	$\mathbf{0.71}^{\pm0.00}$	$0.87^{\pm0.00}$
		50	$\mathbf{56.8}^{\pm0.2}$	$\mathbf{141.85}^{\pm2.51}$	$\mathbf{139.82}^{\pm1.80}$	$0.67^{\pm0.00}$	$\mathbf{0.88}^{\pm0.00}$

Table 8. **Ablation study** on the number of frames to predict (P). The model we choose is highlighted with a blue background.

		δ	Acc↑	FID_{tr} ↓	FID_{te} ↓	Div_{w} ↑	Div ↑
GRAB	RNN	0.005	$78.4^{\pm0.4}$	$152.74^{\pm2.33}$	$\mathbf{21.90}^{\pm1.12}$	$\mathbf{1.17}^{\pm0.01}$	$1.37^{\pm0.01}$
		0.015	$\mathbf{92.6}^{\pm0.6}$	$\mathbf{44.59}^{\pm1.39}$	$38.03^{\pm1.49}$	$1.10^{\pm0.01}$	$1.37^{\pm0.01}$
		0.025	$88.1^{\pm0.4}$	$47.93^{\pm1.83}$	$31.55^{\pm1.26}$	$1.07^{\pm0.01}$	$1.37^{\pm0.01}$
	Tran.	0.005	$51.1^{\pm0.4}$	$276.29^{\pm5.44}$	$111.65^{\pm3.22}$	$1.02^{\pm0.01}$	$1.08^{\pm0.01}$
		0.015	$\mathbf{85.5}^{\pm1.2}$	$48.58^{\pm3.05}$	$\mathbf{25.72}^{\pm2.16}$	$\mathbf{1.05}^{\pm0.01}$	$1.08^{\pm0.01}$
		0.025	$82.7^{\pm1.0}$	$\mathbf{43.04}^{\pm2.46}$	$27.75^{\pm1.65}$	$1.03^{\pm0.01}$	$1.08^{\pm0.01}$
NTU	RNN	0.005	$78.8^{\pm0.2}$	$71.22^{\pm0.63}$	$182.79^{\pm1.97}$	$\mathbf{1.39}^{\pm0.00}$	$2.20^{\pm0.01}$
		0.015	$\mathbf{78.9}^{\pm0.2}$	$\mathbf{51.19}^{\pm0.54}$	$129.20^{\pm1.52}$	$1.29^{\pm0.00}$	$2.20^{\pm0.01}$
		0.025	$76.0^{\pm0.2}$	$72.18^{\pm0.93}$	$\mathbf{111.01}^{\pm1.28}$	$1.25^{\pm0.00}$	$2.20^{\pm0.01}$
	Tran.	0.005	$62.2^{\pm0.2}$	$256.95^{\pm1.55}$	$257.44^{\pm1.90}$	$\mathbf{1.80}^{\pm0.01}$	$2.19^{\pm0.01}$
		0.015	$\mathbf{72.7}^{\pm0.2}$	$\mathbf{77.45}^{\pm0.76}$	$158.81^{\pm2.07}$	$1.38^{\pm0.00}$	$2.19^{\pm0.01}$
		0.025	$71.3^{\pm0.2}$	$83.14^{\pm1.74}$	$\mathbf{114.62}^{\pm0.93}$	$1.25^{\pm0.00}$	$2.19^{\pm0.01}$
BABEL	RNN	0.005	$45.4^{\pm0.2}$	$45.00^{\pm0.10}$	$43.40^{\pm0.30}$	$1.57^{\pm0.00}$	$1.74^{\pm0.00}$
		0.015	$\mathbf{51.1}^{\pm0.3}$	$26.00^{\pm0.24}$	$25.80^{\pm0.34}$	$\mathbf{1.43}^{\pm0.00}$	$1.74^{\pm0.00}$
		0.025	$49.6^{\pm0.4}$	$\mathbf{22.50}^{\pm0.27}$	$\mathbf{22.40}^{\pm0.36}$	$1.35^{\pm0.00}$	$1.74^{\pm0.00}$
	Tran.	0.005	$31.7^{\pm0.1}$	$68.54^{\pm0.32}$	$64.92^{\pm0.27}$	$\mathbf{1.78}^{\pm0.01}$	$1.82^{\pm0.01}$
		0.015	$\mathbf{39.5}^{\pm0.2}$	$31.04^{\pm0.20}$	$30.18^{\pm0.34}$	$1.58^{\pm0.00}$	$1.82^{\pm0.01}$
		0.025	$\mathbf{39.5}^{\pm0.3}$	$\mathbf{20.02}^{\pm0.24}$	$\mathbf{19.41}^{\pm0.35}$	$1.39^{\pm0.00}$	$1.82^{\pm0.01}$
HAct	RNN	0.01	$\mathbf{59.0}^{\pm0.1}$	$\mathbf{129.95}^{\pm0.39}$	$164.38^{\pm2.27}$	$\mathbf{0.74}^{\pm0.00}$	$0.96^{\pm0.00}$
		0.015	$54.2^{\pm0.1}$	$215.68^{\pm2.99}$	$\mathbf{158.94}^{\pm2.09}$	$0.71^{\pm0.00}$	$0.96^{\pm0.01}$
		0.02	$50.4^{\pm0.1}$	$283.92^{\pm4.27}$	$179.30^{\pm2.49}$	$0.68^{\pm0.00}$	$0.96^{\pm0.01}$
	Tran.	0.01	$50.9^{\pm0.1}$	$402.58^{\pm7.32}$	$759.53^{\pm10.26}$	$\mathbf{0.72}^{\pm0.00}$	$0.88^{\pm0.00}$
		0.015	$\mathbf{59.7}^{\pm0.2}$	$\mathbf{97.92}^{\pm1.37}$	$232.52^{\pm3.90}$	$0.69^{\pm0.00}$	$0.88^{\pm0.00}$
		0.02	$56.8^{\pm0.2}$	$141.85^{\pm2.51}$	$\mathbf{139.82}^{\pm1.80}$	$0.67^{\pm0.00}$	$0.88^{\pm0.00}$

