ORALFLOW: A LONGITUDINAL MULTIMODAL BENCHMARK ACROSS CLINICAL WORKFLOWS IN DENTISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have expanded dental AI from isolated image interpretation to open-ended clinical interaction. However, existing dental models often frame diagnosis as a static, modality-specific problem. Complete dental workflows require integrating complementary multimodal evidence and continuously retaining, retrieving, and updating evidence across the typical stages of dental care, including diagnosis, treatment planning, and follow-up. To bridge this gap, we introduce **OralFlow-Bench**, the first benchmark for evaluating longitudinal multimodal dental clinical workflows. It contains 668 cases from two complementary sources: OralFlow-Literature comprises 281 published long-horizon case reports spanning initial presentation, treatment, and follow-up, while OralFlow-Cohort comprises 387 complex, multi-visit hospital cases with 10,326 images across six modalities. We evaluate six MLLMs with six memory methods across five task categories to handle long-context clinical evidence in longitudinal clinical workflows. On end-to-end evaluation across the clinical workflow, models achieve 37.56% diagnosis accuracy and 29.57% treatment and 40.64% follow-up scores, revealing the challenges in handling complete longitudinal dental workflows. Even given ground-truth observations, memory methods yield variation in diagnostic accuracy across MLLMs, ranging from 46.94% to 93.88%, and these differences propagate to subsequent stages, further degrading treatment and follow-up performance. Missing-modality ablations reveal the complementary roles of multiview skeletal model and X-Ray, while noise ablations show that robust performance depends on selective memory updating and retrieval rather than context size alone. Code and benchmark resources are available at `https://anonymous.4open.science/r/OralFlow/`.

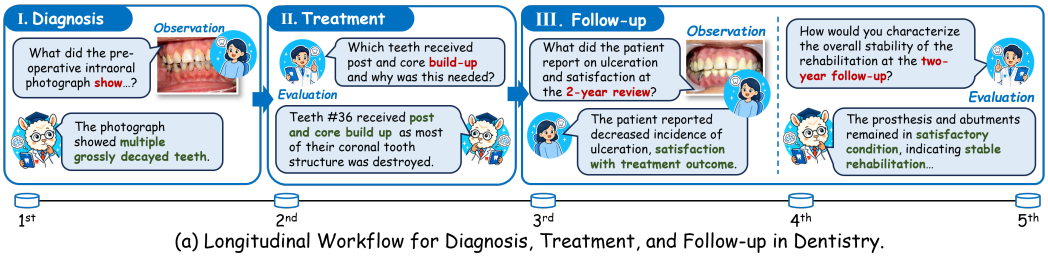
1 INTRODUCTION

Multimodal large language models (MLLMs) have increasingly been adopted across medical domains, ranging from generalist biomedical systems [1, 2] and medical visual question answering [3, 4, 5, 6] to clinical decision-support agents [7, 8, 9]. These advances are reshaping medical AI from isolated image interpretation toward open-ended clinical interaction. In dentistry, this shift raises a fundamental question: Can MLLMs effectively support longitudinal multimodal dental clinical workflows as patient information and clinical evidence evolve over time?

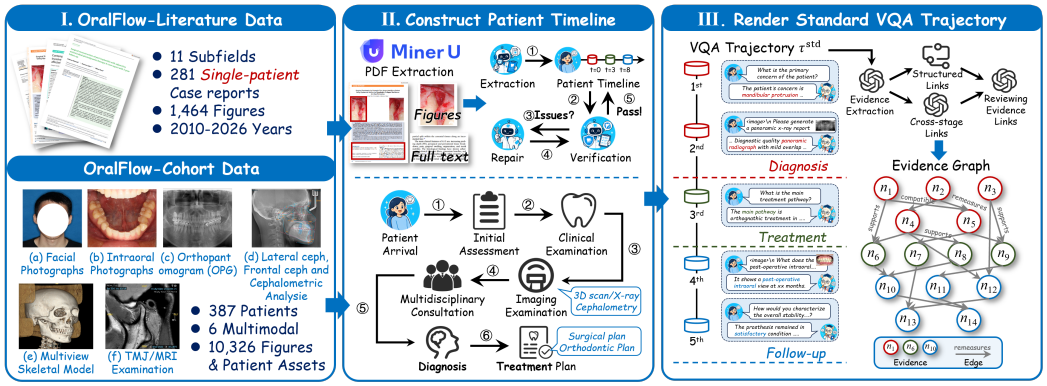
Conventional Dental AI models [10, 11, 12, 13, 14] often frame diagnosis as a static, modality-specific task. While such models demonstrate promise for localized visual assessment, real clinical care requires substantially broader integration. Dentists must synthesize medical history with complementary multimodal evidence spanning teeth, skeletal structures, maxillofacial morphology, and soft tissues to establish a comprehensive clinical picture and formulate an appropriate treatment plan. This gap becomes particularly pronounced in complex dentofacial deformities, where clinical decisions depend on integrating heterogeneous evidence throughout the care process.

Specifically, as shown in Figure 1(a), a complete clinical workflow typically unfolds across three stages: diagnosis, treatment planning, and multiple rounds of postoperative follow-up. Diagnostic assessment informs treatment planning, treatment alters the patient's clinical state, and follow-up

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Longitudinal Workflow for Diagnosis, Treatment, and Follow-up in Dentistry.



(b) OralFlow-Bench data processing, timeline construction and VQA trajectory rendering Pipelines.

Figure 1: **Overview of OralFlow-Bench.** (a) A longitudinal clinical trajectory exposes observations and evaluations across diagnosis, treatment, and follow-up. (b) OralFlow-Bench data are first organized into ordered timelines, then rendered as standard VQA trajectories and used to construct temporally directed evidence graphs.

provides evidence for maintaining, refining, or revising prior decisions. However, a clinically meaningful multimodal benchmark that evaluates the full oral-care process remains unavailable. Bridging existing dental models to longitudinal clinical workflows requires two complementary capabilities. *On the one hand,* models must integrate complementary multimodal evidence, since each modality captures only a partial aspect of the patient's condition. *On the other hand,* they needs to retain, retrieve, and update evidence across clinical stages, since the clinical significance of an early observation may only emerge through subsequent treatment or follow-up outcomes [15, 16].

These requirements create a long and heterogeneous context. Long-term memory systems, including summary memory [17], retrieval-augmented memory [18], and agentic memory [19, 20, 21], provide practical means to manage such streams, enabling MLLMs to integrate evidence and decisions from earlier stages and answer current questions using the information available at that time [22, 23].

To bridge this gap, we introduce **OralFlow-Bench**, the first benchmark for evaluating longitudinal multimodal dental clinical workflows. Specifically, OralFlow-Literature is based on 281 published long-horizon case reports and captures clinical courses from initial presentation through treatment and follow-up. OralFlow-Cohort contains short-horizon, multi-visit cases from 387 patients with complex dentofacial deformities at a hospital in China, comprising 10,326 images across six modalities. We transform both sources into temporally ordered clinical trajectories and evaluate diagnosis, treatment, and follow-up under staged information release. This design moves beyond isolated inputs to test whether models can perceive multimodal observations and maintain clinically meaningful evidence as a workflow evolves. Our contributions are summarized as follows:

- We construct OralFlow-Bench, the first benchmark for longitudinal multimodal dental workflows across diagnosis, treatment planning, and follow-up. It contains 668 cases from OralFlow-Literature with 281 long-horizon case reports and OralFlow-Cohort with 387 multi-visit hospital cases and 10,326 images across six modalities.
- We show that longitudinal clinical workflows expose challenges beyond diagnostic evaluation. Across five task categories, MLLMs show low scores on treatment and follow-up. Perception errors propagate through stages, while missing-modality analyses show complementary roles of multiview skeletal model and X-Ray in planning and evidence grounding.

- We show that robust longitudinal performance requires capabilities beyond memory mechanisms. Evaluations of six MLLMs with six memory methods show that larger context windows or more structured memory alone do not ensure robust performance. Longitudinal clinical workflows instead require stage-specific evidence retrieval, memory updating, and robustness to distractors.

2 ORALFLOW-BENCH DATA AND TRAJECTORY CONSTRUCTION

2.1 LONGITUDINAL ORAL DATA PROCESSING

2.1.1 ORALFLOW-LITERATURE

OralFlow-Literature is constructed from published dental case reports documenting the longitudinal clinical course of individual patients. We collect case reports published between 2010 and 2026 and retain those that contain a patient-specific diagnostic presentation, treatment, and follow-up. Unlike hospital records, reports narrate the clinical course retrospectively, often presenting diagnostic evidence, treatment decisions, and follow-up outcomes together with the corresponding time.

To recover the timeline from reports, MinerU [24, 25] extracts body text, tables, images and captions, after which the timeline extractor organizes the patient-specific material into ordered clinical timepoints. Each timepoint is assigned to diagnostic workup, initial treatment, or follow-up, while preserving clinically relevant temporal, anatomical, and quantitative details, together with links to the figures. Diagnostic workup contains visible observations, initial treatment contains held-out decisions, and follow-up contains outcome observations together with optional held-out management decisions. The construction pipeline therefore reconstructs a clinically valid timeline while preventing later outcomes or retrospective explanations from entering earlier stages.

Then, a separate verifier audits the extracted timeline against the source report for factual support, chronology, stage boundaries, role assignment, figure and answer leakage. High- or medium-severity findings trigger a localized repair of only the affected timepoint span. The repaired timeline is then verified again until it passes or reaches the iteration limit.

