Employing a Vision-Language Model for Structured Evaluation of Oral Surgery Videos: A Four-Level Multimodal Framework with Expert Validation

Huang Qiansu

Academic Supervisor: Teddy Weifa Yang · Laidlaw Scholars Programme, The University of Hong Kong

LAIDLAW UNDERGRADUATE RESEARCH & LEADERSHIP PROGRAMME

1 INTRODUCTION & PROBLEM

- Oral-surgery procedures — cyst enucleation, impacted-tooth removal, implant placement — are routinely video-recorded for training, documentation and medico-legal use, yet almost all this footage is **unstructured and never reused**.
- Manual review is **slow and subjective** (often longer than the operation itself); junior surgeons receive scarce, inconsistent feedback; robust machine understanding of operative video is also a prerequisite for future autonomous surgical robots.
- The oral cavity is uniquely hard to analyse: a confined, deep workspace with blood, irrigation mist, saliva and rapidly moving instruments; subtle tissue boundaries; and a severe lack of public annotated datasets.

Research question: can a multimodal AI system automatically analyse oral-surgery video and produce a structured, clinically meaningful evaluation that human experts can trust?

2 OBJECTIVES

- Design a **four-level hierarchical framework** (perception → workflow → reasoning → OSATS scoring) operationalised as tiered VLM prompts.
- Build a **difficulty-graded, privacy-compliant oral-surgery video dataset** with standardised nomenclature and full provenance.
- Establish an **expert-in-the-loop validation pipeline**: web grading platform, dual independent review, automatic discrepancy detection and third-expert arbitration.
- Define **per-level quantitative metrics** and pilot-test the full end-to-end system.

3 SYSTEM OVERVIEW

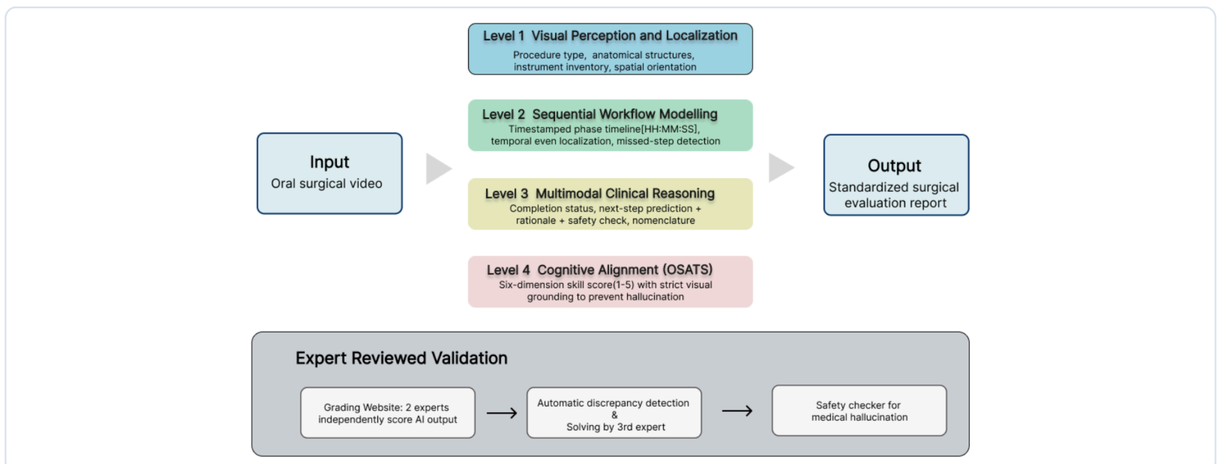


Fig 1. A four-level multimodal AI framework converts a raw oral-surgery video into a standardised evaluation report; an expert-reviewed stage (independent grading, discrepancy detection, third-party arbitration, safety checker) establishes trustworthy ground truth.

4 CURATED VIDEO DATASET

Purposive, criteria-based sampling across publisher, platform and institutional sources, with resumable automated collection, integrity checks and cross-source deduplication.

1,114	8	~500	~2,000
videos cataloged with provenance	distinct sources	retained & annotated	target size

- Sources: PubMed/PMC supplements (95), MedTube (289), YouTube (301), Quintessence Video Journal (122), Alexander Street (142), ClinicalKey (61), plus 2 instructional collections (104).
- Strict privacy rules: no whole-face or patient identifiers; operative-field obscuration < 50%; procedures on real patients only.
- Each video inherits a **difficulty grade** from its procedure type, supporting both AI training and risk-aware teaching.

5 HOW WE DIFFER FROM EXISTING WORK

Approach	What it does	Why it falls short
Clinical video AI (e.g. AMIE)	Expert-level video consultations	Conversations, not operations; no workflow or skill scoring
Phase recognition (laparoscopy)	Segments predefined phases	Fixed events, large datasets needed; no reasoning or scoring
Human OSATS rating	Gold-standard skill assessment	Slow, expensive, inherently subjective
This work	4-level reasoning + OSATS scoring + expert validation	**Oral-surgery-specific**; curated dataset; every prediction expert-checkable