Table 9. **Ablation study** on the stop threshold (δ). The model we choose is highlighted with a blue background.

poses of the history and the first 10 ones of the future motion to form a sequence of length 20. We set the number of DCT bases (M) to 5.

Variable-length motion prediction. To enable predicting variable-length sequences, during training, we make the model generate P additional frames, and supervise these frames with the last pose of the future to encourage the model to generate static poses (i.e., the last pose of the ground-truth motion) after reaching the motion end. We validate the number of additional frames to generate (P). The numerical results are shown in Table 8.

During test time, we stop the prediction when the variance of the last Q consecutive frames falls below a threshold (δ). For all our experiments, Q is set to 5, and the stopping threshold δ ranges from 0.005 to 0.025. In Table 9, we validate this threshold for the different datasets.

		Method	Acc↑	FID_{tr} ↓	FID_{te} ↓	Div_{w} ↑	Div ↑
GRAB	RNN	stop sign	$66.5^{\pm1.0}$	$211.11^{\pm6.89}$	$\mathbf{31.49}^{\pm3.35}$	$\mathbf{1.31}^{\pm0.01}$	$\mathbf{1.44}^{\pm0.01}$
		padding	$\mathbf{92.6}^{\pm0.6}$	$\mathbf{44.59}^{\pm1.39}$	$38.03^{\pm1.49}$	$1.10^{\pm0.01}$	$1.37^{\pm0.01}$
	Tran.	stop sign	$76.3^{\pm1.4}$	$101.07^{\pm4.93}$	$\mathbf{16.02}^{\pm1.03}$	$0.05^{\pm0.00}$	$0.27^{\pm0.00}$
		padding	$\mathbf{85.5}^{\pm1.2}$	$\mathbf{48.58}^{3.05}$	$25.72^{\pm2.16}$	$\mathbf{1.05}^{\pm0.01}$	$\mathbf{1.08}^{\pm0.01}$
NTU	RNN	stop sign	$29.2^{\pm0.1}$	$1188.50^{\pm2.72}$	$1234.79^{\pm3.99}$	$1.13^{\pm0.00}$	$1.50^{\pm0.00}$
		padding	$\mathbf{76.0}^{\pm0.2}$	$\mathbf{72.18}^{\pm0.93}$	$\mathbf{111.01}^{\pm1.28}$	$\mathbf{1.25}^{\pm0.00}$	$\mathbf{2.20}^{\pm0.00}$
	Tran.	stop sign	$\mathbf{77.7}^{\pm0.2}$	$259.22^{\pm2.04}$	$358.79^{\pm2.68}$	$0.00^{\pm0.00}$	$0.01^{\pm0.00}$
		padding	$71.3^{\pm0.2}$	$\mathbf{83.14}^{\pm1.74}$	$\mathbf{114.62}^{\pm0.93}$	$\mathbf{1.25}^{\pm0.00}$	$\mathbf{2.19}^{\pm0.01}$
BABEL	RNN	stop sign	$13.3^{\pm0.1}$	$162.76^{\pm3.23}$	$164.01^{\pm3.25}$	$0.93^{\pm0.00}$	$\mathbf{2.14}^{\pm0.01}$
		padding	$\mathbf{49.6}^{\pm0.4}$	$\mathbf{22.50}^{\pm0.27}$	$\mathbf{22.40}^{\pm0.36}$	$\mathbf{1.35}^{\pm0.00}$	$1.74^{\pm0.00}$
	Tran.	stop sign	$15.2^{\pm0.3}$	$25.43^{\pm0.29}$	$22.26^{\pm0.44}$	$0.46^{\pm0.01}$	$1.23^{\pm0.00}$
		padding	$\mathbf{39.5}^{\pm0.3}$	$\mathbf{20.02}^{\pm0.24}$	$\mathbf{19.41}^{\pm0.35}$	$\mathbf{1.39}^{\pm0.00}$	$\mathbf{1.82}^{\pm0.01}$
HAct	RNN	stop sign	$43.7^{\pm0.2}$	$\mathbf{103.61}^{\pm1.94}$	$222.14^{\pm6.66}$	$\mathbf{0.76}^{\pm0.00}$	$0.84^{\pm0.00}$
		padding	$\mathbf{59.0}^{\pm0.1}$	$129.95^{\pm0.39}$	$\mathbf{164.38}^{\pm2.27}$	$0.74^{\pm0.00}$	$\mathbf{0.96}^{\pm0.00}$
	Tran.	stop sign	$47.7^{\pm0.2}$	$\mathbf{112.33}^{\pm3.10}$	$221.10^{\pm5.62}$	$0.04^{\pm0.00}$	$0.05^{\pm0.00}$
		padding	$\mathbf{56.8}^{\pm0.2}$	$141.85^{\pm2.51}$	$\mathbf{139.82}^{\pm1.80}$	$\mathbf{0.67}^{\pm0.00}$	$\mathbf{0.88}^{\pm0.00}$

