HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models

Supplementary Material

A. Details of HOI-decomposed VQ

In this section, we provide details of our proposed HOI-decomposed VQ. Using the inputs from the hand and object encoders, the HOI-decomposed VQ gradually learns the residual $\mathbf{z}_{res}$ from the learned subject and finally outputs the predicted HOI latent feature $\hat{z}$.

Algorithm 1: HOI-decomposed VQ

Input: z_o, z_l, z_r: the output of the encoders, vector quantizers Q_i for $i \in \{o, l, r\}$, object/hand codebooks C_o, C_h

Output: The quantized HOI feature $\hat{z_N}$

$z \leftarrow z_o + z_l + z_r$

$\hat{z} \leftarrow 0.0$

$z_{res} \leftarrow z$

for i **do**

  if i *is* $\{o\}$ **then**

    $q_i \leftarrow Q_i(\mathrm{C_o}, z_{res})$

  else

    $q_i \leftarrow Q_i(\mathrm{C_h}, z_{res})$

  $z \mathrel{+}= q_i$

  $z_{res} \mathrel{-}= q_i$

return $\hat{z}$

B. Additional Qualitative Results

In this supplementary material, we present additional qualitative results for ablation studies and comparison with other methods, including Text2HOI [3], MotionGPT [4] and T2MGPT [6]. Please refer to the attached video for these qualitative results.

C. Implementation Details

We included additional details about the hyper-parameters and evaluation.

Hyper-parameters. We set the $\lambda = 0.2, \beta = 0.5, \gamma = 1$ and $\alpha = 0.5$ in our experiments.

Evaluation metric. We follow [2] to train the HOI text matching network for feature extraction. Similarly, the motion encoder and text encoder are two bidirectional GRUs. We train the feature extractor for 65 epochs in our combined dataset for evaluation.

D. Dataset Details

We provide additional details about the dataset we used to train and evaluate HOIGPT. Our combined dataset includes two popular HOI datasets: GRAB [5] and Arctic [1], comprising a total of 6.1k HOI sequences with 65 unique objects (*Box, Capsule Machine, Espresso Machine, Ketchup, Laptop, Microwave, Mixer, Notebook, Phone, Scissors, Waffle Iron, Airplane, Alarm Clock, Apple, Banana, Binoculars, Body, Bowl, Camera, Coffee Mug, Cube Large, Cube Medium, Cube Small, Cup, Cylinder Large, Cylinder Medium, Cylinder Small, Doorknob, Duck, Elephant, Eyeglasses, Flashlight, Flute, Frying Pan, Game Controller, Hammer, Hand, Headphones, Knife, Lightbulb, Mouse, Mug, Phone, Piggy Bank, Pyramid Large, Pyramid Medium, Pyramid Small, Rubber Duck, Scissors, Sphere Large, Sphere Medium, Sphere Small, Stamp, Stanford Bunny, Stapler, Table, Teapot, Toothbrush, Toothpaste, Torus Large, Torus Medium, Torus Small, Train, Watch, Water Bottle, Wine Glass, Wristwatch*). From this dataset, 500 samples, including unseen HOI sequences and objects, were selected for testing.

References

[1] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[2] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[3] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 1

[4] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[5] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 1

[6] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1