

Language response O_a



Large Language Model (LLM)



image1

text1

image2

text2



Multi-modal Token

image

token E_v



text

token E_t



Vision-language
Projector (P)

Vision
Encoder (E)

Tokenizer &
embedding



X_v



X_t

User: Output the
color and number
of each box.