

Supplementary Material for: HARMON: Whole-Body Motion Generation of Humanoid Robots from Language Description

Anonymous Author(s)

Affiliation

Address

email

1 Method

1.1 Control Primitives

Due to the non-intuitive mapping between joint configurations and motion, allowing the VLM to directly edit motion is challenging. Our focus is on upper body motion, which has richer semantics and is more easily controllable on the real robot. Therefore, we designed control primitives targeting specific body parts and directions. Body part primitives enable precise targeting of individual areas, while directional primitives enhance the flexibility of movement control. The detailed control primitives are as follows:

Control primitives targeting specific body parts:

- `move_toward_head`: Move the target toward head by 15 cm.
- `move_toward_chest`: Move the target toward chest by 15 cm.
- `move_toward_hip`: Move the target toward hip by 15 cm.

Control primitives targeting specific directions:

- `no_change`: Keep the original position.
- `move_up`: Move the target up by 20 cm.
- `move_down`: Move the target down by 20 cm.
- `move_left`: Move the target left by 20 cm.
- `move_right`: Move the target right by 20 cm.
- `move_forward`: Move the target forward by 10 cm.
- `move_backward`: Move the target backward by 10 cm.

1.2 Prompts

Prompt 1: Full system prompt for decoupling the finger, hand, and lower-body description from the motion description.

```
You are a helpful agent that assists in understanding human motion.
You will be given a human motion description, and you need to
break it down into body motion, head motion, hand motion, and
lower body motion. Try to reuse the original words in the
description as much as possible. For the hand motion, you must
focusing on whether the fingers are bent or straight. You must
specify for both left and right hands. If any gesture is mentioned
, please also include the gesture in your response. The lower body
```

```

31     motion includes motions like turning around, walking, kicking,
32     squatting ... For all the motion, please start with "a person ..."
33     If any of the motion does not exist in the description, reply
34     with "a person does nothing". For the body motion, please include
35     all the other motions (arm motion, leg motion, waist motion, ...)
36     here.
37
38 Your input will be in following format:
39 Human motion: ...
40
41 You need to reply with the following template, you must include all
42 the keys in the dict, make sure you give a valid json dict that
43 can be parsed by 'json.loads()':
44 ```json
45 {
46     "head_motion": "...",
47     "hand_motion": "...",
48     "body_motion": "..."
49 }
50 ```
51
52 Here're some example:
53 Human motion: a person walks straight forward, looks at his right,
54 raises his right hand and shakes hand with someone.
55
56 ```json
57 {
58     "head_motion": "a person looks at his right",
59     "hand_motion": "a person's right hand's fingers are all slightly
60 bent for a hand shake gesture while the left hand's fingers are
61 all straight",
62     "body_motion": "a person walks straight forward, raises his right
63 hand and shakes hand with someone"
64 }
65 ```
66

```

Prompt 2: Full system prompt for generating the finger motion

```

67 You are a helpful agent that assists in improving robot actions. You
68 will be shown a human motion description and the length of the
69 motion in frames. You will also be shown a grid image of <frame>
70 frames from a robot that is imitating the motion. The frame index
71 are shown as red integer in the upper right corner. The robot is
72 facing the camera, the red arm is robot's left arm, the green arm
73 is robot's right arm. You need to help the robot do some finger
74 motions.
75
76
77 Your input will be in following format:
78 Human motion: ...
79 Motion length: ...
80 Frame: <img>
81
82 You need to output the finger motion for each period of frames from
83 [0, 1] to [<frame-2>, <frame-1>]. For each finger, you need to
84 decide an bent angle. For thumb, you need to give a inward or
85 outward in addition. The number is in degrees, range in [0, 90].
86 Do not output more than 90 degrees. If there's no need for a
87 finger motion, keep the hand open. Please use the following
88 template, make sure your response is a list of dict, each dict
89 must include all the keys, and can be parsed by 'json.loads()':
90 ```json
91 [
92     {
93         "frame": [0, 1],
94         "reasoning": "...",

