Supplementary Material: Diverse 3D Hand Gesture Prediction from Body Dynamics by Bilateral Hand Disentanglement

1. Architecture Details

Body Encoder. The body encoder aims to encode the input upper body skeletons into the body features via an MLP-based architecture. The channel dimension of body features is $C = 128$ in practice.

Bilateral Hand Disentanglement Transformers. We design bilateral hand transformers that interacted with a body-specific transformer. Concretely, we leverage the body features Q to match the key features K and value features V in a single-hand-specific transformer via 3 times Multi-Head Attention (MHA) [5], expressed as:

$$MultiHead_{F_{\mathcal{B}\rightarrow\mathcal{SH}}}(Q,K,V) = softmax(\frac{QK}{\sqrt{d}})V, \quad (1)$$

where d is a normalization constant.

2. More Details about MCMC Sampling

Inspired by [1], we leverage an MLP-based sampling header $S_\alpha(w)$ to model the diversification sampling process, where w indicates the perturbation vector and α is the parameter of sampling header. The prior distribution of perturbation is initialized from an isotropic Gaussian reference distribution, expressed as:

$$p_0(w) = \mathcal{N}(0, \sigma_w^2 I), \quad (2)$$

where the hyperparameter σ_w denotes the standard deviation. The whole sampling process is formulated as:

$$p_\alpha(w) \propto exp\left[-S_\alpha(w) - \frac{1}{2\sigma_w^2}\|w\|^2\right], \quad (3)$$

where the $M_\alpha(w) = S_\alpha(w) + \frac{1}{2\sigma_w^2}\|w\|^2$ is defined as the whole sampling function, and α is the learnable parameters of sampling header. For notation simplicity, let $\beta = \{\theta, \alpha\}$. For the i th sample in a training mini-batch with size n, the log-likelihood function of β is defined as:

$$L(\beta) = \sum_{i=1}^{n} log\left[\int p_\alpha(w_i)p_\theta(\tilde{h}_i|h_i, w_i)dw_i\right]. \quad (4)$$

Thus the gradient of $L(\beta)$ is computed as:

$$\bigtriangledown L(\beta) = E_{p_{\beta(w|\tilde{h},h)}}\left[\bigtriangledown_\alpha logp_\alpha(w) + \bigtriangledown_\theta logp_\theta(\tilde{h}|h,w)\right]. \quad (5)$$

We decompose the $\bigtriangledown L(\beta)$ into two parts. The first part is the gradient for the sampling header with parameter α:

$$\begin{aligned} E_{p_{\beta(w|\tilde{h},h)}}\left[\bigtriangledown_\alpha logp_\alpha(w)\right] &= E_{p_\alpha(w)}\left[\bigtriangledown_\alpha S_\alpha(w)\right] \\ &- E_{p_{\beta(w|\tilde{h},h)}}\left[\bigtriangledown_\alpha S_\alpha(w)\right]. \end{aligned} \quad (6)$$

The second part is the gradient for the hand generation model with parameter θ:

$$\begin{aligned} &E_{p_{\beta(w|\tilde{h},h)}}\left[\bigtriangledown_\theta logp_\theta(\tilde{h}|h,w)\right] \\ &= E_{p_\beta(\tilde{h}|h,w)}\left[\frac{1}{\sigma_\epsilon^2}(\tilde{h} - R_\theta(h,w)) \bigtriangledown_\theta R_\theta(h,w)\right]. \end{aligned} \quad (7)$$

In practice, the terms $\bigtriangledown_\alpha S_\alpha(w)$ in Eq. (6) and $\bigtriangledown_\theta R_\theta(h,w)$ in Eq. (7) are directly computed by back-propagation. The intractable expectation terms $E_p(\cdot)$ in Eq. (6) and Eq. (7) are approximately solved by a gradient-based MCMC (Langevin dynamics) [2]. Specifically, the perturbation is obtained from the MLP-based prior sampling process $M_\alpha(w)$, by iterating:

$$\begin{aligned} w^{l+1} =& w^l - \delta \bigtriangledown_w M_\alpha(w^l) + \sqrt{2\delta}e^l, \\ & w_0 \sim p_0(w), e^l \sim \mathcal{N}(0, I), \end{aligned} \quad (8)$$

where l denotes the l th iteration state, and δ is the step size of Langevin sampling. Meanwhile, the posterior distribution $p_\beta(w|\tilde{h},h)$ of the perturbation is computed by iterating:

$$\begin{aligned} w^{l+1} &= w^l - \\ &\delta\left[\bigtriangledown_w M_\alpha(w^l) - \frac{1}{\sigma_\epsilon^2}(\tilde{h} - R_\theta(h,w^l)) \bigtriangledown_w R_\theta(h,w^l)\right] \\ &+ \sqrt{2\delta}e^l, w_0 \sim p_0(w), e^l \sim \mathcal{N}(0, I). \end{aligned} \quad (9)$$

