

SUPERVISED MULTIMODAL POST-TRAINING

01 / Surgical warm-up

Qwen3-VL-4B

SSG-VQA questions
+ CholecT45 frames

Learn surgical scene literacy

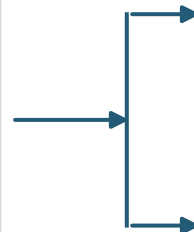


02 / Joint task fine-tuning

FRAME · SEGMENT
· **PROCEDURE**

Image-conditioned chat examples
Loss on answer tokens only

Continue the surgical adapters



03 / SEGMENT

Language + vision + merger LoRA

One merged checkpoint

03 / PROCEDURE

Independent W64 and NW2 adapters

Language LoRA · visual backbone frozen

LoRA r16 / $\alpha 32$

Segment visual mergers use r64

Matched visual preprocessing

Chronological frames · absolute-time clock

Merge for local inference

Independent bfloat16 model artifacts