AI-Assisted Structured Evaluation of Oral and Maxillofacial Surgical Videos

A Four-Level Multimodal Framework with Expert-in-the-Loop Validation

Huang Qiansu Laidlaw Scholar | Supervisor: Teddy Weifa Yang

The University of Hong Kong · Laidlaw Foundation

ABSTRACT

Surgical videos are rich but unstructured sources of clinical information. We present an end-to-end system that automatically converts oral and maxillofacial (OMF) surgical videos into **structured, clinically meaningful evaluations**.

A **four-level multimodal AI framework** — visual perception & localization, sequential workflow modeling, multimodal clinical reasoning, and OSATS-compliant skill scoring — pairs every AI prediction with an **expert-in-the-loop verification loop**: independent grading, discrepancy detection with third-party arbitration, and a medical safety checker.

INTRODUCTION & MOTIVATION

Millions of operations — including OMF procedures such as cyst enucleation, impacted-tooth removal and implant placement — are video-recorded in teaching hospitals each year, yet almost all of this visual data remains **unstructured and unused**.

Manual review is slow and subjective; surgical training depends on feedback that is scarce and inconsistent; high-stakes procedures demand objective, standardized quality assurance.

> Can a multimodal AI system automatically analyze OMF surgical videos and produce a structured, clinically meaningful evaluation that experts can trust?

THE FOUR-LEVEL FRAMEWORK

Level 1 · Classification & Recognition

Identifies operation type, anatomy/pathology, instruments and spatial orientation (L/R; buccal, lingual, palatal) using standard medical nomenclature.

What is happening, where, and with what?

Level 2 · Temporal & Spatial Localization

Reconstructs a timestamped phase timeline and compares it with standard surgical protocols to detect omitted or missed clinical steps.

What was done, in what order, and what was skipped?

Level 3 · Surgical Report Generation & Reasoning

Determines completion status and predicts the exact, medically correct next action with rationale; a safety checker rejects dangerous or hallucinated recommendations.

Is the surgery complete — if not, what should happen next?

Level 4 · Objective Technical Skill Assessment

Scores the six OSATS dimensions (1–5): respect for tissue, time & motion, instrument handling, knowledge of instruments, flow & forward planning, knowledge of the procedure — each justified by a timestamped visual event.

How well was the procedure performed?

EXPERT-IN-THE-LOOP VALIDATION

Every AI prediction is paired with a **human-expert evaluation block** in a structured output template.

- **Independent grading:** two experts score the same video on a web platform.
- **Discrepancy detection:** disagreements are detected and highlighted automatically.
- **Third-party arbitration:** unresolved items are routed to a third expert.
- **Reproducible ground truth:** reference standard is not tied to any single reviewer.

Medical safety checker

Flags dangerous or hallucinated AI recommendations — e.g., a predicted next step that contradicts standard surgical protocol.

Design principle: “fast and automatic” (AI first pass) vs. “authoritative” (expert verification) — letting one AI scale to hundreds of videos while keeping every evaluation trustworthy.

FUTURE WORK

- Scale to a ~200-video stratified corpus with confidence intervals.
- Faster grading platform (keyboard-first, API batch, semi-automatic pre-fill).
- Hard safety constraints & timestamped visual grounding; error-pattern taxonomy.
- Audio/subtitle masking and chunked long-video processing.
- Parameterize the framework by specialty for generalization beyond OMF.

EVALUATION PROTOCOL

Task A · Recognition

- Macro-F1
- Accuracy
- Precision / Recall
- Sensitivity / Specificity

Task B · Localization

- tIoU
- Spatial IoU (0.3/0.5/0.7)
- Dice Coefficient
- mAP@IoU

Task C · Report Quality

- Expert grading of completeness
- Correctness · Terminology
- Grounding · Safety

Task D · Skill Correlation

- SROCC
- LCC
- MAE
- ICC · Fisher z

SYSTEM OVERVIEW

clinically meaningful evaluation that experts can trust?