2.1.2 ORALFLOW-COHORT

OralFlow-Cohort comprises 387 patients with dentofacial deformities from one hospital in China between 2022 and 2026. The records were collected for evaluation under approval from the hospital's Ethics Committee (IRB No. XX-XX-XX-X). For each patient, multimodal imaging and textual records were compiled through multidisciplinary case discussions involving orthognathic surgeons and orthodontists. The textual record for each patient was curated by two attending physicians, each with more than 5 years of clinical experience, and was subsequently reviewed and corrected by one chief physician in orthognathic surgery and one chief physician in orthodontics.

As shown in Figure 1(b), the patient timeline follows the clinical workflow: *Patient Arrival*, *Initial Assessment*, *Clinical Examination*, *Imaging Examination*, *Multidisciplinary Consultation*, *Diagnosis*, and *Treatment Plan*. Initial assessment records textual profile, chief complaint, and history. Clinical and imaging examinations release the corresponding findings. The visible record comprises six clinically distinct multimodal sources, including (i) Facial Photographs (FP)[1], (ii) Intraoral Photographs (DP), (iii) Orthopantomogram (OPG), (iv) Lateral Cephalometric Radiograph (Lateral ceph), Frontal Cephalometric Radiograph (Frontal ceph) and Cephalometric Analysis (v) Multiview Skeletal Model (MVSM), and (vi) TMJ MRI Examination.

Based on these observations, multidisciplinary consultation integrates observations from orthognathic surgeons and orthodontists. Diagnosis and treatment planning record the resulting clinical decisions, including surgical and orthodontic plans, which are held out from the diagnostic stream. To separate observation from downstream decisions, questions about these decisions are asked only after the required diagnostic observations are available, and their reference answers are released at the designated treatment stage, thereby preventing future information leakage.

[1]Facial photographs and other sensitive patient images are de-identified before processing.

2.1.3 RENDERING STANDARD VQA TRAJECTORIES

After extracting the longitudinal clinical timeline for each patient, we render every case as a staged clinical stream. Each released clinical observation is represented as a VQA in a human–assistant dialogue format with corresponding multimodal images, yielding a shared temporally ordered record for all evaluated methods. Source records of treatment, surgery, and follow-up decisions are preserved as hidden decision records rather than added to the observation stream. Formally, a standardized VQA trajectory τ_c^{std} with T stages for each case $c \in \mathcal{C}$ is:

$$\tau_c^{\mathrm{std}} = (s_1, \ldots, s_T), \quad s_t = (\mathcal{I}_t, \mathcal{O}_t, \mathcal{H}_t, \gamma_t), \tag{1}$$

where s_t is the t-th stage, $\mathcal{I}_t$ is the set of multimodal images, $\mathcal{O}_t = \{(u_{t,i}, a_{t,i}, I_{t,i})\}_i$ is the set of visible observation VQA turns, $u_{t,i}$ is the query, $a_{t,i}$ is the reference response validated by a dentist, $I_{t,i} \subseteq \mathcal{I}_t$ is the associated images, $\mathcal{H}_t$ is the set of held-out decision records used to evaluate MLLM π_θ, and γ_t records temporal metadata such as visit order or elapsed follow-up time.

2.2 EVIDENCE GRAPH CONSTRUCTION

To support reliable evaluation, we explicitly identify the atomic evidences that support each answer and extract them from the standard trajectory before generating benchmark tasks [26]. From every visible VQA turn, we record only source-supported findings, measurements, diagnoses, negations, and anatomical qualifiers, without inferring unstated clinical rationale.

For each case c, we first extract atomic evidence and let $\mathcal{K}_c$ denote its index set. The j-th evidence node is $n_{c,j} = (t_{c,j}, f_{c,j})$, where $j \in \mathcal{K}_c$, $t_{c,j}$ is the stage or timepoint at which the evidence becomes available, and $f_{c,j}$ is its clinical content. Normalized attributes used for matching and audit, including values, units, anatomical sites, and laterality, are associated with each node. Based on this atomic evidence, we then construct the temporally directed evidence graph $\mathcal{G}_c$:

$$\begin{gathered} \mathcal{G}_c = (\mathcal{N}_c, \mathcal{L}_c), \quad \mathcal{N}_c = \{n_{c,j} : j \in \mathcal{K}_c\}, \quad \mathcal{L}_c = \mathcal{L}_c^{\mathrm{intra}} \cup \mathcal{L}_c^{\mathrm{cross}}, \\ \mathcal{L}_c^{\mathrm{cross}} \subseteq \{(n_{c,j}, n_{c,k}, r_{j,k}) \in \mathcal{N}_c \times \mathcal{N}_c \times \mathcal{R} : t_{c,j} < t_{c,k}\}. \end{gathered} \tag{2}$$

where $\mathcal{N}_c$ and $\mathcal{L}_c$ are respectively the node and directed-edge sets, $r_{j,k}$ is the relation from $n_{c,j}$ to $n_{c,k}$, and $\mathcal{R}$ denotes allowable clinical relationships, including confirmation, support, and conflict.

We construct two types of candidate evidence edges. Intra-stage evidence edges $\mathcal{L}_c^{\mathrm{intra}}$ connect atomic evidence from the same stage when they share a normalized clinical attribute or anatomical scope. Cross-stage evidence edges $\mathcal{L}_c^{\mathrm{cross}}$ connect earlier and later evidence when normalized attributes match or when the earlier evidence factually supports, confirms, refines, or conflicts with a later observation or interpretation. All candidate evidence edges are independently reviewed. Each retained cross-stage evidence edge satisfies $t_{c,j} < t_{c,k}$, preserving clinical information flow [27].

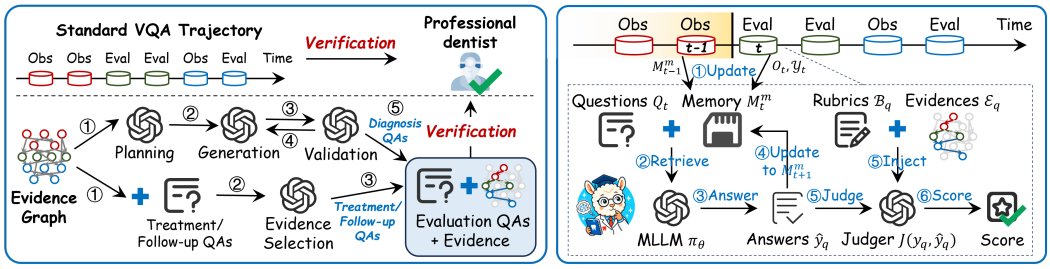
2.3 TRAJECTORY VARIANT CONSTRUCTION

To test robustness under controlled changes to the information stream, OralFlow-Bench derives three variant trajectories while preserving task identity, temporal placement, and evidence provenance.

Model-perception trajectories $\tau_c^{\mathrm{mp},\theta}$. For each evaluated MLLM π_θ, a model-perception trajectory replaces the reference answer of each image-grounded VQA turn $(u_{t,i}, a_{t,i}, I_{t,i})$ in τ_c^{std} with its generated answer, $\pi_\theta(u_{t,i}, I_{t,i})$. Non-image observations and hidden decision records remain fixed, isolating how perception errors propagate to downstream clinical decisions.

Missing-modality trajectories $\tau_c^{\mathrm{miss}(D)}$. For each standard trajectory τ_c^{std}, a missing-modality variant removes a subset D of multimodal images and their associated visible VQA turns while retaining the remaining observations and original temporal metadata.

Noisy trajectories $\tau_c^{\mathrm{noise}(k)}$. For noise robustness, we construct a frozen pool $\mathcal{P}_{\mathrm{noise}}$ of 50 distractor VQA pairs, with 10 examples in each of 5 categories: scheduling and billing logistics, general lifestyle advice, benign repeated history, postoperative concerns, and modality-adjacent distractors. Each noisy trajectory inserts k deterministically sampled pairs from $\mathcal{P}_{\mathrm{noise}}$ into the standard trajectory. Modality-adjacent distractors are inserted into matching stages when possible, whereas generic distractors are distributed across non-empty visible stages.



(a) Benchmark Task Generation and Validation Pipeline. (b) Streaming Model Evaluation pipeline.

Figure 2: **OralFlow-Bench construction and evaluation.** (a) The evidence graph supports task generation and verification by professional dentists. (b) The resulting tasks are evaluated by streaming each trajectory through a memory method m and the evaluated MLLM π_θ.

3 ORALFLOW-BENCH

3.1 TASK GENERATION

To construct clinically valid questions that are supported by available evidence and protected from wording or future leakage, we automatically generate benchmark tasks from the evidence graph. Each task is temporally constrained and tests how the information stream is retained, retrieved, integrated, and updated. As shown in Figure 2(a), we first plan the evidence and temporal constraints of a task, then generate its clinical question and reference answer, and finally subject the candidate to independent validation. For a task indexed by q, let:

$$q = (x_q, y_q, \mathcal{E}_q, \alpha_q, \beta_q), \quad \mathcal{E}_q \subseteq \{n_{c,j} \in \mathcal{N}_c : t_{c,j} \leq \alpha_q\}, \quad \alpha_q < \beta_q, \tag{3}$$

where x_q is the question, y_q is the reference answer, $\mathcal{E}_q$ is the required evidence set, α_q is the ask stage, and β_q is the release stage at which y_q becomes eligible to enter subsequent memory states. The evidence constraint ensures that every selected item has become observable no later than α_q, while $\alpha_q < \beta_q$ ensures that the reference answer is unavailable when the answer is generated.

