HERO: Human Reaction Generation from Videos

Supplementary Material

7. More Details on ViMo Dataset

Explanation of Utilizing Synthetic Data. As described in Sec. 4, since some of the data are difficult to acquire from publicly available datasets, we adopt generative models to obtain some desired data to form part of our dataset. To guarantee the quality of the data, we leverage several cutting-edge generative models [1, 7, 20, 28, 49, 65, 83, 91] that are known for their superior performance, and manually screen the synthetic data. It is worth mentioning that synthetic data is customizable and low-cost. In addition, the videos generated by the models have a wide variety of characters and backgrounds. Another advantage of using synthetic data is that it does not violate the subject's right to likeness because none of the persons in the synthetic videos are real. Although there may be some distribution differences between real data and synthetic data, previous works (e.g., HDM [99] and InterTrack [100]) use synthetic data to train and then test on real data, which demonstrate the reliability of using such data.

Data Annotation. Ego-Humans [42] and Harmony4D [43] provide genuine video-motion pairs, so we don't need manual pairing. For Inter-X [102], we render the actor's motion into video and convert it to a realistic style using Runway Gen3 [77], then use the reactor's motion as GT reactions, which ensures reaction correctness. For videos without raw paired GT reactions, the annotation criteria consider two main dimensions: **motion quality** and **reaction plausibility**, to ensure physically plausible and contextually correct reaction motions. The entire procedure consists of labeling instruction preparation, pre-labeling trial, labeling instruction update, post-labeling, and checks by inspectors. After annotation, a user study rates the motion quality (4.51) and reaction plausibility (4.49) of GT pairs in our dataset (main paper Fig. 4).

Pose Representation. We use the same pose representation as in HumanML3D [26], which is over-parameterized, expressive, and neural network friendly [54], and has been widely adopted in recent works [12, 28, 71, 90, 115]. Each pose in the motion sequence is defined by $(\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^v, j^r, c^f)$, where $\dot{r}^a \in \mathbb{R}$ is root angular velocity along the Y-axis; $\dot{r}^x \in \mathbb{R}$ and $\dot{r}^z \in \mathbb{R}$ are root linear velocities on the XZ-plane; $r^y \in \mathbb{R}$ is root height; $j^p \in \mathbb{R}^{3j}, j^v \in \mathbb{R}^{3j}$ and $j^r \in \mathbb{R}^{6j}$ are the local joints positions, velocities and rotations in root space, with j indicating the number of joints; $c^f \in \mathbb{R}^4$ are binary features representing foot ground contacts derived from thresholding the velocities of the heel and toe joints.

Data Distribution. Fig. 9 and Fig. 10 provide some information on the distribution of the videos and the motion data, respectively.

8. More Details on Evaluation Metrics

In line with prior practices [26, 90], each experiment is conducted 20 times, and the reported values of the metrics indicate the mean along with a confidence interval of 95%.

FID. Frechet Inception Distance (FID) is adopted as the principal metric to evaluate the overall quality of the generation, which is calculated between the feature distribution of the generated motions and the feature distribution of the real motions. The feature extractor is from [26].

Diversity. Diversity measures the variability and richness of motions, which is calculated by averaging the Euclidean distances of 300 randomly sampled pairs of motions.

MultiModality. MultiModality measures the diversity of human motion generated from the same video. Specifically, it represents the average variance for a single video by computing the Euclidean distances of 10 sampled pairs of generated motions. For each video, we generate the motion 30 times.

Explanation of the Absence of Action Recognition Accuracy. [58, 103] use a pretrained model to classify the generated motions and calculate the action recognition accuracy. However, this does not apply to our task. In our setup, motions generated from videos in one category may belong to multiple categories, as long as they are plausible reactions to the videos. For example, the reactions to "*walking towards happily*" might be a wave, a handshake, or even a hug. Besides, the reactions to "*hitting*" might be to dodge, parry, or counterattack... Therefore, we do not use action recognition accuracy as one of our evaluation metrics.

Some other works [21, 59] of human reaction generation do not use it either. [58, 103] use specific pose representations, they train the classification model to obtain the feature extractor mainly for measuring FID, which is incidentally used to measure accuracy. In contrast, the motions in our ViMo dataset are provided with the same pose representation as in HumanML3D [26]. Thus, we can naturally employ the high-quality feature extractor from [26], which is trained on much more motion data than in our dataset.

