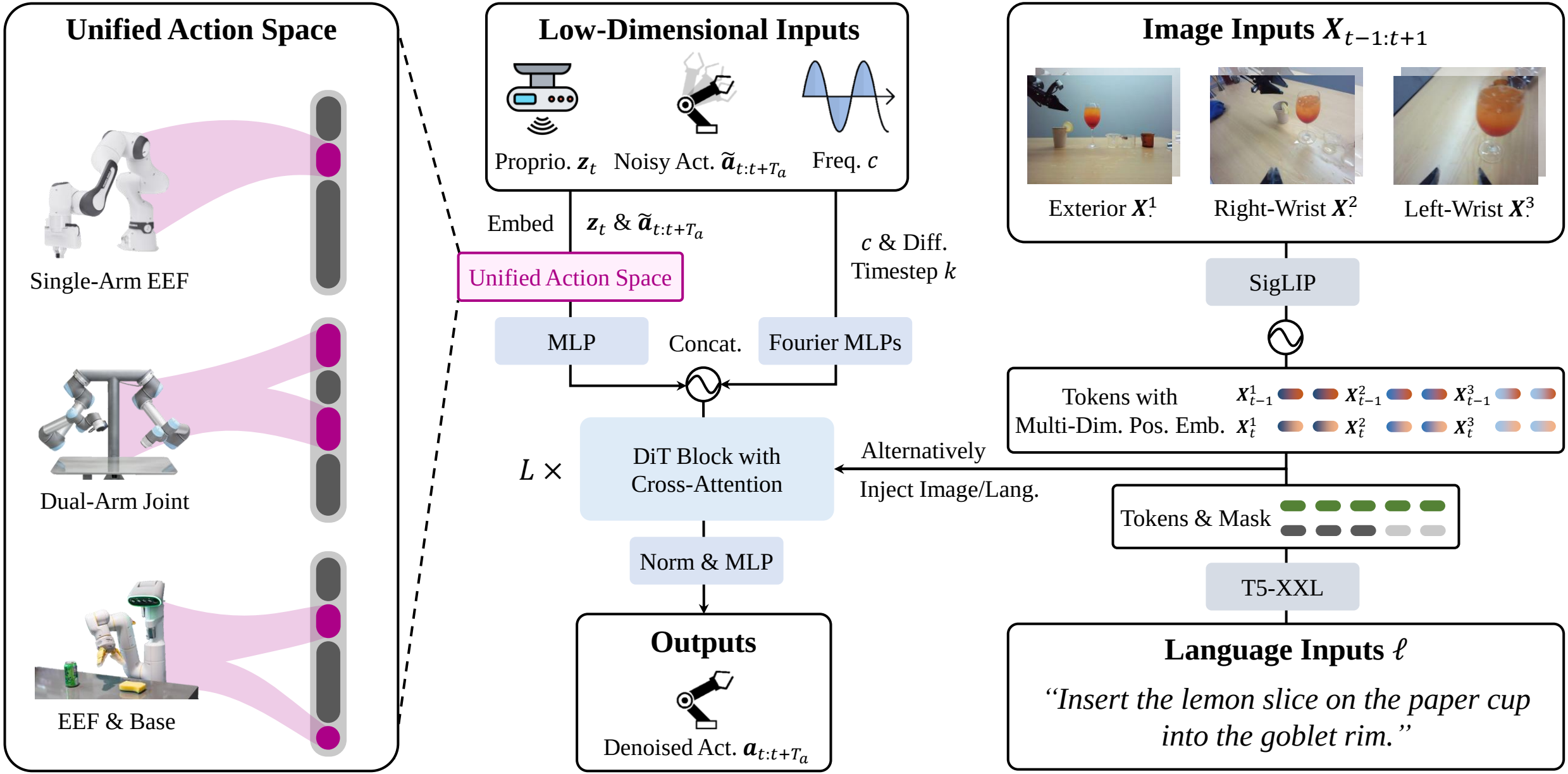
Unified Action Space
Single-Arm EEF
Dual-Arm Joint
EEF & Base
Low-Dimensional Inputs
Proprio. z_t
Noisy Act. $\tilde{a}_{t:t+T_a}$
Freq. c
Embed
z_t & $\tilde{a}_{t:t+T_a}$
Unified Action Space
MLP
Concat.
Fourier MLPs
c & Diff. Timestep k
$L\times$
DiT Block with Cross-Attention
Norm & MLP
Outputs
Denoised Act. $a_{t:t+T_a}$
Alternatively
Inject Image/Lang.
Image Inputs $X_{t-1:t+1}$
Exterior X^1
Right-Wrist X^2
Left-Wrist X^3
SigLIP
Tokens with Multi-Dim. Pos. Emb.
X^1_{t-1} X^2_{t-1} X^3_{t-1}
X^1_t X^2_t X^3_t
Tokens & Mask
T5-XXL
Language Inputs ℓ
"Insert the lemon slice on the paper cup into the goblet rim."