In the experiments, we set the total iteration state as 6, and the Langevin step sizes of the prior and posterior are 0.4 and 0.1, respectively.

Table 1. Statistics of the B2H, TED Gestures, and our newly collected TED Hands datasets.

Dataset	Speaker Identities	Interest Shots Length	Sequence Numbers	Frame Numbers in a Sequence	Speaker Identities in a Sequence
B2H [3]	8	71.2h	120,188	64	Only one
TED Gestures [6,7]	1,766	106.1h	252,109	34	Multi-identities
TED Hands	1,755	99.6h	134,456	64	Only one

3. More Details about Datasets

The original TED Gestures dataset [6,7] only contains 10 upper body joints without elaborate fingers of two hands. We newly collect a TED Hands dataset based on the raw videos of TED talking speeches. The videos are captured from the official TED channel on YouTube[1]. To obtain reliable 3D hand joints and their corresponding upper body skeletons, we leverage a state-of-the-art 3D human pose estimator Fankmocap [4] for annotation. In particular, we acquire 8 upper body joints and 30 figure joints in our dataset.

Concretely, we split the videos into 64-frame sequences under the following criteria:

- The above-mentioned 38 joints are visible for more than 48 frames in a sequence. Then, we interpolate the sequence to 64 frames.
- Since there might be multiple speakers in a single video of TED talking speeches (*e.g.,* conversation between two speakers). To guarantee the continuity of body-hand movements, we only select the sequence that the joints of 64 frames belonging to a single speaker.

Finally, we obtain 1,755 videos with 134,456 sequences in our TED Hands dataset. The statistics of B2H, TED Gesture, and our TED Hands datasets are reported in Tab. 1. For our TED Hands dataset, the numbers of sequences in each data partition are:

- Training set: 94,125.
- Validation set: 13,446.
- Testing set: 26,885.

4. Additional Visualization Results

Here, we provide more visual results of our method as well as other competitors in the demo video. For more details, please refer to our *project page*. Since all the comparison methods are designed without the diversification setting, we divide the comparisons into two parts. In the first part, we visualize the results of various competitors and the initial predictions of our method. In the second part, we visualize the diversification results based on our initial prediction from stage one.

[1]We obey the TED Talks Team's Creative Commons License (CC BY-NC-ND 4.0 International). In this work, all the videos from TED talking speeches are only used for research.

Moreover, to demonstrate the effectiveness of our proposed loss functions and components, we visualize vital frames of the generated motions based on stage one predictions. As illustrated in Fig. 1 and Fig. 2, we can clearly observe that all combinations of the different loss functions and components have positive impacts on 3D hand predictions.

5. Additional Results on Model Complexity

We calculate the GFlops and inference time on a single NVIDIA RTX 2080 GPU, as reported in Tab. 2. Due to the bilateral hand disentanglement process, the GFlops of our model are moderately higher than the second-best-performing method MRT. However, our method consistently outperforms other methods by large margins on L2, FHD, and MPJRE. The inference time of our model is around 32.921 ms (*i.e.*, faster than 30 FPS). This inference speed allows our method to be deployed in real-time applications.

Table 2. Comparison of model complexity, inference time, and performance on the TED Hands dataset.

