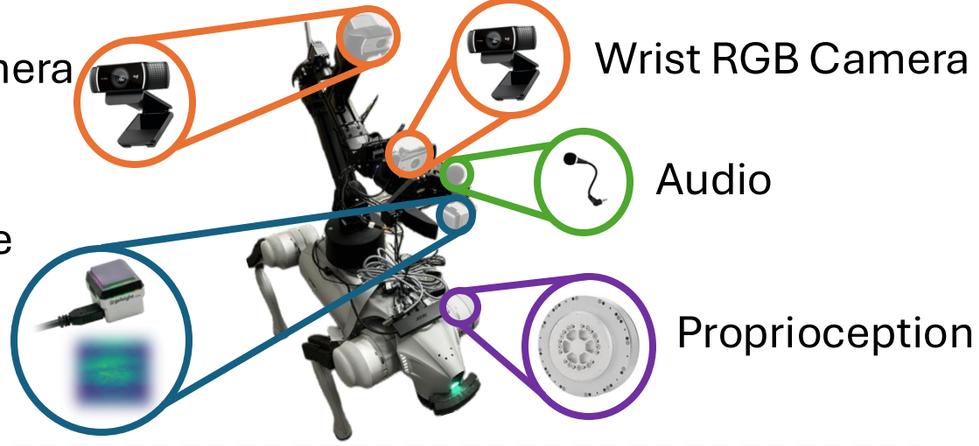
Unitree Go2 Robot
With D1 Arm



PRISM

RGB | Depth | Tactile | Proprioception | Text | Audio

Temporal Multi-Sensory Encoder

Context Tokens
$(B, T_{obs}, 1024)$

Query Tokens
$(16, 512)$ Learned
$T_{pred} = 16$

Noise
$(B, 64) \rightarrow (B, 16, 512)$
Projection + Expand

Pos Encoding
$(16, 512)$ Sinusoidal
Temporal ordering

Performer Transformer

Action Output

T 16 | T 16 | T 16 | T 16

Output
Linear $(512 \rightarrow 7)$

Actions
$(B, 16, 7)$

Performer Block (N=6)

Input

Self-Attention
FAVOR+ Linear O(L)

Cross-Attention
Context → Queries

Feed-Forward
512→2048→512

LayerNorm (x3)

Output

FAVOR+: 256 features,
Kernal, Bidirectional

RS-IMLE Loss
Batch-global ϵ-rejection
Robust distance EMA ϵ calibration

Top K Loss
Diversity
Weight: 0.02
Huber loss

Top Loss
$L_{hard} + \lambda L_{soft}$
$\lambda = 0.02$
End-to-end

Training Objectives