OralFlow-Bench evaluates five task categories. ❶ *Modality perception* reads findings directly visible or stated in one modality at one stage. ❷ *Longitudinal evidence recall* retrieves earlier evidence when answering a later-stage question. ❸ *Cross-modal or cross-temporal reasoning* integrates evidence from at least two modalities or timepoints into a supported clinical interpretation. ❹ *Memory update and correction* uses later evidence to confirm, refine, or correct the current clinical memory state. These four categories are evidence-grounded tasks scored by Accuracy. ❺ *Treatment and follow-up planning* produces a comprehensive decision from the available history, and is scored using source-specific rubrics because clinically valid decisions contain multiple components.

For evidence-grounded diagnostic tasks, the planner selects evidence under constraints. A generator maps the plan to a clinical question and reference answer. A verifier checks source support, medical validity, task-category alignment, answer leakage, and temporal feasibility. Minor defects are directly repaired. Otherwise, structured feedback guides regeneration, and invalid candidates at the iteration limit are discarded. For decision tasks, OralFlow-Bench uses the corresponding hidden decision record in $\mathcal{H}_t$ as the source-grounded clinical decision y_q, links it to evidence available by α_q, and constructs a 100-point grading rubric. Further details can be found in *Appendix B*.

3.2 BENCHMARK COMPOSITION

OralFlow-Bench comprises 668 cases, including 387 de-identified hospital patients in OralFlow-Cohort, with 6,958 tasks and 10,326 images averaging 26.7 images per patient, and 281 published case reports in OralFlow-Literature, with 5,591 tasks and 1,464 report images averaging 5.2 images per case. Their standard trajectories yield 14,208 VQA pairs and 12,549 evaluation tasks. Figure 3(a) shows the distribution of modalities and VQA pairs across diagnosis, treatment, and follow-up.

To assess whether the benchmark is clinically accurate and suitable for evaluation [28], a dental team reviewed tasks from 20 cases in each subset, covering 795 questions. Each question is rated from 1 (unacceptable) to 5 (excellent) across five dimensions. Figure 3(b) shows the expert-quality assessment, supporting benchmark quality. Additional details are included in *Appendix C*.

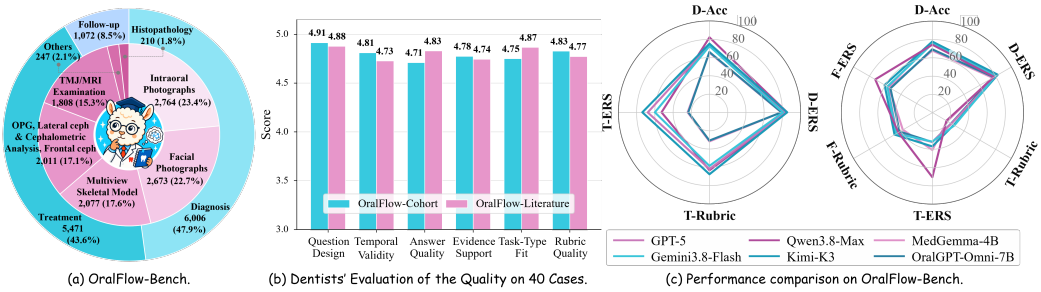


Figure 3: (a) Composition of OralFlow-Bench. (b) Dentist-assessed generation quality across five dimensions for benchmark subsets: four shared dimensions and one task-specific dimension. (c) Model-level standard-trajectory performance on diagnosis, treatment, and follow-up metrics.

3.3 Streaming Evaluation and Metrics

We evaluate each memory method as an online clinical MLLM. Let the pre-answer memory state M_t^m denote the state of method m after ingesting the information released at stage s_t and before answering tasks scheduled at that stage, with initial state $M_0^m = \emptyset$. The method writes the newly visible observations and eligible released answers into memory:

$$M_t^m = U_m\big(M_{t-1}^m, \mathcal{O}_t \cup \mathcal{Y}_t\big)\,, \tag{4}$$

where U_m is the update operator and the release set $\mathcal{Y}_t = \{(x_q, y_q) : \beta_q = t\}$ contains the reference answers to previously evaluated tasks whose release stage is t. The task set $\mathcal{Q}_t = \{q : \alpha_q = t\}$ contains all questions scheduled at stage t. For each $q \in \mathcal{Q}_t$, the MLLM π_θ produces the answer:

$$\hat{y}_q^m = \pi_\theta(x_q, R_m(M_t^m, x_q))\,, \quad q \in \mathcal{Q}_t. \tag{5}$$

where R_m is the retrieval operator. A task never observes its own reference answer before answering. Instead, each answer is generated from only the information available at its scheduled stage, mirroring the sequential evidence acquisition and decision-making process in real clinical care.

For each evidence-grounded task q, **Accuracy** is defined by the clinical judge, and for each treatment or follow-up planning task q, the normalized **Rubric** score is [29, 30]:

$$\mathrm{Acc}_q(m, \pi_\theta) = J(\hat{y}_q^m, y_q), \qquad \mathrm{Rubric}_q(m, \pi_\theta) = \frac{\sum_{r \in \mathcal{B}_q} b_{q,r}(\hat{y}_q^m)}{\sum_{r \in \mathcal{B}_q} b_{q,r}^{\max}}. \tag{6}$$

Here, $J(\hat{y}_q^m, y_q)$ is the binary correctness decision of the clinical judge given the model and reference answers. $\mathcal{B}_q$ is the rubric-criterion set for task q, where $r \in \mathcal{B}_q$ indexes a criterion, $b_{q,r}(\hat{y}_q^m)$ is its awarded score, and $b_{q,r}^{\max}$ is its maximum score. Across all tasks, **Evidence Recall Score (ERS)** measures whether the answer covers the task's gold evidence:

$$\mathrm{ERS}(m, \pi_\theta) = \frac{\sum_{q \in \mathcal{Q}} |\widehat{\mathcal{E}}_q \cap \mathcal{E}_q|}{\sum_{q \in \mathcal{Q}} |\mathcal{E}_q|}, \tag{7}$$

where $\mathcal{Q} = \bigcup_t \mathcal{Q}_t$ is the full task set, $\widehat{\mathcal{E}}_q$ is the subset judged to be covered by $\hat{y}_q^m$. For model-perception trajectories, we additionally evaluate generated answers by **precision (Prec.)**, **evidence recall score**, **F1**, and **hallucination control (Hallu.)** before downstream memory evaluation [31].

4 Experiments

4.1 Experimental Setup

MLLMs and Memory Methods. All experiments are conducted on NVIDIA A800 GPUs. We evaluate GPT-5 [32], Gemini3.8-Flash [33], Qwen3.8-Max [34], Kimi-K3 [35], MedGemma-4B [36], and OralGPT-Omni-7B [13]. The memory methods are Full Context [37], Summary Memory [17], Vector Memory [18], Mem0 [38], LangMem [39], and Graphiti [40]. GPT-5-mini [32] is used as the scoring judge because it gives stable assessments and reliably follows the grading instructions.

Table 1: *Main standard-trajectory* results on OralFlow-Bench for MLLMs and memory methods. **Bold** and underlined indicate the best and second-best results, respectively.

