Training Stage I

Weng earns 12/60 = 0.2...

Encoding → <vision embedding>

Visual Projection Head 🔥

$\mathcal{L}_{MSE}$

$\mathcal{L}_{CE}$

Question → LLM Backbone ❄️ → Answer

<latent embedding>|<img_end|>
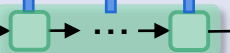
Training Stage II

<projected embedding>

Visual Projection Head ❄️

$\mathcal{L}_{CE}$

Question → LLM Backbone 🔥 → Answer

<latent embedding>|<img_end|>