Methods	GFlops ↓	Time (ms) ↓	L2 ↓	FHD ↓	MPJRE ↓
Body2hands	0.068	2.823	2.551	1.174	11.371
MRT	0.211	4.341	2.325	0.877	10.314
BTM	0.052	11.089	2.350	1.111	10.440
LTD	0.113	5.962	2.482	1.367	11.078
MotionMixer	0.110	19.071	2.324	0.910	10.427
SPGSN	0.174	52.436	2.435	0.990	10.887
Ours	0.503	32.921	**2.037**	**0.258**	**8.888**

References

[1] Tian Han, Yang Lu, Song-Chun Zhu, and Ying Nian Wu. Alternating back-propagation for generator network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 1

[2] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011. 1

[3] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11865–11874, June 2021. 2

[4] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF*

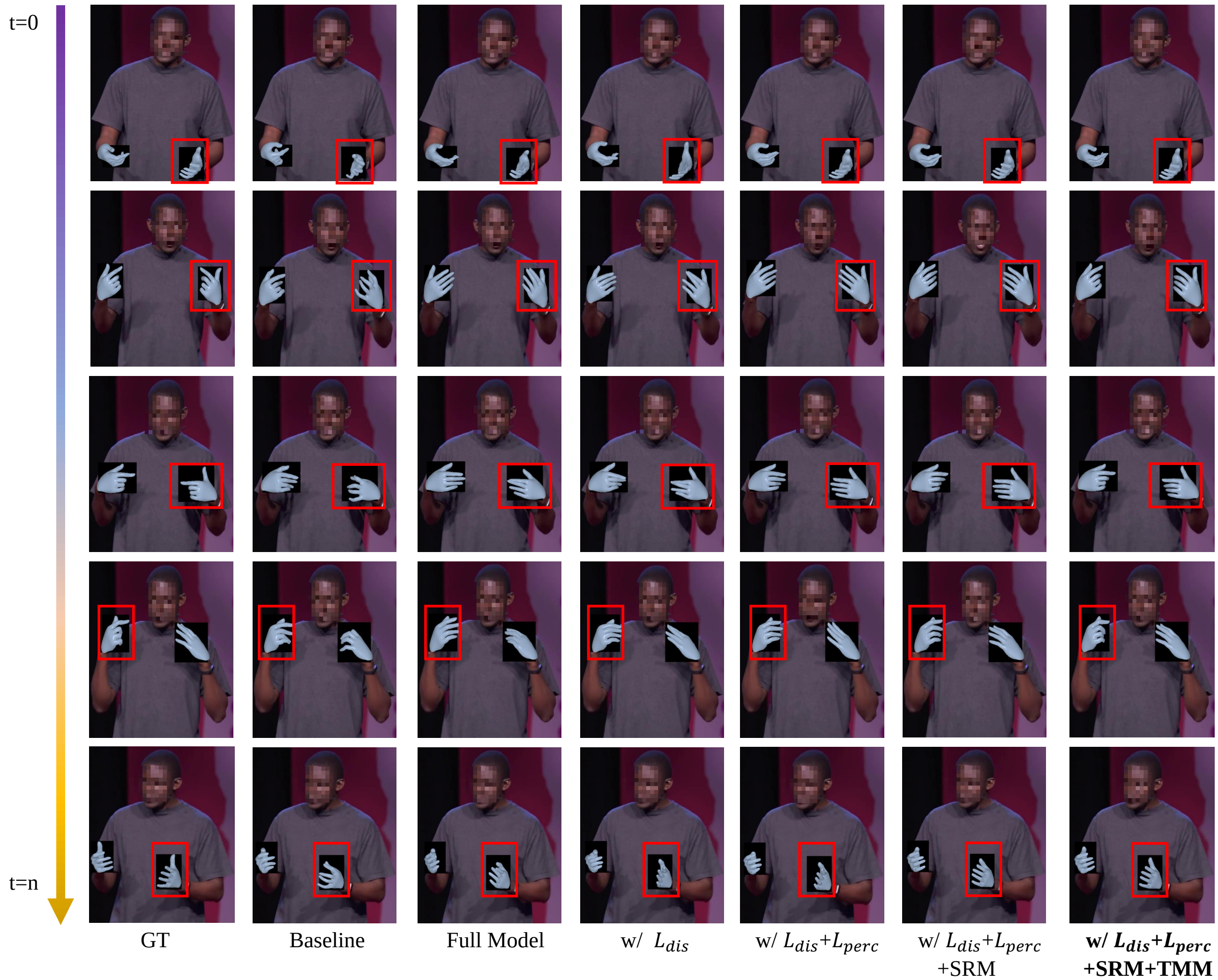

Figure 1. Visual comparisons of ablation study on our newly collected TED Hands dataset. We show the key frames of the generated motions based on stage one initial predictions. Best view on screen.

