Response

A: `<html> <head> ... </html>`

Large Language Model

Vision Output   Latent Output

Feed Forward Network

x L   Self-Attention

Layer Norm   Layer Norm

Learned Positional Encoding

Vision Input   Latent Input

Visual Tokens   Latent Tokens

Q: What is the website's HTML code?

Instruction

Our Projector

Vision Encoder   Latent Queries

Websites   **Photos**   **Documents**