Model	Memory	OralFlow-Cohort				OralFlow-Literature					
		Diagnosis		Treatment		Diagnosis		Treatment		Follow-up	
		Acc	ERS	Rubric	ERS	Acc	ERS	Rubric	ERS	Rubric	ERS
GPT-5 [32]	Vector	96.33	**99.32**	72.91	81.19	98.89	99.34	28.22	40.84	51.89	69.64
	Mem0	26.41	41.58	47.41	47.23	52.64	59.67	7.85	6.19	39.00	47.58
	Graphiti	70.71	87.63	51.47	55.02	59.14	71.74	25.74	30.99	39.53	43.98
	LangMem	62.85	78.13	55.76	61.77	66.90	74.99	22.34	29.31	41.92	52.78
	Summary	90.93	95.88	71.28	80.49	82.16	87.68	32.05	43.69	43.96	54.53
	Full Context	95.38	99.02	72.24	80.80	99.64	99.83	28.03	41.81	54.19	69.99
Gemini3.8 -Flash [33]	Vector	97.70	99.82	66.34	67.37	99.37	99.48	34.71	43.00	56.27	74.63
	Mem0	17.63	29.30	47.04	51.71	39.86	46.82	7.54	4.35	31.22	38.03
	Graphiti	69.77	85.79	49.73	49.53	62.12	70.75	26.94	26.59	38.06	44.30
	LangMem	59.09	73.57	56.33	56.32	68.96	75.86	26.03	30.78	41.68	54.38
	Summary	91.85	96.10	65.87	67.66	82.84	87.61	37.08	42.26	48.29	56.45
	Full Context	98.11	99.83	65.97	67.01	99.57	99.70	35.26	42.97	59.35	75.09
Qwen3.8 -Max [34]	Vector	**99.66**	98.65	68.96	66.26	98.81	99.39	20.50	**89.03**	51.83	89.45
	Mem0	73.39	68.64	59.84	42.83	43.18	43.03	10.26	31.51	34.12	53.42
	Graphiti	56.07	77.64	53.52	29.57	45.20	59.24	16.11	59.76	33.47	48.79
	LangMem	72.35	74.09	58.73	44.05	69.40	73.45	16.31	68.94	37.64	68.64
	Summary	90.87	93.74	66.64	63.38	88.73	93.71	20.54	85.86	47.55	82.22
	Full Context	99.40	98.91	72.56	69.87	99.64	99.75	23.11	88.94	54.85	**91.91**
Kimi-K3 [35]	Vector	96.47	98.69	**75.34**	**83.13**	99.21	99.44	32.93	48.84	55.51	75.73
	Mem0	45.02	63.26	64.26	75.39	54.58	61.27	7.94	6.38	41.38	51.10
	Graphiti	68.56	85.66	56.54	56.18	63.34	73.90	27.42	31.77	39.36	46.76
	LangMem	55.04	70.80	60.72	63.34	68.41	76.14	27.01	37.56	45.34	56.38
	Summary	89.43	94.61	74.56	81.76	81.69	87.17	**34.63**	48.15	46.96	57.45
	Full Context	96.24	98.80	75.25	82.49	**99.84**	**99.89**	32.80	48.73	**58.77**	75.02
MedGemma -4B [36]	Vector	79.16	87.32	34.28	24.28	88.97	93.83	21.08	48.56	39.95	58.45
	Mem0	59.69	69.07	25.84	25.43	44.60	50.98	15.85	24.70	32.33	46.15
	Graphiti	44.27	71.60	25.86	18.26	40.21	56.05	16.85	41.83	30.38	39.95
	LangMem	58.83	72.75	29.14	21.58	64.77	74.13	17.78	45.02	35.55	58.20
	Summary	70.71	84.99	33.91	29.24	79.72	88.21	19.95	56.19	41.37	62.15
	Full Context	79.16	88.06	31.88	25.04	88.97	94.23	20.95	49.46	45.31	63.08
OralGPT -Omni-7B [13]	Vector	78.90	88.13	32.81	25.98	87.54	92.88	**25.47**	33.12	42.90	55.63
	Mem0	57.88	67.76	29.38	19.80	48.40	56.08	21.12	22.69	36.79	44.69
	Graphiti	45.74	72.97	26.32	17.78	43.89	59.58	23.05	32.72	36.55	39.56
	LangMem	59.26	71.84	28.02	20.53	67.38	75.78	24.64	33.62	41.69	54.35
	Summary	72.35	85.47	36.96	27.86	79.24	88.70	25.31	38.84	45.28	60.80
	Full Context	82.00	88.70	32.10	27.35	88.61	94.01	25.32	32.91	46.56	58.91
Overall		72.42	82.17	52.10	50.21	73.51	79.56	23.30	41.33	43.52	58.89

Implementation Details. The answer phase streams trajectories through each memory method and freezes model answers. The scoring phase reads the frozen answers and applies the clinical judge. Visual observations are generated before memory evaluation and verified against the same evidence used downstream. Detailed metrics and hyperparameters are provided in *Appendix D*.

4.2 Overall Performance

Finding 1. Verified observations support diagnosis, but treatment and follow-up planning remain challenging. Table 1 and Figure 3(c) show that diagnosis is stronger than treatment and follow-up planning. High diagnostic scores arise because standard trajectories present dentist-validated observations in temporal order, allowing models to reason over explicit evidence. This indicates that current MLLMs can retrieve and integrate verified clinical findings effectively when relevant evidence is explicit. However, this capability does not translate into complete clinical management. Treatment and follow-up require selecting among alternatives, satisfying patient-specific

Table 2: *Model-perception trajectory* results on OralFlow-Bench. Δ denotes the change from the corresponding *standard trajectory*. Higher values are better for all metrics.

Model	Memory	OralFlow-Cohort				OralFlow-Literature					
		Diagnosis		Treatment		Diagnosis		Treatment		Follow-up	
		Acc	ERS	Rubric	ERS	Acc	ERS	Rubric	ERS	Rubric	ERS
Qwen3.8 -Max [34]	Vector	**14.73**	**37.60**	64.34	**45.41**	**82.09**	**89.07**	19.06	79.55	50.16	84.28
	Summary	2.33	21.66	61.39	34.60	60.62	70.44	19.20	74.96	31.99	47.58
	Full Context	14.47	37.31	**64.45**	44.86	80.66	88.61	18.79	**80.13**	**51.85**	**85.46**
MedGemma -4B [36]	Vector	10.68	17.45	25.95	11.74	66.07	76.30	20.36	41.53	37.35	52.07
	Summary	8.79	18.39	27.63	15.12	56.47	69.89	19.79	52.16	34.37	45.47
	Full Context	10.59	17.28	25.74	13.14	66.79	76.98	20.40	42.78	42.73	55.17
OralGPT -Omni-7B [13]	Vector	11.02	18.46	20.56	12.27	61.33	71.55	24.65	27.61	38.70	46.83
	Summary	10.42	22.18	29.21	14.47	47.33	62.83	24.12	32.07	35.91	39.06
	Full Context	10.08	17.85	21.74	13.99	61.68	73.27	**24.81**	28.43	42.65	51.50
Overall		10.35	23.13	37.89	22.84	64.78	75.44	21.24	51.02	40.64	56.38
Δ (Model − Standard)		-73.23	-67.31	-7.68	-17.07	-24.13	-18.42	-1.23	-7.08	-5.54	-12.80

Table 3: *Model-perception trajectory* quality on OralFlow-Cohort and OralFlow-Literature. **Bold** and underlined indicate the best and second-best results, respectively.

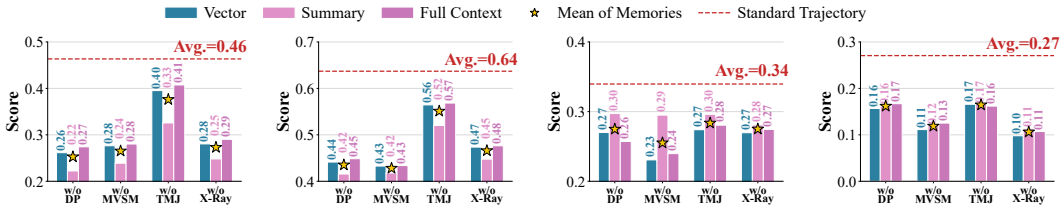
Model	OralFlow-Cohort				OralFlow-Literature							
	Diagnosis				Diagnosis				Follow-up			
	Prec.	Recall	F1	Hallu.	Prec.	Recall	F1	Hallu.	Prec.	Recall	F1	Hallu.
GPT-5 [32]	41.35	74.66	53.22	61.08	27.87	65.79	39.16	39.44	33.58	78.04	46.95	41.26
Gemini3.8-Flash [33]	**43.94**	**88.16**	**58.65**	52.28	**49.18**	**92.81**	**64.29**	**64.48**	**53.12**	**93.06**	**67.63**	**71.12**
Qwen3.8-Max [34]	43.17	85.90	57.46	59.40	30.55	83.70	44.76	49.07	32.88	89.95	48.15	45.83
Kimi-K3 [35]	42.05	82.98	55.81	**61.47**	29.39	75.00	42.23	48.40	32.27	88.01	47.23	44.80
MedGemma-4B [36]	19.19	40.13	25.96	32.27	18.41	46.66	26.41	31.23	24.44	52.89	33.43	31.17
OralGPT-Omni-7B [13]	17.27	44.36	24.86	37.30	14.25	54.81	22.62	33.06	16.84	58.86	26.19	30.20
Overall	34.50	69.37	45.99	50.63	28.28	69.80	39.91	44.28	32.19	76.80	44.93	44.06

anatomical and procedural constraints, sequencing interventions, anticipating treatment response, and maintaining consistency with prior decisions and evolving clinical states. These outputs must also cover multiple rubric components rather than identify a single finding. The broader specialty and procedural diversity of long-horizon reports further expands the decision space, explaining why complete treatment and follow-up planning remains difficult.

Finding 2. Perception errors degrade end-to-end clinical performance, especially diagnosis. Table 2 evaluates model-perception trajectories, exposing the perception-to-decision pipeline. Performance declines across all task families, with the most severe diagnostic degradation on patients. Diagnosis depends on integrating complementary findings across modalities. Even a few omitted, incorrect, or inconsistent observations can break multiple evidence links and distort the clinical state used by later stages. Treatment and follow-up remain weak because downstream reasoning cannot recover evidence that was never extracted correctly, while erroneous observations can influence subsequent decisions. Plausible but generic plans may retain partial rubric credit, yet they do not restore patient-specific evidence grounding. This cascade shows that the bottleneck in end-to-end longitudinal care is constructing a faithful clinical representation from dense multimodal inputs.