In the context of human reaction generation, the accuracy is mainly used to measure reaction plausibility, i.e., whether the generated reaction is plausible for the input. However, as mentioned earlier, the accuracy is not applicable to our task. To compensate for the lack of reaction plausibility evaluation, we investigate it through visualization and user study (Sec. 5.2) following [21]. In the future, more

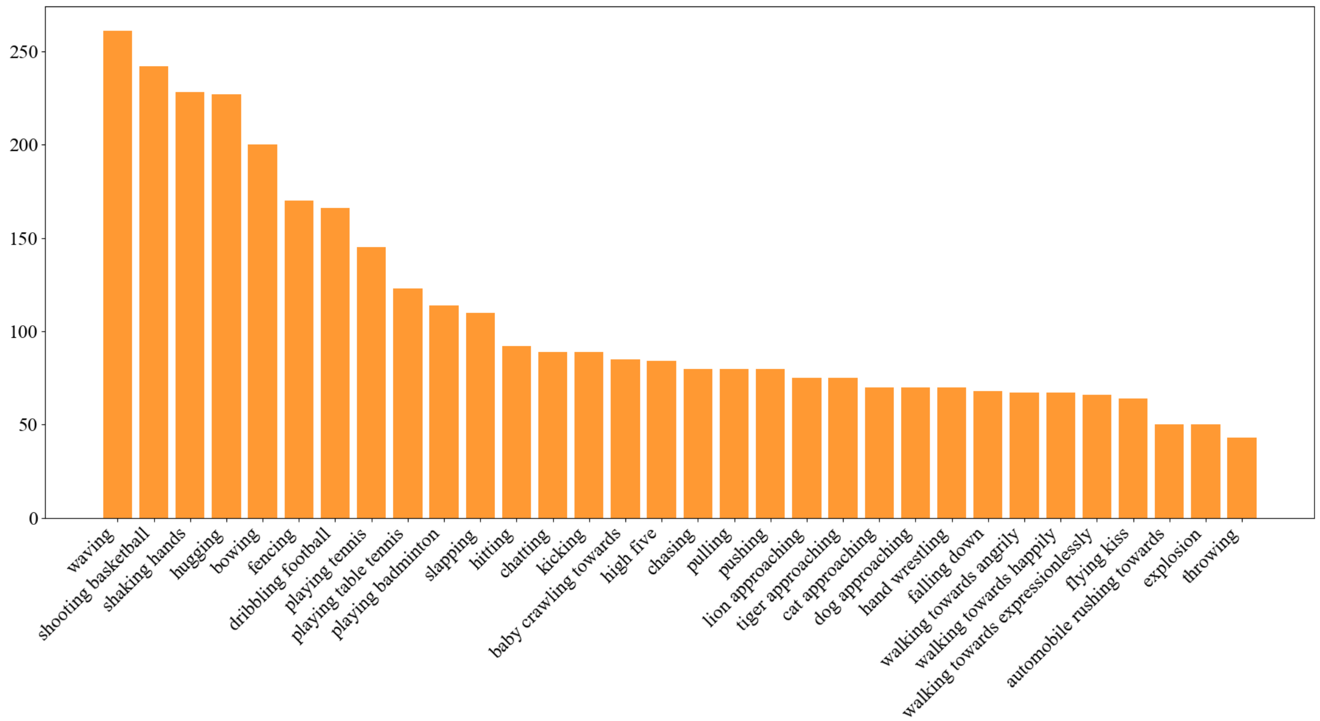

Figure 9. **Distribution of the video data in ViMo dataset.**

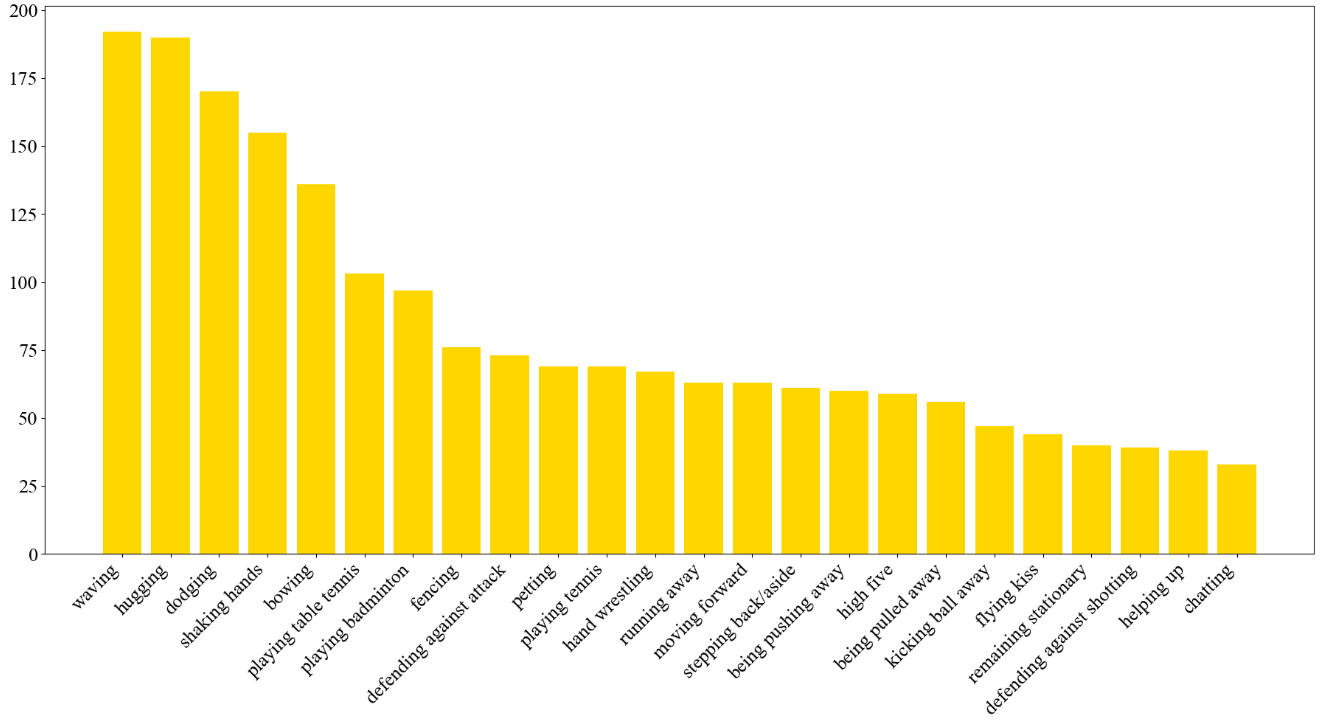

Figure 10. **Distribution of the motion data in ViMo dataset.**

powerful multimodal large language models (MLLMs) may also be employed to evaluate it.

Method	Reconstruction		Generation	
	FID↓	MPJPE (mm)↓	FID↓	Re.Plau.↑
VQ-VAE	$0.484^{\pm 0.012}$	38.7	$0.625^{\pm 0.036}$	3.66
RVQ-VAE	$\mathbf{0.133}^{\pm \mathbf{0.002}}$	**32.8**	$\mathbf{0.427}^{\pm \mathbf{0.014}}$	**3.82**

Table 3. **VQ-VAE v.s. RVQ-VAE.** ± indicates 95% confidence interval. **Bold** face indicates the best result.

9. More Details on Implementation

Our models are implemented using PyTorch. TC-CLIP [45] is initialized with the weights of CLIP [73] with ViT-B/16 and then pretrain on Kinetics-400 [41], a large-scale action recognition dataset with a total of 400 action classes and around 240k training videos. Therefore, the video encoder's action recognition capability is enhanced on the basis of its general recognition ability. This applies to our scenario, as most of the data in our ViMo dataset are about human-human interactions. Following [45, 96], 16 frames are sampled from each video before entering the video encoder, and all frames are resized to a uniform resolution of 224 × 224. Our RVQ-VAE has the same architecture as in MoMask [28] and the pose representation in ViMo is the same as in HumanML3D [26], so the RVQ-VAE can be initialized with the weights obtained by pretraining on HumanML3D. As for Transformers, we set the number of Transformer decoder units $\mathrm{N}_{\text{layers}} = 6$ in our models. Our models are trained with the AdamW [60] optimizer. The learning rate reaches 2e-4 after 20 and 250 iterations with a linear warm-up schedule for training RVQ-VAE and Transformers, respectively. All models are trained using 2 NVIDIA A40 GPUs.