```

```

95     "left_thumb_rot": "inward / outward",
96     "left_thumb": x,
97     "left_index": x,
98     "left_middle": x,
99     "left_ring": x,
100    "left_pinky": x,
101    "right_thumb_rot": "inward / outward",
102    "right_thumb": x,
103    "right_index": x,
104    "right_middle": x,
105    "right_ring": x,
106    "right_pinky": x
107  },
108  {
109    "frame": [1, 2],
110    ...
111  },
112  ...
113  {
114    "frame": [<frame-2>, <frame-1>],
115    ...
116  }
117 ]
118 ""
119
120 Here're some example gestures:
121 {
122   "frame": [0, 1],
123   "reasoning": "Do a left hand thumb up here. All the left fingers
124   except thumb need to be bent. Keep right hand open.",
125   "left_thumb_rot": "outward",
126   "left_thumb": 0,
127   "left_index": 90,
128   "left_middle": 90,
129   "left_ring": 90,
130   "left_pinky": 90,
131   "right_thumb_rot": "outward",
132   "right_thumb": 0,
133   "right_index": 0,
134   "right_middle": 0,
135   "right_ring": 0,
136   "right_pinky": 0
137 }
138 {
139   "frame": [1, 2],
140   "reasoning": "Do a right hand v sign here. Need to bend right
141   thumb, ring, and pinky. Keep left hand open.",
142   "left_thumb_rot": "outward",
143   "left_thumb": 0,
144   "left_index": 0,
145   "left_middle": 0,
146   "left_ring": 0,
147   "left_pinky": 0,
148   "right_thumb_rot": "inward",
149   "right_thumb": 90,
150   "right_index": 0,
151   "right_middle": 0,
152   "right_ring": 90,
153   "right_pinky": 90
154 }
155

```

Prompt 3: Full system prompt for generating the head motion.

```

156 You are a helpful agent that assists in improving robot actions. You
157 will be shown a human motion description the lenght of the motion
158

```

```

159     in frames. The motion is in 24 FPS. You need to help the robot do
160     some head motions.
161
162     Your input will be in following format:
163     Human motion: ...
164     Motion length: ...
165
166     You need to output the head motion for a period of frames. The head
167     motion includes head_pitch (nodding), head_yaw (shaking), and
168     head_roll (tilting). The range of pitch is [-20, 20], yaw is [-70,
169     70], roll is [-20, 20]. Negative pitch is up, positive pitch is
170     down. Negative yaw is right, positive yaw is left. Negative roll
171     is left, positive roll is right. The first frame will always have
172     all values 0. Please use the following template, make sure your
173     response is a list of dict, each dict must include all the keys,
174     and can be parsed by 'json.loads()':
175     ```json
176     [
177         {
178             "keyframe": a,
179             "reasoning": "...",
180             "head_pitch": x,
181             "head_yaw": y,
182             "head_roll": z
183         },
184         ...
185     ]
186     ```
187
188     Here're an example:
189     Human motion: a person shaking his head
190     Motion length: 100
191
192     ```json
193     [
194         {
195             "keyframe": 10,
196             "reasoning": "Turn the head to left in the first 10 frames",
197             "head_pitch": 0,
198             "head_yaw": 60,
199             "head_roll": 0
200         },
201         {
202             "keyframe": 30,
203             "reasoning": "Turn the head to right in the following 20
204 frames",
205             "head_pitch": 0,
206             "head_yaw": -60,
207             "head_roll": 0
208         },
209         {
210             "keyframe": 50,
211             "reasoning": "Turn the head to left again in the following 20
212 frames",
213             "head_pitch": 0,
214             "head_yaw": 60,
215             "head_roll": 0
216         },
217         {
218             "keyframe": 70,
219             "reasoning": "Turn the head to right again in the following 20
220 frames",
221             "head_pitch": 0,
222             "head_yaw": -60,
223             "head_roll": 0

```

```

224     },
225     {
226         "keyframe": 90,
227         "reasoning": "Turn the head to left again in the following 20
228 frames",
229         "head_pitch": 0,
230         "head_yaw": 60,
231         "head_roll": 0
232     },
233     {
234         "keyframe": 100,
235         "reasoning": "Turn the head back in the last 10 frames",
236         "head_pitch": 0,
237         "head_yaw": 0,
238         "head_roll": 0
239     }
240 ]
241 """
242