Table 10. **Ablation study** on generating variable-length future motions. The model we choose is highlighted with a blue background.

As an alternative to the above-mentioned padding approach to variable-length prediction, we draw inspiration from the NLP literature. However, while the standard strategy for variable-length outputs in NLP is to predict a stop token, there is no such "stop token" for human pose. Thus, we make the model output one more value together with the human pose at each prediction step. The additional value is supervised by a binary label indicating whether the motion ends or not. Such a value is then used as a "stop sign" during testing. We compare the results of two the strategies described above (stop sign vs padding) in Table 10. In general, the models with padding perform better than those with the stop sign.

7. Stopping Strategy for Action2Motion

Due to the jitter produced by the Action2Motion [2] model, we have observed our variance-based stopping criterion to be sub-optimal. We therefore test a stopping strategy based on the difference between consecutive frames. That is, if the difference between the latest 2 frames falls below certain threshold δ, the model stops predicting the future motions. In Table 11, we compare the performance of the variance-based strategy and the adjacent frames' difference-based one with different stopping thresholds. Overall, the stopping strategy based on the difference of adjacent frames outperforms the variance-based one. In the main paper, we report the values highlighted with a blue background.

8. Details of the Datasets

We evaluate our method on four different datasets, i.e., GRAB [1,10], NTU RGB-D [5,9], BABEL [8] and HumanAct12 [2].

License. These four datasets all are for non-commercial scientific research use only. For the details of their individual licenses, please refer to their official websites:

- GRAB: `https://grab.is.tue.mpg.de/`
- NTU RGB-D: `http://rose1.ntu.edu.sg/Datasets/actionRecognition.asp`
- BABEL: `https://babel.is.tue.mpg.de/data.html`
- HumanAct12: `https://jimmyzou.github.io/publication/2020-PHSPDataset`

9. Additional Qualitative Results

We provide additional qualitative results in the supplementary video.