10. More Experimental Results

Ablation Study. We ablate the RVQ-VAE choice, as reported in Tab. 3, indicating that RVQ-VAE effectively improves both reconstruction and generation quality. We also present experimental results with different number of Transformer decoder units $\mathrm{N}_{\text{layers}}$ in Tab. 4. We notice that HERO achieves the best overall performance when setting $\mathrm{N}_{\text{layers}} = 6$.

Different Broad Categories of Interactions. The quantitative results on different broad categories of interactions are reported in Tab. 5. Numerically, HERO performs well in the broad category of human-human interactions. The suboptimal performance of HERO in other broad categories may be due to limited data in these categories. It is also possible that these broad categories are harder for the model to learn. Nevertheless, from the visualization results shown in Fig. 5 of the main text, HERO can generate plausible reactions in all broad categories.

$\mathrm{N}_{\text{layers}}$	FID↓	Diversity→	MModality↑
	-	$7.954^{\pm 0.074}$ (R.)	-
4	$0.659^{\pm 0.014}$	$7.549^{\pm 0.051}$	$1.572^{\pm 0.059}$
5	$0.639^{\pm 0.018}$	$7.604^{\pm 0.075}$	$1.642^{\pm 0.046}$
6	$\mathbf{0.427}^{\pm \mathbf{0.014}}$	$\mathbf{7.801}^{\pm \mathbf{0.061}}$	$1.614^{\pm 0.040}$
7	$0.589^{\pm 0.015}$	$7.621^{\pm 0.068}$	$\mathbf{1.649}^{\pm \mathbf{0.051}}$
8	$0.683^{\pm 0.014}$	$7.499^{\pm 0.074}$	$1.550^{\pm 0.041}$

Table 4. **Ablation studies on the number of Transformer decoder units $\mathbf{N}_{\textbf{layers}}$.** → means the closer to the real motions the better. R. means real motions.

Category	FID↓	Diversity→	MModality↑
Scene-Human R.	-	$3.666^{\pm 0.111}$	-
Scene-Human	$3.726^{\pm 0.249}$	$3.676^{\pm 0.184}$	$1.916^{\pm 0.063}$
Animal-Human R.	-	$6.755^{\pm 0.387}$	-
Animal-Human	$1.809^{\pm 0.075}$	$6.695^{\pm 0.380}$	$1.362^{\pm 0.044}$
Human-Human R.	-	$7.807^{\pm 0.063}$	-
Human-Human	$0.421^{\pm 0.011}$	$7.737^{\pm 0.043}$	$1.454^{\pm 0.042}$

Table 5. **Evaluation results on different broad categories of interactions.**

Category	FID↓	Diversity→	MModality↑
Angrily R.	-	$3.411^{\pm 0.215}$	-
Angrily	$5.347^{\pm 0.430}$	$4.234^{\pm 0.241}$	$2.014^{\pm 0.064}$
Happily R.	-	$5.074^{\pm 0.290}$	-
Happily	$5.215^{\pm 0.627}$	$3.835^{\pm 0.216}$	$1.831^{\pm 0.056}$
Expressionlessly R.	-	$1.548^{\pm 0.082}$	-
Expressionlessly	$0.900^{\pm 0.146}$	$1.383^{\pm 0.123}$	$0.589^{\pm 0.039}$

Table 6. **Evaluation results on "*walking towards*" with different emotions.**

The Same Action with Different Emotions. We can find in Tab. 6 that the reaction generation of "*walking towards angrily*" and "*walking towards happily*" is much more difficult to learn than that of "*walking towards expressionlessly*". However, Fig. 8 of the main text shows that HERO is able to synthesize distinct plausible reactions according to different emotions, even if the actions in the videos are the same.

Generalization Ability. We train the models on the **Seen** set and evaluate them on the **Unseen** set. The quantitative comparisons reported in Tab. 7 show that HERO achieves the best FID score, and values of diversity and multimodality comparable to those of other methods. Some visualized cases can be found in Fig. 7 of the main text.