SYSTEM OVERVIEW

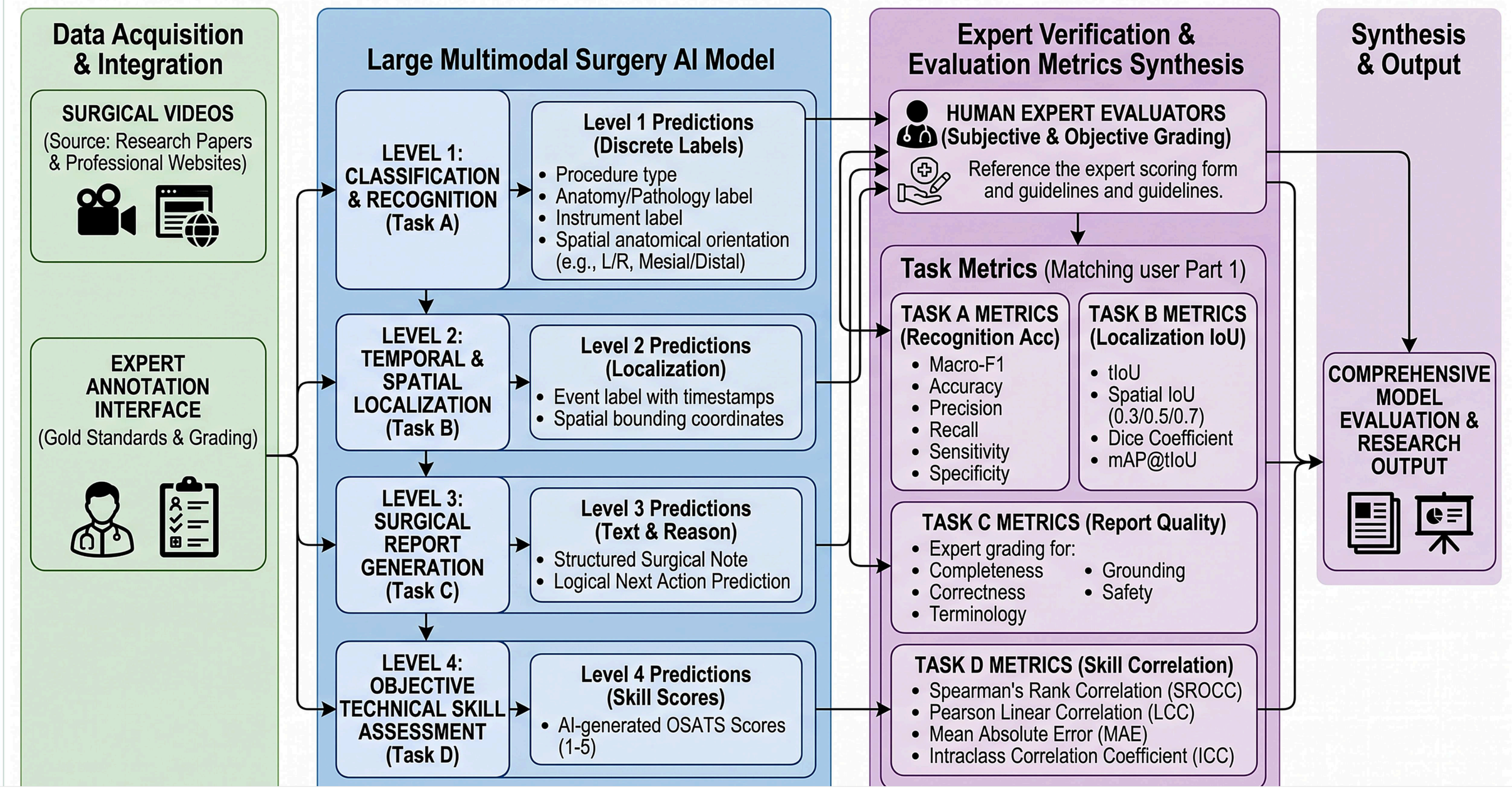

Figure 1. Data acquisition & integration → four-level multimodal analysis (classification & recognition, temporal & spatial localization, surgical report generation, OSATS skill scoring) → expert verification & evaluation-metric synthesis → comprehensive model evaluation and research output.

PILOT RESULTS & DATASET

- Pipeline ran end-to-end on all **5 pilot videos**; every AI output verifiable via the grading platform.
- Strongest at **Level 1** perception; largest residual risk at Level 2 boundary precision and borderline Level 4 OSATS ratings — matching where human experts disagree most.
- AI-assisted search returned **5,000+** candidate OMF videos → ~500 retained (target ≈2,000), with standardized nomenclature and a difficulty-gradient classification.

COMPARISON WITH EXISTING WORK

Text & image medical AI

Expert-level in clinical conversations, but analyzes consultations rather than operations — no workflow reconstruction or skill scoring.

Automated phase recognition

Accurate on fixed event sets in other specialties, yet rigid, data-hungry, and lacking clinical reasoning.

Human OSATS

The clinical gold standard — but expensive, slow, and inherently subjective.

Our system

Four-level breadth, expert-in-the-loop ground truth, low marginal cost, adaptable beyond OMF.

AI augments — not replaces — human expertise.

Objective, reproducible, and scalable feedback for trainees and practitioners who would otherwise receive little, and standardized quality assurance across institutions of any size.

ACKNOWLEDGEMENT

KEY REFERENCES