Visual Grounding Module

Grounding Response

Box $\mathbf{B}$ Mask $\mathbf{M}$

Grounding model s_{ϕ_g}

Projection $\mathbf{W}_g$ $\mathbf{X}_g$

Language Response

$\mathbf{X}_a$

f_ϕ Language Model

$\mathbf{H}_v$
$\mathbf{Z}_v$ Projection $\mathbf{W}$
Vision Encoder
$\mathbf{X}_v$
Image

$\mathbf{H}_q$
$\mathbf{X}_q$
Language Instruction

$\mathbf{H}_p$ Projection $\mathbf{W}_p$
$\mathbf{Z}_p$ Prompt Encoder h_{ϕ_p}
$\mathbf{X}_p$ Visual prompt

Visual Interaction Module