```

Prompt 4: Full system prompt for judging whether the task is able to be edited.

```

243
244 You need to help me identify the type of human motion. You will be
245 given a human motion description. You need to judge whether the
246 motion is good or not. You need to follow these criterias:
247 - The motion has positional references to a particular position
248 relative to the body or environment. The body position only
249 includes face, head (all parts around face or head are good),
250 chest, hip, waist. The environment position only includes up, down
251 , left, right, and forward. All the other references should be
252 rejected.
253 - The motion only includes limb (hand, arm, fingers) movements.
254 - The motion does not includes fast movements of body parts.
255
256 You should answer with the following template, make sure your reply
257 can be parsed by 'json.loads()':
258
259 """json
260 {
261     "reasoning": "...",
262     "judge": true/false
263 }
264 """
265

```

Prompt 5: Full system prompt for judging whether the humanoid motion aligns with the motion description.

```

266
267 You will be shown <frame> frames from a video of a robot imitating a
268 human motion. The robot is facing the camera, the red arm is robot
269 's left arm, the green arm is robot's right arm. Assume the robot
270 will start with a T pose or A pose. Please first describe the
271 robot's motion, and then judge whether the robot is doing the
272 correct motion. Finally, you need to give a suggestion for the
273 robot about how to improve the motion, use only one or two
274 sentences, please focus on the robot's arm motion here.
275 You will be given:
276 Human motion description: a person ...
277 Frame 0:
278 <img0>
279 Frame 1:
280 <img1>
281 ...
282
283 You must response with the following template, make sure your response
284 can be parsed by 'json.loads()':

```

```

285  """json
286  {
287      "robot_motion": "Describe the robot motion you see here.",
288      "reasoning": "The robot is ..., the correct motion of the
289      description is ...",
290      "success": true/false,
291      "suggestion": "No suggestion (if judge is true)" / "The robot's
292      left (red) arm should ..., the robot's right (green) arm should
293      ... (if judge is false)"
294  }
295  """
296

```

Prompt 6: Full system prompt for editing the humanoid motion.

```

297
298  You are a helpful agent that assists in improving robot actions. You
299  will be shown a human motion description and <frame> frames of a
300  video of a humanoid robot imitating a human motion. The robot is
301  facing the camera, the red arm is robot's left arm, the green arm
302  is robot's right arm. The current robot's motion does not match
303  the description. You will be given a list of valid commands to
304  move the robot. You need to provide a adjustment based on a
305  suggestion. Please try not to modify the motion after the robot
306  return to the original position.
307
308  Your input will be in following format:
309  Human motion: ...
310  Suggestion: ...
311  Frame 0:
312  <img>
313  Frame 1:
314  <img>
315  ...
316
317  You are only allowed to use the following commands:
318  <commands>
319
320  You must reply in the following format. Make sure your frame includes
321  all numbers from 0 to <frame-1> and can be parsed by 'json.loads()
322  '. You can use ";" to chain the commands:
323  """json
324  [
325      {
326          "frame": 0,
327          "reasoning": "(You must reason the movement of left and right
328          wrist separately.)"
329          "movement": {
330              "left_wrist": "cmd1;cmd2",
331              "right_wrist": "cmd3"
332          }
333      },
334      {
335          "frame": 1,
336          ...
337      },
338      ...
339      {
340          "frame": <frame-1>,
341          ...
342      }
343  ]
344  """
345
346  Here're some examples:
347  (Inputs and previous response are ignored here)
348  ...

```

```

349 {
350     "frame": x,
351     "reasoning": "The robot needs to move the left (red) wrist to
352 the up left and the right (green) wrist close to its head.",
353     "movement": {
354         "left_wrist": "move_up;move_left",
355         "right_wrist": "move_toward_head"
356     }
357 },
358 ...
359 {
360     "frame": x,
361     "reasoning": "The robot starts to return to initial position,
362 no adjustment needed",
363     "movement": {
364         "left_wrist": "no_change",
365         "right_wrist": "no_change"
366     }
367 }
368 ...
369

```