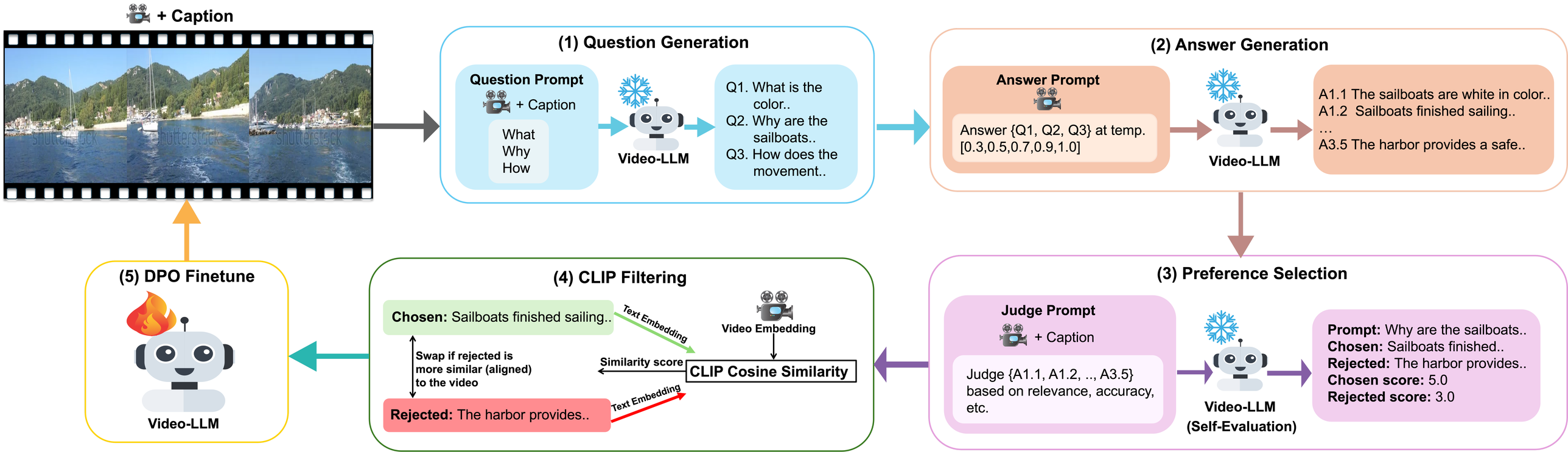
+ Caption
(1) Question Generation
Question Prompt
+ Caption
What
Why
How
Video-LLM
Q1. What is the color..
Q2. Why are the sailboats..
Q3. How does the movement..
(2) Answer Generation
Answer Prompt
Answer {Q1, Q2, Q3} at temp. [0.3,0.5,0.7,0.9,1.0]
Video-LLM
A1.1 The sailboats are white in color..
A1.2 Sailboats finished sailing..
…
A3.5 The harbor provides a safe..
(3) Preference Selection
Judge Prompt
+ Caption
Judge {A1.1, A1.2, .., A3.5} based on relevance, accuracy, etc.
Video-LLM
(Self-Evaluation)
Prompt: Why are the sailboats..
Chosen: Sailboats finished..
Rejected: The harbor provides..
Chosen score: 5.0
Rejected score: 3.0
(4) CLIP Filtering
Chosen: Sailboats finished sailing..
Text Embedding
Video Embedding
Swap if rejected is more similar (aligned) to the video
Similarity score
CLIP Cosine Similarity
Rejected: The harbor provides..
Text Embedding
(5) DPO Finetune
Video-LLM