Training on More Pairs. In addition to ViMo-base mentioned in Sec. 5.1, we manually pair each video and two motions in the training set of ViMo-base to form twice as many training data as in the base dataset, that is, 5600 video-

Method	FID$\downarrow$	Diversity$\rightarrow$	MModality$\uparrow$
Real	-	$7.935^{\pm 0.061}$	-
BAMM [71]	$2.541^{\pm 0.079}$	$7.143^{\pm 0.082}$	$\mathbf{2.192}^{\pm\mathbf{0.062}}$
MoMask [28]	$2.477^{\pm 0.049}$	$\mathbf{7.151}^{\pm\mathbf{0.084}}$	$2.009^{\pm 0.064}$
HERO	$\mathbf{2.156}^{\pm\mathbf{0.054}}$	$7.095^{\pm 0.066}$	$2.093^{\pm 0.057}$

Table 7. **Quantitative evaluation on the Unseen set.**

Num. of T.P.	FID$\downarrow$	Diversity$\rightarrow$	MModality$\uparrow$
Real	-	$7.954^{\pm 0.074}$	-
2800	$0.427^{\pm 0.014}$	$7.801^{\pm 0.061}$	$1.614^{\pm 0.040}$
5600	$0.398^{\pm 0.012}$	$7.815^{\pm 0.069}$	$1.529^{\pm 0.058}$
8400	$0.376^{\pm 0.014}$	$7.833^{\pm 0.064}$	$1.422^{\pm 0.049}$

Table 8. **Evaluation results on different number of training pairs.** T.P. means training pairs.

motion pairs (ViMo-T2). Similarly, we also obtain 8400 video-motion pairs (ViMo-T3) for training, which is three times the number of training pairs in the base dataset. Note that ViMo-base, ViMo-T2, and ViMo-T3 share the same test set.

As shown in Tab. 8, training on more data pairs leads to consistent improvements in FID and diversity. The decay in multimodality implies a decrease in the diversity of reactions generated by the model to a single video, but perhaps indirectly indicates that the model is becoming more confident in the generated motions. Although multimodality is undoubtedly important, [28] emphasizes its role as a secondary metric that should be evaluated alongside primary performance metrics like FID. Theoretically, we can pair more than 100 k video-motion pairs. The lack of training pairs can be alleviated to some extent by pairing more data.

11. Discussion

The Modular MLLM-based Pipeline. Another implementation to deal with our task might be to utilize a combination of a multimodal large language model (MLLM) and a text-to-motion generative model. Specifically, the MLLM gives a textual description of the corresponding reaction based on the input video, and the text-to-motion generative model synthesizes the motion based on this description. However, after extensive testing of MLLMs [11, 40, 55, 63], we observe that despite being able to describe the video content well, these MLLMs struggle to output texts that accurately describe reactive motions in detail. In addition, MLLMs tend to generate redundant information (which reduces the quality of the text used for motion generation), despite being asked not to. These issues prevent them from outputting textual descriptions like those written by humans in HumanML3D, making it difficult for them to perform as an ideal preceding module for the text-to-motion generative model. In contrast, our framework adopts an end-to-end manner that does not require manually crafted text prompts and is much more efficient.

Future Work. Although as discussed above, an interesting direction is to utilize the multi-turn conversation capability of MLLMs to instruct the model for motion reasoning, generation, and editing [38]. In addition, modeling the details of hand motions to synthesize more expressive reactions is worth exploring [61]. Moreover, generating a highly diverse set of reactions from the same input video, as well as producing motions that are aligned with the videos in both spatial and temporal dimensions, remains challenging.

Broader Impacts. According to Rosenblum [76], vision is the dominant sense in human perception, accounting for approximately 80%–90% of sensory input. Similarly, Jensen [36] estimates that around 83% of information acquired by humans comes from vision. Compared to texts, videos contain denser information. Our work is no longer limited to enabling machines to passively follow instructions to generate motion. Instead, we aim to empower them to proactively explore how to interact with the world through visual signals. We believe that our work has the potential for a wide range of applications, especially in the emerging fields of AR/VR and embodied intelligence.