Grounding Hierarchical Vision-Language-Action Models Through Explicit Language-Action Alignment

Theodor Wulff* Federico Tavella Rahul Singh Maharjan Manith Adikari
Angelo Cangelosi

The University of Manchester
Manchester, United Kingdom

A. Prompt

We use a conversational structure to prompt the high-level VLM for low-level instructions. Figure 1 depicts the corresponding prompt template. The answer not provided during inference.

```
System: You are controlling a
robotic agent. Your task is to
<high-level instruction>.
User: What should the robot do next?
Answer: <low-level instruction>
```

Figure 1. Prompt template used for robotic agent instructions.

B. Data Augmentation

The data augmentation techniques we applied during training are listed in Table 1.

Modality	Augmentation	Probability
Images	Brightness	0.5
	Contrast	0.5
	Saturation	0.5
	Crop and resize	0.6
	Vertical translation	0.4
	Horizontal translation	0.4
	Scale (zoom in/out)	0.3
Actions	Noise	0.7

Table 1. Data augmentation techniques and their application probabilities, grouped by modality.

C. Hyperparameters

Table 2 lists the hyperparameters we applied to during training on the different model variants.

General	
Gradient norm clipping:	1.0
Action cut-off threshold:	0.1
High-level VLM (fine-tuning)	
Steps:	1,500
Learning rate:	10^{-5}
Effective Batch size:	64
Optimizer:	AdamW
Horizon:	8
Low-level VLA (fine-tuning)	
Steps:	15,000
Learning rate:	10^{-5}
Effective Batch size:	64
Optimizer:	AdamW
Horizon:	8
GPLA	
Steps:	100
Learning rate:	10^{-7}
Batch size:	64
Optimizer:	AdamW
Horizon:	8
Action-Conditioned Grounding Model	
Steps:	50,000
Learning rate:	10^{-4}
Effective batch size:	256
Optimizer:	Adam
Horizon:	8
Initial logit scale factor:	0.1
Label smoothing:	None
Diversity weight:	0.01
Model dimension:	64
N_FiLM layers:	4

Table 2. Hyperparameters grouped by model variant.

*Corresponding author: theodor.wulff@manchester.ac.uk