6 VLM PIPELINE (GEMINI)

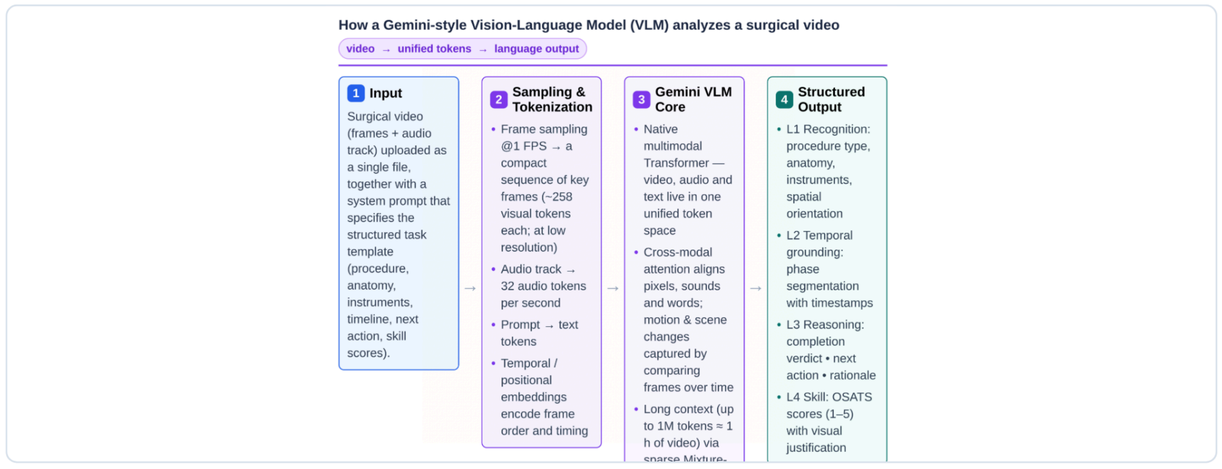


Fig 2. Frames sampled at 1 FPS (258 visual tokens each) + audio (32 tokens/s) are unified with temporal position encoding into a single ~1M-token sequence (~1 h of video); a native multimodal MoE Transformer generates structured outputs autoregressively.

7 FOUR-LEVEL ASSESSMENT FRAMEWORK

Level 1 Perception & Localization	**What is happening, where, with what?** Procedure type; anatomical & pathological structures; instruments in order; spatial orientation (left/right, buccal/lingual/palatal).
Level 2 Workflow Modeling	**What was done, in what order, what was skipped?** Timestamped phase timeline compared against standard protocol; missed-step detection.
Level 3 Clinical Reasoning	**Is it complete? What next?** Completion verdict + medically correct next action with rationale; a safety checker flags dangerous or hallucinated recommendations.
Level 4 OSATS Skill Scoring	**Cognitive alignment with experts.** Six OSATS dimensions rated 1–5; every score must be justified by a concrete, timestamped visual event to suppress hallucination.

8 EXPERT GRADING PLATFORM

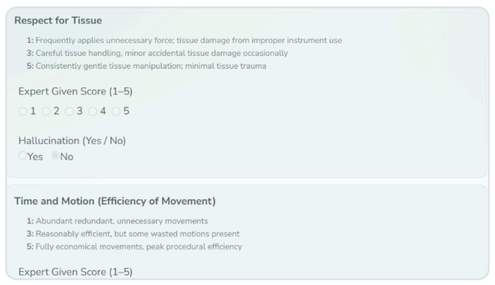
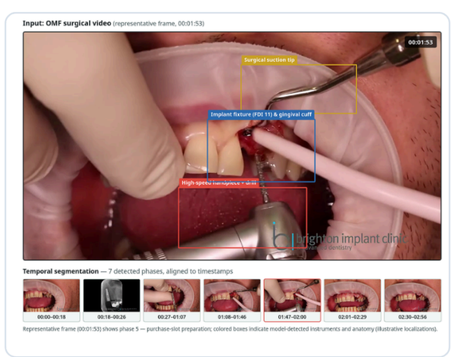


Fig 3. Web platform for trustworthy ground truth: two oral surgeons grade independently, the system highlights disagreements, unresolved items go to a third expert, and each OSATS justification is flagged for hallucination.

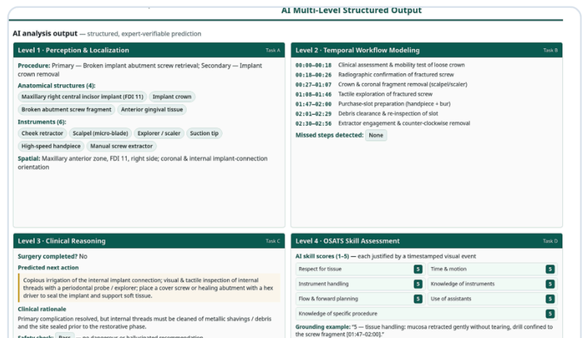
Design principle — **fast & automatic** (AI scales to hundreds of videos) is deliberately separated from **authoritative** (experts verify). This division enables scale without sacrificing trust.

9 PRELIMINARY RESULTS — CASE STUDY

Broken implant abutment-screw retrieval (FDI 11), 00:00–02:56.



Input: frame at 01:53 (osteotomy preparation) with detected instrument/anatomy boxes; thumbnail bar maps the 7 detected phases across the full timeline.



Four-level AI output: L1 — procedure, 4 structures, 6 instruments, orientation; L2 — 7-phase timeline, no missed steps; L3 — incomplete, correct next step predicted, safety check **Pass**; L4 — OSATS scores of 5, each grounded in a timestamped event.

10 CONCLUSION & FUTURE WORK

Achieved

First end-to-end multimodal framework turning unstructured oral-surgery video into standardised, expert-verifiable evaluations — supporting objective, scalable surgical training and quality assurance.

Next

Expand the corpus toward ~2,000 videos, add multi-model comparison, harden safety guardrails, test cross-specialty generalisation and run multi-institutional validation.

Limitations

Pilot on 5 videos only (low statistical power); oral-surgery-specific; not yet formally validated across multiple institutions.