4.3 Ablation Study

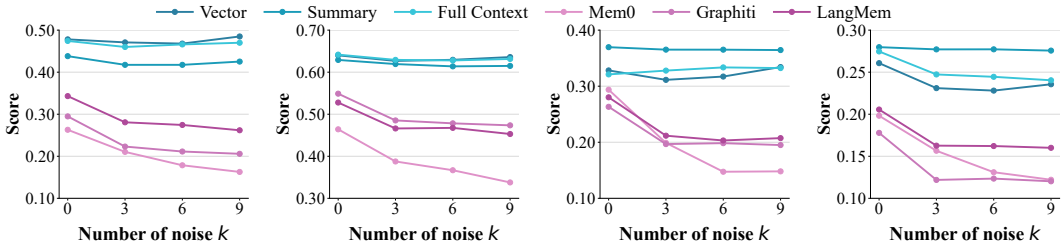
Finding 3. The standard-to-model-perception gap is concentrated in diagnosis. Table 3 reports model-perception trajectory quality. The results are consistent with diagnostic reasoning's dependence on image interpretation. Because this intervention changes only image-grounded reference responses while holding non-image observations and hidden decision records fixed, it isolates how visual errors enter memory and propagate to downstream queries. The larger patient-side gap is consistent with its denser set of images and measurements, where one early perception error can invalidate multiple later evidence links. Treatment and follow-up rubrics decline less because non-image history preserves partial planning credit. The alignment between report-diagnosis perception quality and downstream diagnostic performance links observation quality to later reasoning.



(a) D-Acc **(b)** D-ERS **(c)** T-Rubric **(d)** T-ERS

Figure 4: **Missing-modality ablation on OralFlow-Cohort.** Each panel reports OralGPT-Omni-7B with three memory methods across trajectories without DP, MVSM, TMJ, or X-Ray. **D-**, **T-**, and **F-** denote diagnosis, treatment, and follow-up, respectively.

Finding 4. Missing MVSM causes the broadest degradation, whereas missing X-Ray most weakens treatment evidence grounding. Figure 4 shows that MVSM is most consequential for comprehensive treatment planning, whereas X-Ray is most important for grounding treatment evidence. TMJ has the smallest diagnostic effect because fewer diagnosis tasks directly require TMJ evidence. These complementary effects show why no single modality can substitute for the multimodal record: different sources support different stages of the clinical workflow.



(a) D-Acc **(b)** D-ERS **(c)** T-Rubric **(d)** T-ERS

Figure 5: **Noise-trajectory ablation on OralFlow-Cohort.** Performance of OralGPT-Omni-7B is reported as the number of injected noise pairs $k \in \{0, 3, 6, 9\}$ increases for six memory methods.

Finding 5. Noise robustness depends on memory update and retrieval rather than context size. Figure 5 shows that Mem0 is more sensitive to distractors than Summary Memory, whereas Vector and Full Context remain stable as noise increases. This suggests that item-level writing and retrieval can admit irrelevant observations that compete with clinical evidence, while recursive compression can suppress repeated distractors through a stage-level state. Robust longitudinal memory requires selective evidence control, not merely a larger context window or structured memory.

4.4 Case Study

We further illustrate four representative standard VQA trajectories, including two from OralFlow-Cohort and two from OralFlow-Literature, with visualizations of their temporally directed evidence graphs. These graphs highlight how atomic clinical observations are connected through cross-stage evidence relations and aggregated across visits, supporting subsequent treatment decisions and follow-up assessments. The associated benchmark questions and answers are highlighted for each trajectory. Full case details are provided in *Appendix G*.

5 Conclusion

We introduce OralFlow-Bench, the first benchmark for longitudinal multimodal dental workflows. We evaluate six MLLMs with six memory methods across five task categories for long-context clinical evidence in dental workflows. Verified observations support diagnosis more strongly than treatment and follow-up planning, revealing the gap between recognition and clinical decision-making. In end-to-end evaluation, performance degrades across all task families, with errors concentrated in diagnosis, particularly for image-dense patient records. Missing-modality analyses reveal complementary roles of MVSM and X-Ray, while noise ablations show that robustness to distractors depends on selective memory updating and retrieval. This work advances longitudinal dental AI through multimodal evaluation across clinical workflows and reliable clinical decision-making.

AI USE STATEMENT

In this work, we used AI tools to assist with benchmark construction and language editing, including grammar checking, sentence refinement, and improving the clarity and readability of the manuscript. We also used generative AI tools to assist with parts of the code implementation and debugging.

All AI-assisted content was manually reviewed and revised by the authors. All benchmark content was further reviewed by professional dentists to ensure its clinical accuracy and relevance. The code was executed and tested, and all experimental results and quantitative claims in the manuscript were verified against the actual experiments and records. We take full responsibility for the final content of this work, including all text and benchmark content developed with the assistance of AI tools.

ETHICS STATEMENT

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the XXX hospital (IRB No. XX-XX-XX-X). The clinical data used in this study were de-identified prior to research use, and all data were collected and processed in accordance with applicable ethical requirements to protect patient privacy and confidentiality.

REFERENCE

[1] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138. URL https://doi.org/10.1056/AIoa2300138.

[2] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024. doi: 10.1038/s41591-024-03185-2. URL https://doi.org/10.1038/s41591-024-03185-2.

[3] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. URL https://arxiv.org/abs/2306.00890.

[4] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: A multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. URL https://arxiv.org/abs/2307.15189.

[5] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *arXiv preprint arXiv:2102.09542*, 2021. URL https://arxiv.org/abs/2102.09542.

[6] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. URL https://arxiv.org/abs/2305.10415.

[7] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

[8] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *arXiv preprint arXiv:2404.15155*, 2024. URL https://arxiv.org/abs/2404.15155.

[9] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025. URL https://arxiv.org/abs/2502.02673.

[10] Hao Ding, Jiamin Wu, Wuyuan Zhao, Jukka P. Matinlinna, Michael F. Burrow, and James K. H. Tsoi. Artificial intelligence in dentistry–a review. *Frontiers in Dental Medicine*, 4:1085251, 2023. doi: 10.3389/fdmed.2023.1085251. URL https://doi.org/10.3389/fdmed.2023.1085251.

[11] Ibrahim Ethem Hamamci, Sezgin Er, Omer Faruk Durugol, Gulsade Rabia Cakmak, Ezequiel de la Rosa, Enis Simsar, Atif Emre Yuksel, Sadullah Gultekin, Serife Damla Ozdemir, Kaiyuan Yang, Mehmet Berke Isler, Mustafa Salih Gucez, Shenxiao Mei, Chenglong Ma, Feihong Shen, Kaidi Shen, Huikai Wu, Han Wu, Lanzhanju Mei, Zhiming Cui, Niels van Nistelrooij, Khalid El Ghoul, Steven Kempers, Tong Xi, Shankeeth Vinayahalingam, Kyoungyeon Choi, Jaewon Shin, Eunyi Lyou, Lanshan He, Yusheng Liu, Lisheng Wang, Tudor Dascalu, Shaqayeq Ramezanzade, Azam Bakhshandeh, Lars Bjørndal, Bulat Ibragimov, Hongwei Bran Li, Sarthak Pati, Bernd Stadlinger, Albert Mehl, Mehmet Kemal Ozdemir, Mustafa Gundogar, and Bjoern Menze. Dentex: Dental enumeration and tooth pathosis detection benchmark for panoramic x-ray. *arXiv preprint arXiv:2305.19112*, 2023. URL https://arxiv.org/abs/2305.19112.

[12] Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lizhuo Lin, Jinrong Yang, Qi Yong H. Ai, Lun M. Wong, Hao Tang, and Kuo Feng Hung. Towards better dental ai: A multimodal benchmark and instruction dataset for panoramic x-ray analysis. *arXiv preprint arXiv:2509.09254*, 2025. URL https://arxiv.org/abs/2509.09254.

[13] Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, Qiankun Li, Hao Tang, James Kit-Hon Tsoi, Linlin Shen, and Kuo Feng Hung. Oralgpt-omni: A versatile dental multimodal large language model. *arXiv preprint arXiv:2511.22055*, 2025. URL https://arxiv.org/abs/2511.22055.

[14] Yuxuan Fan, Jing Hao, Hong Chen, Jiahao Bao, Yihua Shao, Yuci Liang, Kuo Feng Hung, and Hao Tang. Oralgpt-plus: Learning to use visual tools via reinforcement learning for panoramic x-ray analysis. *arXiv preprint arXiv:2603.06366*, 2026. URL https://arxiv.org/abs/2603.06366.

[15] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

[16] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. EHRXQA: A multi-modal question answering dataset for electronic health records with chest x-ray images. In *Advances in Neural Information Processing Systems*, volume 36, pages 3867–3880, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0c007ebef1d11fd48da6ce4f54687db6-Abstract.html.

[17] Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.11805*, 2021. URL https://arxiv.org/abs/2109.11805.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen tau Yih, Tim Rockt"aschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020. URL https://arxiv.org/abs/2005.11401.

[19] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023. URL https://arxiv.org/abs/2305.10250.

[20] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

[21] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[22] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

[23] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. URL https://arxiv.org/abs/2404.06654.

[24] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024. URL https://arxiv.org/abs/2409.18839.

[25] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fUtxNAKPDV.

[26] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.14463.

[27] Steven Bethard, Guergana Savova, Wei Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1052–1062, 2016. doi: 10.18653/v1/S16-1165. URL https://aclanthology.org/S16-1165/.

[28] John Mongan, Louis Moy, and Charles E. Kahn. CLAIM: A checklist for artificial intelligence in medical imaging. *Radiology: Artificial Intelligence*, 2(2):e200029, 2020. doi: 10.1148/ryai.2020200029. URL https://doi.org/10.1148/ryai.2020200029.

[29] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

[30] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023. URL https://arxiv.org/abs/2306.05685.

[31] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, 2023. doi: 10.18653/v1/2023.conll-1.21. URL https://aclanthology.org/2023.conll-1.21/.

