

(a) Rendering Method

CoT Plain Text

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

Rendering

Font
Size

Text
Color

Image
Height

...

Rendered Image

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

Vision Encoder ❄️

<vision embedding>

\mathcal{L}_{MSE}

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

<question>

Visual Projection Head

<answer>

\mathcal{L}_{CE}

10 Answer

Question Text

LLM Backbone

<|img_begin|>

<latent reasoning>

<|img_end|>

(b) Latent Reasoning Method