

Frequency-Enhanced Wavelet-based Transformer in Imitation Learning for Humanoid Robot

Jiaxin Huang^{1*}, Hanyu Liu^{1*}, Jian Shen¹, Jixiang Li¹, Jiang Han¹, Baishu Wan¹, Pan Li¹,
Yilin Zheng¹, Jiayi Wen¹, Yunsheng Ma¹, Jiejun Hou¹, Dachuan Xiao¹, Zhigong Song^{1†}

*These authors contributed equally to this work.

†Corresponding author: song_jnu@jiangnan.edu.cn.

¹Jiangsu Provincial Key Laboratory of Food Advanced Manufacturing Equipment Technology, School of Mechanical Engineering, Jiangnan University, Wuxi 214122, China.

Project Website: <https://humanoid-black-knight.github.io/>.

APPENDIX

A. Comparison of Parameters and FLOPs

Table A1 shows that, based on the ACT Policy baseline and Backbone (ResNet18), the FE-EMA module reduces the computational complexity (FLOPs) compared to the EMA module. Where the Simulation Tasks (ACT baseline) are *Cube Transfer* and *Bimanual Insertion*, respectively.

Table A1. Comparison of Parameters and FLOPs for One Epoch.

Method	Backbone	Simulation Tasks	
		#.Param.	FLOPs
Baseline (ACT)	ResNet18	60.7566 M	37570.51 M
EMA + ACT		60.7592 M	37616.59 M
FE-EMA + ACT (ours)		60.7621 M	37596.91 M

B. Training Details

We adopt the hyperparameters reported in the ACT Policy paper, with the sole modification of reducing the batch size from 8 to 4. The complete hyperparameter settings are summarized in **Table A2**. We deploy our policy with inference on a desktop with an NVIDIA RTX 3060 GPU.

Table A2. Hyperparameters of the FEWT (based on ACT Policy).

Hyperparameter	FEWT
Learning rate	1e-5
Batch size	4
# encoder layers	4
# decoder layers	7
Feedforward dimension	3200
Hidden dimension	512
# heads	8
Chunk size	100
KL weights	10

For the two simulation tasks, *Cube Transfer* and *Bimanual Insertion*, the dynamic weight α is consistently maintained at approximately 0.5 during training (see Fig. A1). This suggests that the time-domain and frequency-domain feature weights are well balanced during model training. Dynamically fusing the time-domain and frequency-domain features allows effective capture of information across different scales.

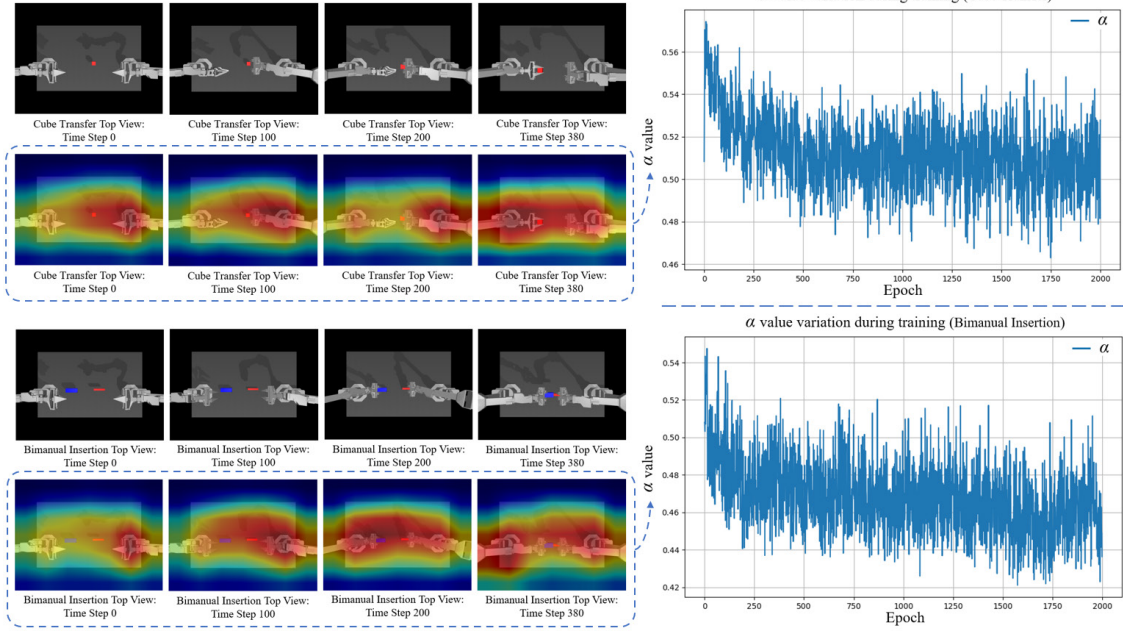


Fig. A1: Visualization of the dynamic weight α in the FE-EMA module (training phase, effect of 2000 epoch under different simulation tasks).

C. The Construction of Humanoid Black Knight

The construction of *Humanoid Black Knight* (HBK) mainly consists of two 7-dof (arm + gripper) robotic arms (*ViperX-300*), a two-wheeled differential mobility chassis (*Diablo*), three RGB cameras (Logitech C922x webcams), and an inertial measurement unit (IMU sensor). The purpose of the chassis linkage restraint parts is to prevent skidding during chassis steering movements (See Fig. A2).

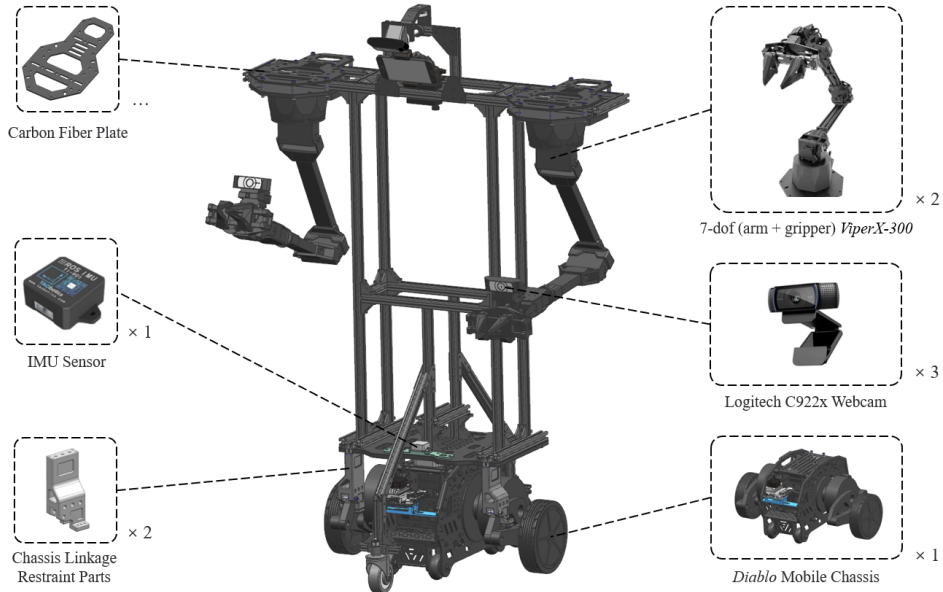


Fig. A2: The design of humanoid robot (*Humanoid Black Knight*, HBK).