[32] OpenAI. OpenAI GPT-5 System Card. https://openai.com/index/gpt-5-system-card/, 2025. Accessed: 2026-08-28.

[33] Google DeepMind. Gemini 3.8 flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, 2026. Accessed: 2026-09-10.

[34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

[35] Kimi Team. Kimi-k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

[36] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Rame, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Leonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

[37] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. URL https://arxiv.org/abs/2005.14165.

[38] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025. URL https://arxiv.org/abs/2504.19413.

[39] LangChain AI. Langmem. https://github.com/langchain-ai/langmem, 2025. Accessed: 2026-08-28.

[40] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

[41] Jason J. Lau, S. Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251. URL https://doi.org/10.1038/sdata.2018.251.

[42] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. doi: 10.48550/arXiv.2003.10286. URL https://arxiv.org/abs/2003.10286.

[43] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6: 317, 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

[44] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

[45] Jing Hao, Andrew Nalley, Andy Wai Kan Yeung, Ray Tanaka, Qi Yong H. Ai, Walter Yu Hang Lam, Zhiyi Shan, Yiu Yan Leung, Abeer AlHadidi, Michael M. Bornstein, James Kit Hon Tsoi, Colman McGrath, and Kuo Feng Hung. Characteristics, licensing, and ethical considerations of openly accessible oral-maxillofacial imaging datasets: A systematic review. *npj Digital Medicine*, 8:412, 2025. doi: 10.1038/s41746-025-01818-5.

[46] J. Bao, X. Zhang, S. Xiang, H. Liu, M. Cheng, Y. Yang, X. Huang, W. Xiang, W. Cui, H. C. Lai, S. Huang, Y. Wang, D. Qian, and H. Yu. Deep learning–based facial and skeletal transformations for surgical planning. *Journal of Dental Research*, 103(8):809–819, 2024. doi: 10.1177/00220345241253186. URL https://doi.org/10.1177/00220345241253186.

[47] Jiahao Bao, Huazhen Liu, Yu Zhuang, Leran Tao, Xinyu Xu, Yongtao Shi, Mengjia Cheng, Yiming Wang, Congshuang Ku, Ting Zeng, Yilang Du, Siyi Chen, Shunyao Shen, Suncheng Xiang, and Hongbo Yu. PhysSFI-Net: Physics-informed geometric learning of skeletal and facial interactions for orthognathic surgical outcome prediction. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-03075-6. URL https://doi.org/10.1038/s41746-026-03075-6.

[48] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, 2024. URL https://aclanthology.org/2024.acl-long.747/.

[49] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024. URL https://arxiv.org/abs/2410.10813.

[50] Jing Hao, Siyuan Dai, Yongxin Zhang, Yuci Liang, Jiamin Wu, Jiahao Bao, Yuxuan Fan, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Ming Hu, Liang Zhan, James Kit Hon Tsoi, Linlin Shen, Junjun He, and Kuo Feng Hung. Oralagent: Integrating reasoning, tools, and knowledge for interactive dental image analysis. *arXiv preprint arXiv:2605.27378*, 2026. URL https://arxiv.org/abs/2605.27378.

[51] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

[52] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

[53] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

[54] Howard H Wang, Chien Lam, and Srinivas R Myneni. Resolution of a cystic endodontic-periodontal lesion utilizing combined apicoectomy and guided tissue regeneration: a case report. *Clinical Advances in Periodontics*, 12(2):94–100, 2022.

[55] Daniel D Gober and Matthew J Fien. Flapless extraction socket healing around an immediate implant placed into a mandibular molar site without the use of regenerative materials: A case report. *International Journal of Periodontics & Restorative Dentistry*, 36(2):219, 2016.

Appendix

A Related Works **17**

A.1 Multimodal Dental MLLMs and Benchmarks 17

A.2 Longitudinal Memory for LLMs 17

B Details of OralFlow-Bench Construction **17**

B.1 Report-Timeline Canonicalization and Leakage Guards 17

B.2 Atomic Evidence and Edge Quality Control 17

B.3 Task Categories and Records 18

B.4 Task Acceptance and Evidence Scoping 18

C OralFlow-Bench Analysis and Visualization **19**

C.1 Detailed Benchmark Composition 19

C.2 Trajectory Vocabulary 19

C.3 Expert Quality Review 20

D Implementation Details **22**

D.1 Model Serving and Memory Methods. 22

D.2 Model-Perception Metrics 22

D.3 Noise Pool Construction and Examples. 22

E Effectiveness and Stability of Evaluation **23**

E.1 Effectiveness 23

E.2 Stability 24

F Memory Efficiency and Operational Analysis **24**

F.1 Measurement Protocol. 25

F.2 Operational Profile 25

F.3 Efficiency Frontier and Method-Specific Trade-offs 26

G Case Studies **27**

G.1 Case 1: Severe skeletal Class III with mandibular excess 28

G.2 Case 2: Skeletal Class II with mandibular retrognathia and staged rehabilitation 35

G.3 Case 3: Endodontic-periodontal lesion with regenerative surgery 42

G.4 Case 4: Immediate implant placement with open healing 46

H Prompt Templates **52**

A Related Works

A.1 Multimodal Dental MLLMs and Benchmarks

Medical MLLMs have expanded visual–textual reasoning from image understanding to open-ended clinical interaction. Representative models and medical VQA benchmarks include LLaVA-Med [3], Med-Flamingo [4], MedGemma [36], VQA-RAD [41], PathVQA [42], SLAKE [5], and PMC-VQA [6]. These models and benchmarks establish image-grounded medical capability, but their evaluations are generally organized around static image–text pairs or isolated diagnostic prompts rather than end-to-end evaluation of an entire longitudinal oral-health trajectory from perception to clinical decision making [43, 44].

Dental research has begun to provide specialized imaging benchmarks. Available resources span openly accessible Oral-Maxillofacial imaging datasets [45], panoramic X-ray benchmarks for dental enumeration and pathology detection such as DENTEX [11], and multimodal dental benchmarks and instruction datasets including MMOral [12] and MMOral-Uni in OralGPT-Omni [13]. OralGPT-Plus [14] further studies visual tool use for panoramic X-ray analysis. These benchmarks and methods evaluate dental image detection, diagnosis, surgical planning [46, 47], or isolated multimodal tasks, rather than end-to-end evaluation of an entire longitudinal oral-health trajectory from perception to clinical decision making.

A.2 Longitudinal Memory for LLMs

Long-term memory systems for LLMs address limited context through several complementary mechanisms. Summary memory recursively compresses interaction histories [17], retrieval memory grounds generation in external text collections [18], and LLM memory systems include MemoryBank [19], MemGPT [20], Mem0 [38], LangMem [39], and Graphiti [40]. These systems motivate the memory methods evaluated in OralFlow-Bench, including full-context, summary, vector, and graph-based memory methods, but they do not specify a clinical protocol for when a decision can be observed, queried, or written back into memory.

Memory benchmarks have progressed from short-context recall to sustained interactive evaluation, such as LoCoMo [48] and LongMemEval [49]. Medical and dental LLM systems, including OralAgent [50], MDAgents [8], and MedRAX [9], add knowledge integration, collaboration, or tool-mediated clinical reasoning, but are evaluated on decision making or a specific imaging workflow rather than longitudinal, evidence-grounded memory under staged information release. Their limitations for this setting include the absence of an explicit answer-release protocol, limited evaluation of memory revision after later clinical evidence, and no benchmark-wide future-leakage control.

B Details of OralFlow-Bench Construction

B.1 Report-Timeline Canonicalization and Leakage Guards

Before rendering OralFlow-Literature, we apply deterministic structural checks to the extracted timeline. Whole-figure records are removed when their captions contain multiple dated panels that cannot be safely released as independent timepoints. Within a timepoint, repeated records for the same figure are merged, and observation records are ordered before evaluation records. The canonicalizer assigns each valid timepoint a unique ordinal stage identifier and enforces a nondecreasing progression from perception to treatment and follow-up. Treatment timepoints contain held-out evaluation records only, follow-up timepoints contain at least one visible observation. Image-grounded questions are additionally checked against captions and preceding text to reject caption-only questions, duplicate figure use, and premature figure references. When verification identifies high- or medium-severity issues, only the implicated contiguous timepoint span is repaired. Iteration stops after a clean verification result, a repeated candidate timeline, or the repair limit.

B.2 Atomic Evidence and Edge Quality Control

Atomic evidence is extracted independently from each visible observation turn. Each record retains its source-turn provenance, introduction stage, modality, fact type, clinical dimension, and normal-

ized attributes. The normalization preserves units, tooth identifiers, laterality, and negation, and assigns deterministic evidence identifiers from the case, stage, source turn, and normalized field. Cross-stage evidence edges are constructed through a hybrid procedure. Deterministic candidates connect repeated normalized attributes across successive observed stages, labeling equal values as *confirms* and differing repeated values as *remeasures*.

A separate proposal model adds candidate clinical associations, after which an independent reviewer classifies valid candidates as clinical support or contextual consistency. We reject self-links, unknown evidence identifiers, empty rationales, non-forward temporal links, and duplicate edges, then cap the retained reviewed candidates at 40 per case. Context-consistency edges remain in the graph for audit but are excluded from the evidence links exposed during task planning. The atomic-evidence and cross-stage evidence-edge prompts are shown in Figures 24–26.

