

Question: What is **the title** of the book?

Prompt x



Chosen m_w

... "Trail of the Dead by Jon." ... The specific book title can be determined ... if the book's cover ... are visible ...

Chosen y_w

... "Trail of the Dead by Jon." It is **evident from the book being displayed with a distinctive feature like a blue ruler** shaped like a ruler ...

Rejected y_l

Text-based Preference Pairs

Candidate Set



Randomly perturb chosen image n times

Rejected image Selection
(CLIP Score with m_w)

Question: What is **the title** of the book?

Prompt x



Rejected m_l



Chosen m_w

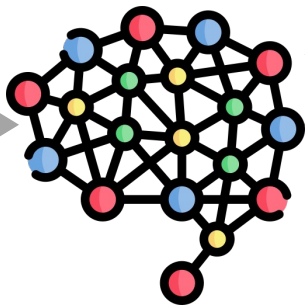
... "Trail of the Dead by Jon." ... The specific book title can be determined ... if the book's cover ... are visible ...

Chosen y_w

Perception-Enhanced
Preference Pairs

$$\mathcal{M}_{\theta}^r(x, m_w, y_w, y_l) > \gamma_r$$

Response quality
Preference
Optimization



$$\mathcal{M}_{\theta}^m(x, m_w, m_l, y_w) > \gamma_m$$

Visual context
Preference
Optimization