	Stop criterion	δ	Acc↑	$FID_{\text{tr}}\downarrow$	$FID_{\text{te}}\downarrow$	$Div_{\text{w}}\uparrow$	$Div\uparrow$
GRAB	adjacent frames	0.01	$64.3^{\pm1.2}$	$163.85^{\pm10.14}$	$56.38^{\pm7.02}$	$0.40^{\pm0.00}$	$0.76^{\pm0.01}$
		0.02	$64.7^{\pm1.2}$	$159.14^{\pm9.82}$	$54.35^{\pm6.93}$	$0.43^{\pm0.00}$	$0.76^{\pm0.01}$
		0.04	$\mathbf{70.6}^{\pm1.3}$	$\mathbf{80.22}^{\pm6.64}$	$\mathbf{47.81}^{\pm1.09}$	$\mathbf{0.50}^{\pm0.00}$	$0.76^{\pm0.01}$
	variance based	0.01	$64.3^{\pm1.2}$	$163.85^{\pm10.14}$	$56.38^{\pm7.02}$	$0.40^{\pm0.00}$	$0.76^{\pm0.01}$
		0.02	$64.3^{\pm1.2}$	$163.69^{\pm10.12}$	$56.34^{\pm7.02}$	$0.40^{\pm0.00}$	$0.76^{\pm0.01}$
		0.04	$69.8^{\pm1.2}$	$100.76^{\pm7.78}$	$35.20^{\pm3.04}$	$0.49^{\pm0.01}$	$0.76^{\pm0.01}$
NTU	adjacent frames	0.01	$60.5^{\pm0.2}$	$354.92^{\pm1.77}$	$442.99^{\pm3.80}$	$0.67^{\pm0.01}$	$1.19^{\pm0.01}$
		0.02	$61.9^{\pm0.2}$	$316.37^{\pm1.87}$	$415.69^{\pm3.14}$	$0.68^{\pm0.01}$	$1.19^{\pm0.01}$
		0.04	$\mathbf{66.3}^{\pm0.2}$	$\mathbf{144.98}^{\pm2.44}$	$\mathbf{113.61}^{\pm0.84}$	$\mathbf{0.75}^{\pm0.01}$	$1.19^{\pm0.01}$
	variance based	0.01	$60.5^{\pm0.2}$	$354.86^{\pm1.75}$	$442.80^{\pm3.78}$	$0.67^{\pm0.01}$	$1.19^{\pm0.01}$
		0.02	$60.6^{\pm0.2}$	$350.17^{\pm1.80}$	$439.70^{\pm3.67}$	$0.67^{\pm0.01}$	$1.19^{\pm0.01}$
		0.04	$64.3^{\pm0.2}$	$177.21^{\pm0.96}$	$236.82^{\pm2.28}$	$0.70^{\pm0.01}$	$1.19^{\pm0.01}$
BABEL	adjacent frames	0.01	$\mathbf{15.9}^{\pm0.2}$	$62.95^{\pm0.40}$	$57.39^{\pm0.31}$	$0.76^{\pm0.01}$	$1.10^{\pm0.01}$
		0.02	$\mathbf{15.9}^{\pm0.2}$	$62.81^{\pm0.39}$	$57.26^{\pm0.30}$	$0.76^{\pm0.01}$	$1.10^{\pm0.01}$
		0.04	$14.8^{\pm0.2}$	$\mathbf{42.02}^{\pm0.40}$	$\mathbf{37.41}^{\pm0.47}$	$\mathbf{0.79}^{\pm0.01}$	$1.10^{\pm0.01}$
	variance based	0.01	$\mathbf{15.9}^{\pm0.2}$	$62.95^{\pm0.40}$	$57.39^{\pm0.31}$	$0.76^{\pm0.01}$	$1.10^{\pm0.01}$
		0.02	$\mathbf{15.9}^{\pm0.2}$	$62.93^{\pm0.40}$	$57.38^{\pm0.31}$	$0.76^{\pm0.01}$	$1.10^{\pm0.01}$
		0.04	$15.5^{\pm0.2}$	$56.48^{\pm0.20}$	$51.16^{\pm0.24}$	$0.76^{\pm0.01}$	$1.10^{\pm0.01}$
HAct	adjacent frames	0.01	$25.1^{\pm0.2}$	$388.98^{\pm15.31}$	$585.59^{\pm21.07}$	$0.26^{\pm0.00}$	$0.60^{\pm0.01}$
		0.02	$24.5^{\pm0.1}$	$245.34^{\pm7.13}$	$298.06^{\pm10.79}$	$\mathbf{0.31}^{\pm0.00}$	$0.60^{\pm0.01}$
		0.04	$20.5^{\pm0.2}$	$292.83^{\pm6.37}$	$\mathbf{134.47}^{\pm2.69}$	$\mathbf{0.31}^{\pm0.00}$	$0.60^{\pm0.01}$
	variance based	0.01	$25.0^{\pm0.2}$	$395.59^{\pm15.56}$	$598.76^{\pm21.49}$	$0.25^{\pm0.00}$	$0.60^{\pm0.01}$
		0.02	$\mathbf{25.2}^{\pm0.1}$	$337.77^{\pm12.19}$	$486.46^{\pm16.87}$	$0.28^{\pm0.00}$	$0.60^{\pm0.01}$
		0.04	$22.6^{\pm0.2}$	$\mathbf{224.47}^{\pm5.67}$	$150.82^{\pm4.00}$	$\mathbf{0.31}^{\pm0.00}$	$0.60^{\pm0.01}$

Table 11. **Ablation study** on the stopping criterion for Action2Motion [2]. The model we report results with is highlighted with a blue background.

References

[1] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 4

[2] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1, 3, 4

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 2

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 2

[5] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 42(10):2684–2701, 2019. 4

[6] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010. 2

[7] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*, pages 10985–10995, October 2021. 1

[8] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, June 2021. 4

[9] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, pages 1010–1019, 2016. 4

[10] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 4

[11] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *ECCV*, pages 346–364. Springer, 2020. 1

[12] Shihao Zou, Xinxin Zuo, Yiming Qian, Sen Wang, Chi Xu, Minglun Gong, and Li Cheng. 3d human shape reconstruction from a polarization image. *ECCV*, 2020. 1