User is pointing at *a clock*

Q: What is this? A. clock. B. Price tag. C. Photo. D. Basket.

A. clock. (ground truth)

Gemini 3 — D. Basket.

Qwen3-VL — B. Pricetag.

Qwen3-VL (Fine-tuned) — A. Clock.

User is pointing at *a white jacket*

Q: what is the color of the object I am pointing at?

White. (ground truth)

Gemini 3 — [...], it appears to be a brown jacket.

Qwen3-VL — [...], you are pointing at is brown.

Qwen3-VL (Fine-tuned) — White.