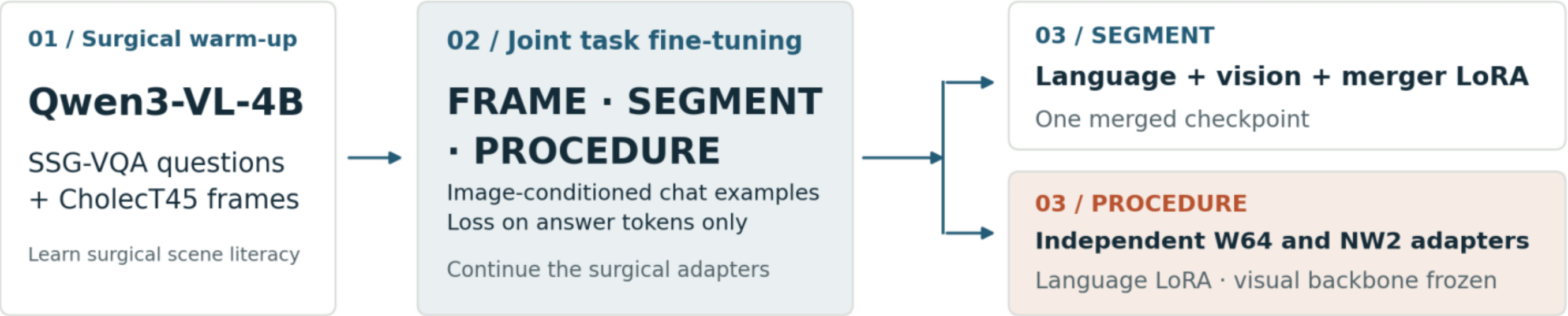


# Surgical VLM post-training

## SUPERVISED MULTIMODAL POST-TRAINING



| LoRA r16 / $\alpha$ 32         | Matched visual preprocessing               | Merge for local inference            |
|--------------------------------|--------------------------------------------|--------------------------------------|
| Segment visual mergers use r64 | Chronological frames · absolute-time clock | Independent bfloat16 model artifacts |

# Answer, then zoom in

## COARSE-TO-FINE TEMPORAL INFERENCE

### Whole procedure

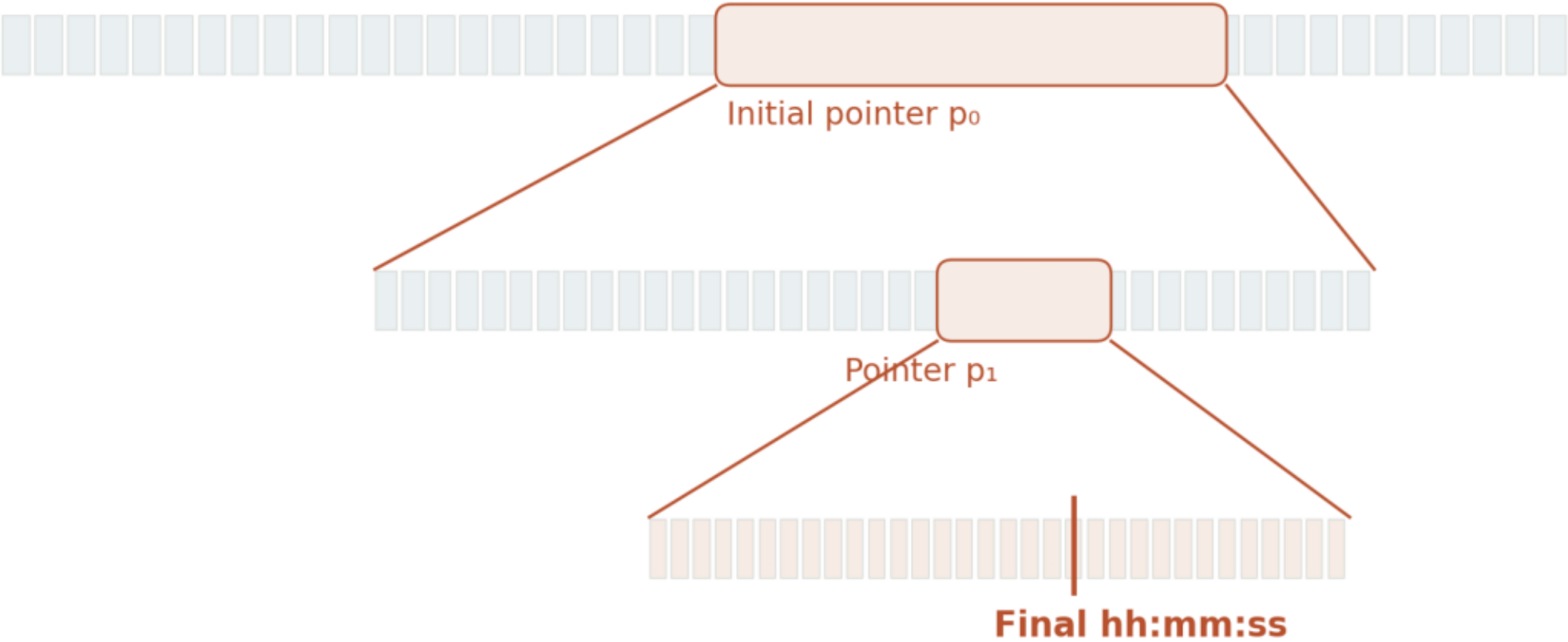
W64 · 128 frames

### 20-minute window

NW2 · 128 frames

### 100-second window

NW2 · 64 frames



One model resident at a time · absolute video clock across all passes · schematic windows

