

## (a) Rendering Method

### CoT Plain Text

Weng earns  $12/60 = \$\ll 12/60=0.2 \gg 0.2$  per minute. Working 50 minutes, she earned  $0.2 \times 50 = \$\ll 0.2*50=10 \gg 10$ .

Rendered Image

Rendering



Weng earns  $12/60 = \$\ll 12/60=0.2 \gg 0.2$  per minute. Working 50 minutes, she earned  $0.2 \times 50 = \$\ll 0.2*50=10 \gg 10$ .

Vision Encoder

<vision embedding>

Weng earns \$12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

Question Text

<question>

Visual Projection Head

<answer>

10

Answer

$$\mathcal{L}_{MSE}$$

$$\mathcal{L}_{CE}$$

LLM Backbone

<|img\_begin|>

<latent reasoning>

<|img\_end|>

## (b) Latent Reasoning Method