

*Stage 1: Pre-training with Textual Preference Data*

Bridging Task Gap

*Stage 2: Fine-tuning with Image Caption-based Preference Data*

Bridging Modality Gap

*Stage 3: Fine-tuning with Visual Preference Data*

Instruction: Can you determine the missing number in the sequence: 2, 6, 14, 30, 62, --?

Chosen Response: The missing number in the sequence is 126.

Rejected Response: The sequence is 6.

Instruction: Describe the image in detail.

Image Caption: There are four mangosteens, one of which has been cut open.

Chosen Response: There are several mangosteens on a wooden table.

Rejected Response: This is a dark fruit, possibly a variety of pitaya.



Instruction: Describe the image in detail.

Chosen Response: There are several mangosteens on a wooden table.

Rejected Response: This is a dark fruit, possibly a variety of pitaya.