B.3 Task Categories and Records

The task planner constructs each benchmark instance from the verified evidence graph and assigns it to one of five clinical abilities. For each case, the standard task allocation includes three modality-perception tasks, two longitudinal evidence-recall tasks, two cross-modal or cross-temporal reasoning tasks, and two memory-update tasks. Treatment, surgery, and follow-up planning tasks are generated separately because their answers contain multiple clinically meaningful components and are evaluated with source-specific rubrics. The five task categories are summarized in Table 4.

Table 4: OralFlow-Bench task categories.

Task category	**Clinical ability tested**
Modality perception	Reading findings directly visible or stated in one modality at one stage.
Longitudinal evidence recall	Retrieving earlier evidence when answering a later-stage question.
Cross-modal/-temporal reasoning	Integrating evidence from at least two modalities or timepoints into a supported clinical interpretation.
Memory update and correction	Using later evidence to confirm, refine, or correct the current clinical memory state.
Treatment and follow-up planning	Comprehensive decision tasks scored with source-specific rubrics.

The task records preserve the case identifier, task category, question, reference answer, selected evidence, ask stage, release stage, and scoring metadata. The task generator first selects temporally eligible evidence, then produces the question and answer, and finally applies independent validation for source support, temporal consistency, category assignment, and answer leakage. For a task q, these records explicitly identify the question, reference answer, evidence set, ask stage, and release stage used by the streaming evaluator. These fields are also the interfaces consumed by the streaming evaluator, which freezes answers before scoring.

B.4 Task Acceptance and Evidence Scoping

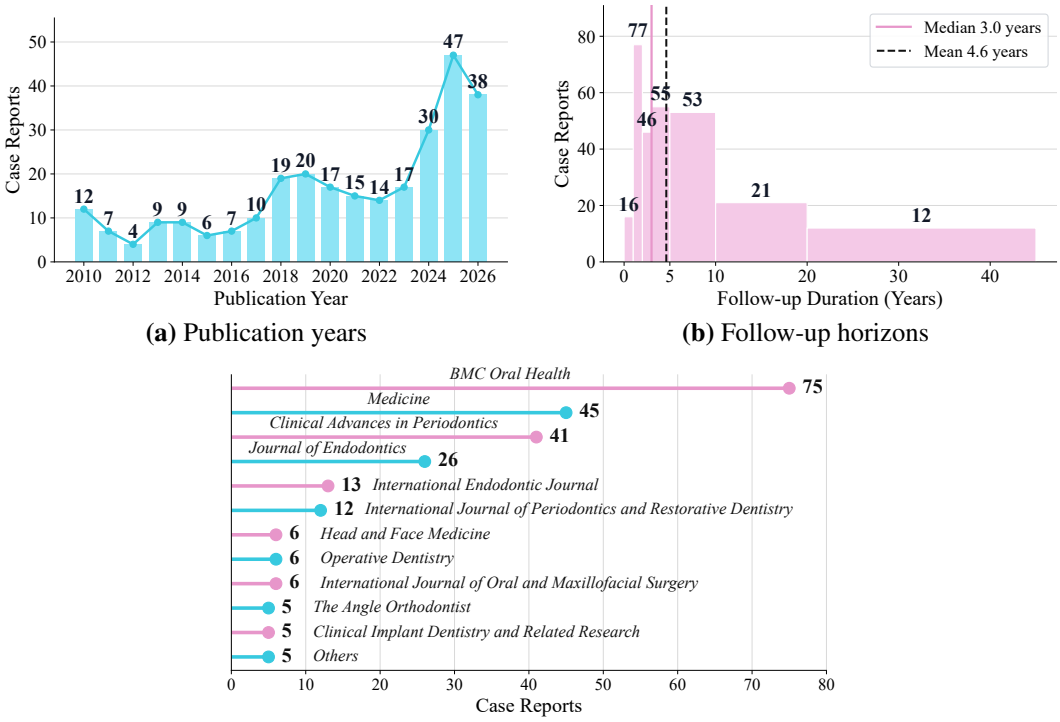
For each non-decision category, the planner oversamples candidate specifications and applies deterministic preflight checks before question generation. A candidate must select nonempty evidence available at its ask stage, satisfy the required number of stages, and, when required by the category, include evidence from an earlier stage. The generator and verifier then iterate for up to three rounds. Verifier feedback is returned to regenerate the question, and verifier-supplied corrections may replace the question or reference answer before acceptance. For source decision tasks, required evidence is selected only from the evidence subgraph observable at the ask stage. Invalid or empty evidence selections are retried with a constrained repair request that accepts only identifiers from that available catalog. This procedure makes the selected evidence set both a task-construction constraint and the fixed reference set for downstream evidence-recall scoring.

C ORALFLOW-BENCH ANALYSIS AND VISUALIZATION

C.1 DETAILED BENCHMARK COMPOSITION

OralFlow-Cohort contains 2,322 visible diagnostic stages and 387 held-out treatment stages. Its 10,326 assets comprise 2,673 facial photographs, 2,292 intraoral photographs, 1,131 radiographs, 570 cephalometric-analysis tables, 1,852 multiview skeletal-model images, and 1,808 TMJ/MRI images. The 6,958 tasks include 1,161 modality-perception, 774 longitudinal-recall, 774 cross-modal/-temporal reasoning, 774 memory-update, and 3,475 treatment or surgical-planning tasks. Due to the use of real patient data, OralFlow-Cohort is not publicly released.

OralFlow-Literature spans 11 dental specialties and contains 355 diagnostic-workup, 546 treatment, and 1,248 follow-up stages. Follow-up ranges from 0.5 to 43.0 years, with a median of 3.0 years. Its 5,591 tasks include 843 modality-perception, 560 longitudinal-recall, 560 cross-modal/-temporal reasoning, 560 memory-update, 1,996 treatment, and 1,072 follow-up-management tasks. The report figures are categorized using a pretrained OralCLIP model. Figure 6 summarizes the publication years, follow-up horizons, and journal distribution of the reports. The report subfields are listed in Table 5.



(a) Publication years **(b)** Follow-up horizons

(c) Top journals

Figure 6: Statistics of the OralFlow-Literature collection.

C.2 TRAJECTORY VOCABULARY

The stage-specific word clouds summarize the clinical content of the standard trajectories across diagnosis, treatment, and follow-up. Figure 7 presents the three stages separately.

- **Diagnostic** stages are dominated by terms describing dentofacial morphology and imaging findings, such as mandibular, maxillary, facial, skeletal, asymmetry, and panoramic.
- **Treatment** stages emphasize treatment planning and intervention, with frequent terms including orthodontic, surgical, orthognathic, occlusal, correction, and osteotomy.
- **Follow-up** stages are characterized by temporal and outcome-related terms, including months, year, treatment, tooth, bone, implant, healing, and recurrence.

Table 5: Oral and dental subfields represented in OralFlow-Literature.

No.	Subfields	# Cases	Description
1	Periodontics	58	Diseases and treatments involving the periodontium, including periodontal inflammation, bone loss, and related interventions.
2	Oral Oncology and Pathology	45	Oral neoplasms, mucosal lesions, histopathological findings, and diagnostic or therapeutic management of oral disease.
3	Implantology	40	Implant placement, implant-supported rehabilitation, peri-implant conditions, and associated surgical or restorative procedures.
4	Endodontics	38	Diseases of the dental pulp and periapical tissues, root canal treatment, and endodontic complications.
5	General Dentistry and Oral Health	27	General oral conditions, preventive care, common dental diseases, and broad oral-health management.
6	Orthodontics	25	Malocclusion, tooth movement, dentofacial growth, and orthodontic diagnosis and treatment.
7	Oral Medicine and Pathology	19	Non-surgical diagnosis and management of oral mucosal, inflammatory, infectious, and systemic manifestations.
8	Oral and Maxillofacial Surgery*	12	Surgical management of dentofacial deformities, trauma, jaw lesions, and other oral and maxillofacial conditions.
9	Restorative Dentistry	8	Restoration of damaged or missing tooth structure using direct or indirect restorative procedures.
10	Prosthodontics	5	Replacement of missing teeth and oral structures with fixed, removable, or implant-supported prostheses.
11	Pediatric Dentistry	4	Oral health, dental diseases, and treatment considerations specific to children and adolescents.

* *All OralFlow-Cohort Cases belong to this subfield.*

This progression reflects the intended clinical workflow from observation and diagnosis to treatment planning and longitudinal outcome assessment.



(a) The word cloud for diagnosis

(b) The word cloud for treatment

(c) The word cloud for follow-up

Figure 7: **Stage-specific vocabulary in standard OralFlow trajectories.** Each panel shows the word cloud for a different clinical stage, with word size reflecting frequency.

C.3 Expert Quality Review

A dental team with over five years of clinical experience reviewed the finalized package of 795 tasks sampled from 20 OralFlow-Cohort cases and 20 OralFlow-Literature cases. The package contains 360 patient tasks and 435 report tasks, and every task received a completed review. Each task is rated on a five-point scale, from 1 (unacceptable) to 5 (excellent).

The reported dimensions are summarized in Table 6. All tasks are assessed for Question Design, Temporal Validity and Answerability, Gold Answer Quality, and Evidence Support. Non-decision tasks additionally receive a Task-Type Fit rating, whereas treatment and follow-up decision tasks receive a Rubric Quality rating.

Table 6: Reported dimensions in the expert quality review.