# Evaluate the evidence policy

## LOCAL OFFLINE ABLATIONS

### SEGMENT

#### Answer-guided temporal refinement

n = 1,577 · Single-timestamp questions

### SEGMENT

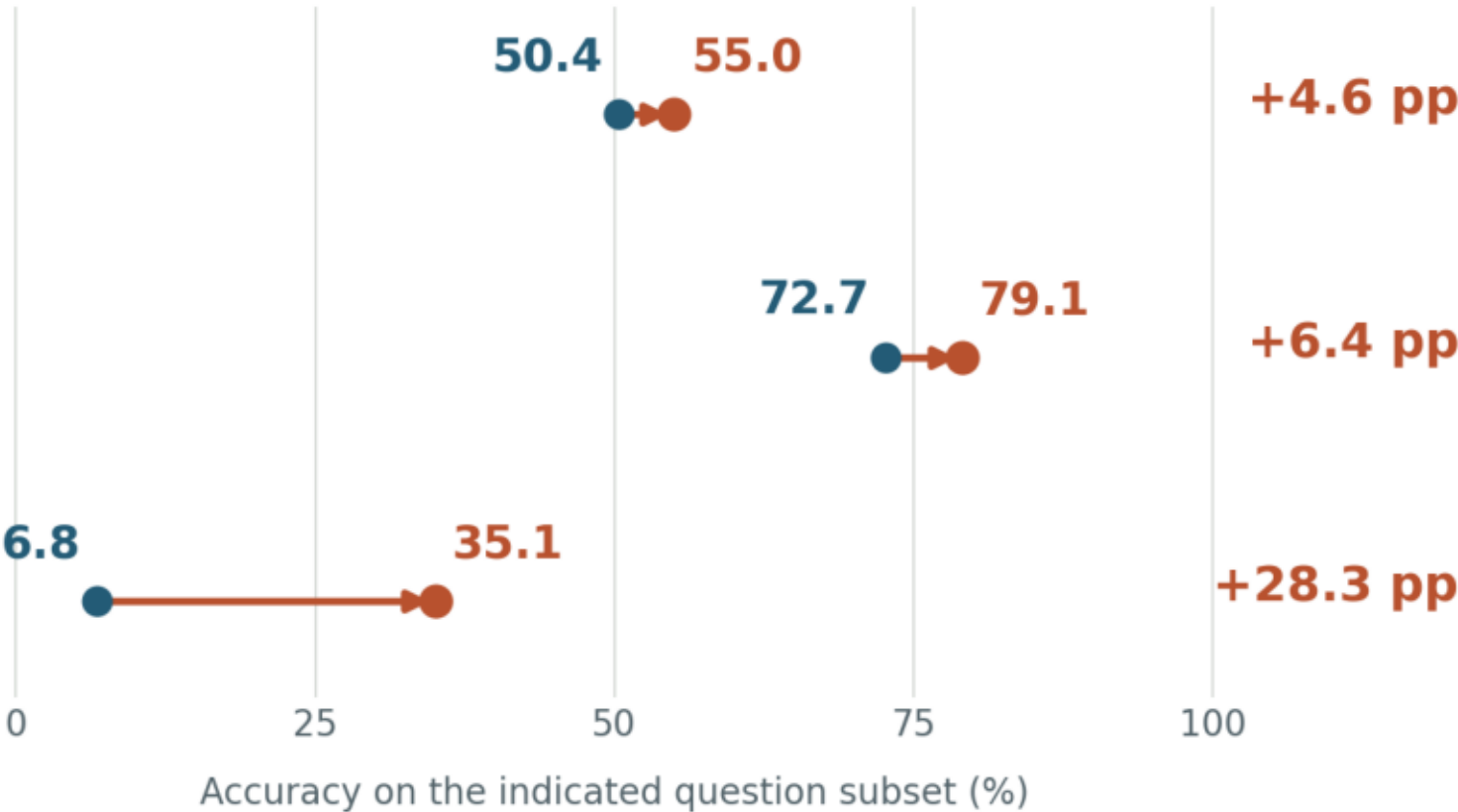
#### Question-anchored evidence windows

n = 422 · Questions that state a timestamp

### PROCEDURE

#### Coarse-to-fine temporal search

n = 746 · Single-timestamp questions



Separate question sets · Procedure search study uses one checkpoint, before the W64/NW2 split