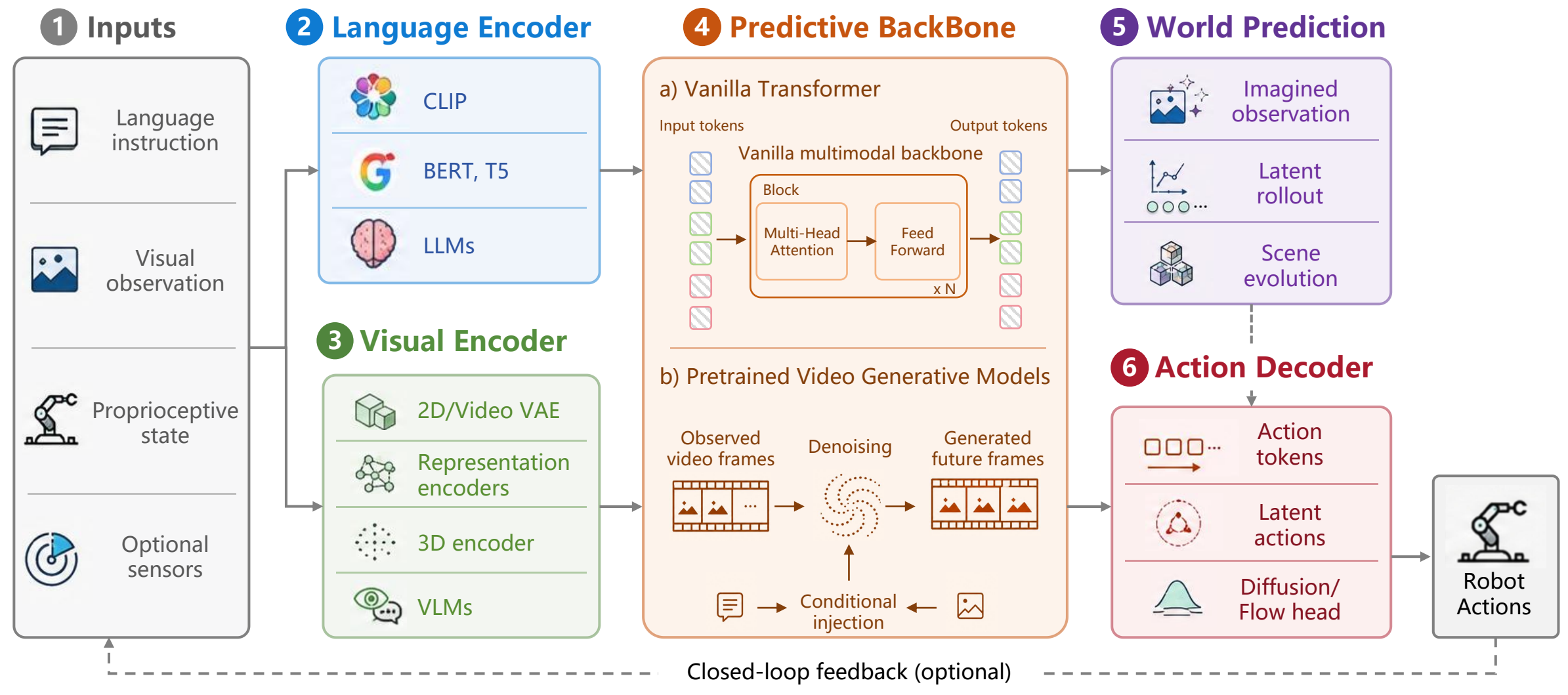
1 Inputs
Language instruction
Visual observation
Proprioceptive state
Optional sensors
2 Language Encoder
CLIP
BERT, T5
LLMs
3 Visual Encoder
2D/Video VAE
Representation encoders
3D encoder
VLMs
4 Predictive BackBone
a) Vanilla Transformer
Input tokens
Output tokens
Vanilla multimodal backbone
Block
Multi-Head Attention
Feed Forward
x N
b) Pretrained Video Generative Models
Observed video frames
Denoising
Generated future frames
Conditional injection
5 World Prediction
Imagined observation
Latent rollout
Scene evolution
6 Action Decoder
Action tokens
Latent actions
Diffusion/ Flow head
Robot Actions
Closed-loop feedback (optional)