International Conference on Computer Vision (ICCV) Workshops, pages 1749–1759, October 2021. 2

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[6] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 2

[7] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 2

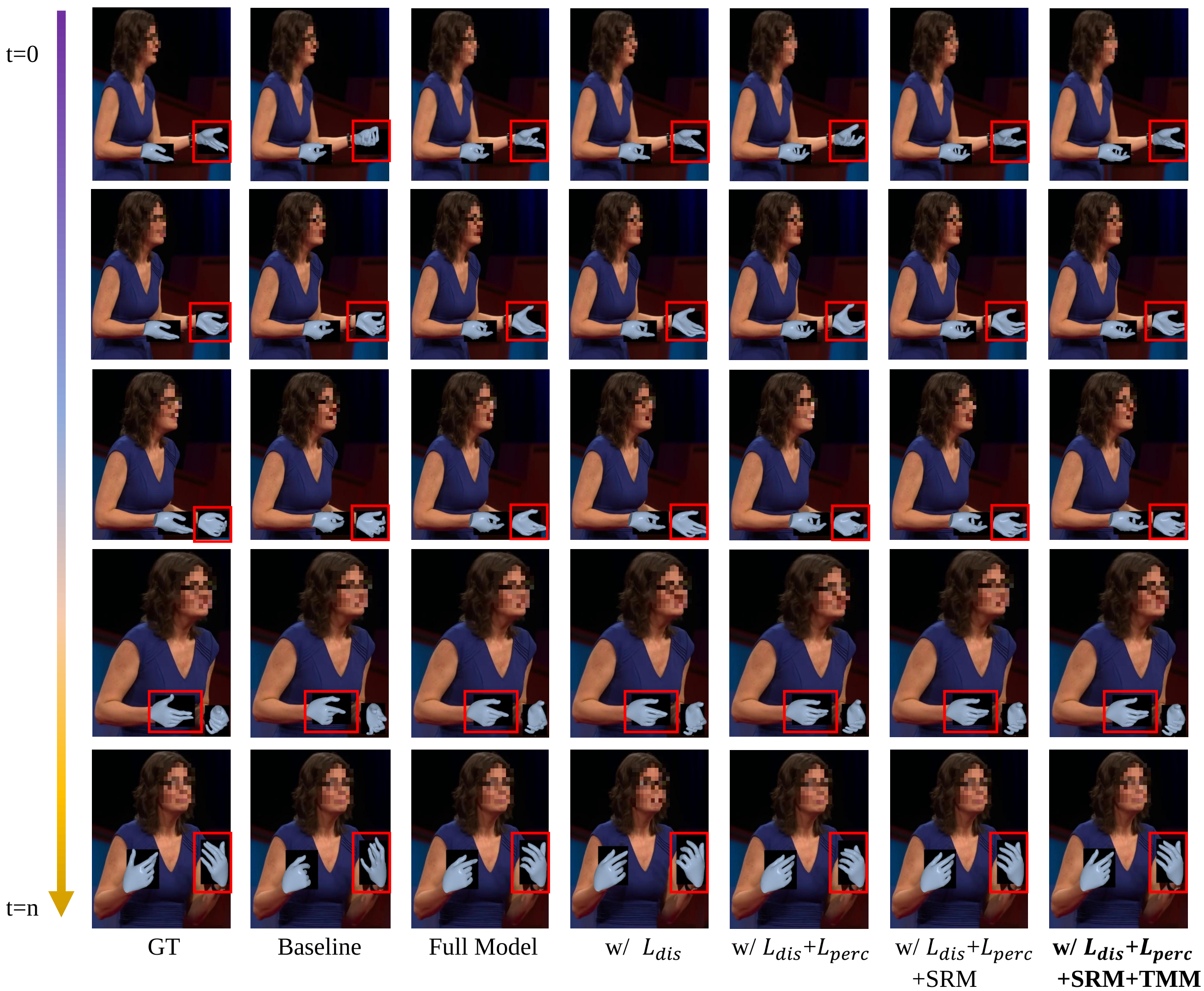

Figure 2. Visual comparisons of ablation study on our newly collected TED Hands dataset. We show the key frames of the generated motions based on stage one initial predictions. Best view on screen.