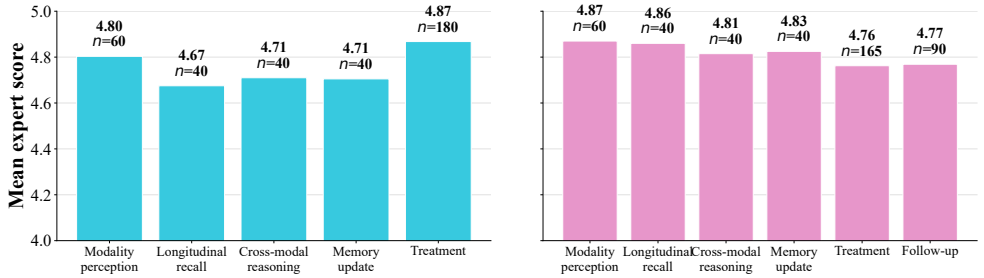
No.	Dimension	Evaluation focus
1	Question Design	Clinical relevance, clarity, answerability, and non-redundancy of the question.
2	Temporal Validity and Answerability	Appropriate clinical ordering, evidence availability at the ask stage, and absence of future-answer leakage.
3	Gold Answer Quality	Correctness, necessary qualifiers, and consistency with the source-grounded clinical decision or observation.
4	Evidence Support	Sufficiency, atomicity, temporal availability, and direction of the selected evidence.
5	Task-Type Fit	Alignment between the question and its assigned non-decision task category.
6	Rubric Quality	Clinical coverage, criterion specificity, and weighting of decision-task rubric criteria.

The dental team reviewed the question, reference answer, supporting evidence, available timeline, and, for decision tasks, the rubric. A task could be marked as pass, pass after minor revision, redo, or exclude. Critical factual errors, future leakage, unsupported answers, image mismatches, and identifying patient information trigger revision, regeneration, or exclusion. Table 7 reports the finalized review coverage, score distribution, and recorded verdicts.

Table 7: Summary of finalized expert quality review.

Subset	Mean ± Std.	Pass	Minor revision	Redo	Exclude	Unspecified
OralFlow-Cohort	4.80±0.34	284	55	2	1	18
OralFlow-Literature	4.80±0.30	405	24	0	0	6

Figure 8 reports the mean task-level review score for each evaluated task category. It shows that the review covers all task categories available in each subset, including treatment for both subsets and follow-up for OralFlow-Literature. The consistently high scores across task categories support the clinical quality, temporal validity, and overall reasonableness of OralFlow-Bench.



(a) OralFlow-Cohort **(b)** OralFlow-Literature

Figure 8: Task-category expert-quality scores. Bars show mean task-level scores, and labels report the number of reviewed tasks in each category. Follow-up is available only in OralFlow-Literature.

D Implementation Details

D.1 Model Serving and Memory Methods

In experiments, open-weight models are served with vLLM [51], whereas closed models are queried through their official APIs. Mem0[2], LangMem[3], and Graphiti[4] are evaluated from their open-source implementations. Full Context stores all released stage texts without compression. Summary Memory incrementally rewrites a structured clinical summary with Qwen2.5-7B-Instruct [52], capped at 1,024 output tokens per update. Vector Memory encodes released stage texts and task queries with Qwen3-Embedding-0.6B [53] into 1,024-dimensional vectors, then retrieves the top eight entries by cosine similarity. Mem0 uses the same embedding model with a local Qdrant store. LangMem uses the same embedding model for retrieval and Qwen2.5-7B-Instruct [52] for memory updates. Graphiti stores temporally ordered clinical episodes in Neo4j and uses the same embedding model for retrieval. The key construction and evaluation hyperparameters are listed in Table 8.

Table 8: Key hyperparameters used by the OralFlow-Bench construction and evaluation.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Answer tokens	2,048	Treatment tokens	4,096
Judge tokens	2,048 / 4,096	Summary tokens	1,024
Top-k	8	Embedding dim.	1,024
Graph store	Neo4j	Graph edge cap	40
Noise pool	50	Noise levels	3, 6, 9
Noise seed	42	Planning iterations	3
Generation iterations	3	Validation iterations	3
Evidence selection attempts	2	Rubric attempts	2
Embedding model	Qwen3-Embedding-0.6B	Update model	Qwen2.5-7B-Instruct

D.2 Model-Perception Metrics

For each image-grounded VQA turn $(u_{t,i}, a_{t,i}, I_{t,i}) \in \mathcal{O}_t$, let $\mathcal{C}_{t,i}$ be the gold atomic-claim set from the reference answer $a_{t,i}$, $\widehat{\mathcal{C}}_{t,i}$ be the claims extracted from $\pi_\theta(u_{t,i}, I_{t,i})$, and $\mathcal{U}_{t,i} \subseteq \widehat{\mathcal{C}}_{t,i}$ be the claims judged unsupported by the image or source evidence. Claim-level precision, recall, F1, and hallucination control for the VQA turn are:

$$\mathrm{Prec.}(u_{t,i}, \pi_\theta) = \frac{|\widehat{\mathcal{C}}_{t,i} \cap \mathcal{C}_{t,i}|}{|\widehat{\mathcal{C}}_{t,i}|}, \quad \mathrm{Recall}(u_{t,i}, \pi_\theta) = \frac{|\widehat{\mathcal{C}}_{t,i} \cap \mathcal{C}_{t,i}|}{|\mathcal{C}_{t,i}|}, \tag{8}$$

$$\mathrm{F1}(u_{t,i}, \pi_\theta) = \frac{2\,\mathrm{Prec.}\,\mathrm{Recall}}{\mathrm{Prec.} + \mathrm{Recall}}, \quad \mathrm{Hallu.}(u_{t,i}, \pi_\theta) = 1 - \frac{|\mathcal{U}_{t,i}|}{|\widehat{\mathcal{C}}_{t,i}|}. \tag{9}$$

Precision measures the proportion of generated claims matching the gold atomic claims, whereas recall measures the proportion of gold claims recovered by the generated observation. Hallucination control measures the fraction of generated claims that are supported by the image or source evidence, so a larger value indicates fewer hallucinated observations before downstream memory evaluation. The multimodal perception prompts are shown in Figures 22–23.

D.3 Noise Pool Construction and Examples

For the noise-trajectory ablation, we use a fixed pool of 50 distractor VQA pairs, with 10 pairs in each of five categories. For every noisy trajectory, we select the required number of pairs without replacement and append them as observation turns. Modality-adjacent distractors are inserted into a compatible perception stage whenever one is available. All other distractors are distributed across non-empty perception stages. Table 9 shows one representative pair from each category.

[2] https://github.com/mem0ai/mem0

[3] https://github.com/langchain-ai/langmem

[4] https://github.com/getzep/graphiti

Table 9: **Examples from the fixed noise pool.** Each category contains 10 distractor VQA pairs.

No.	Category	Question	Response
1	Scheduling and billing logistics	Could we move my next appointment to a later week? Something came up at work.	Let me check the calendar and find a slot a week or two later, then I will send the new confirmation.
2	General lifestyle advice	I brush twice a day. Is that enough, or should I be doing more?	Twice a day is a solid routine. Add flossing once daily if possible. Together they cover most people's needs well.
3	Benign repeated history	Just to be clear, I do not have any other major health conditions.	Thanks for confirming. That is already noted from earlier, so we are all set on that front.
4	Postoperative concerns	I am worried about facial swelling after surgery.	Some swelling is normal and usually peaks in the first few days. Ice and keeping the head elevated can help.
5	Modality-adjacent distractors (FP)	Did the lighting in my facial photos turn out okay?	The lighting looked good and the photos came out clear. If needed, we can simply retake them.

E Effectiveness and Stability of Evaluation

We examine whether GPT-5-mini can serve as a reliable evaluator by comparing its judgments with those of an experienced dentist. We study both *effectiveness*, which measures agreement with human evaluation, and *stability*, which measures score variation across repeated judging runs. The answer-generation and answer-judging prompts are shown in Figures 27–30.

E.1 Effectiveness

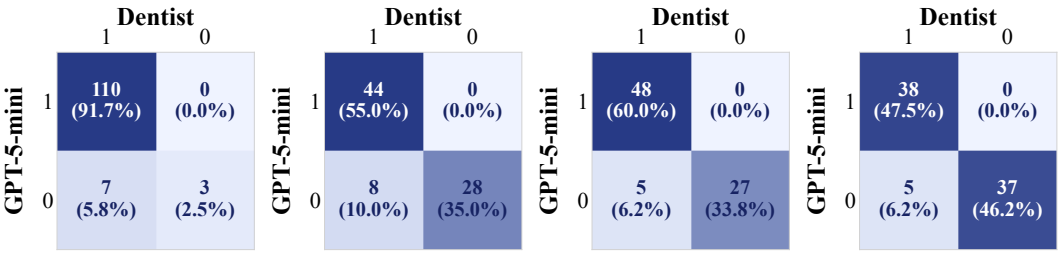
We sample 20 cases from each of OralFlow-Cohort and OralFlow-Literature using frozen standard-trajectory answers from OralGPT-Omni-7B with Vector Memory. Cases are selected to represent intermediate automatic performance (approximately 50%) while requiring both correctly and incorrectly scored tasks per case. The resulting package maintains an approximately balanced distribution of positive and negative tasks. For the four task categories, we convert dentist and GPT-5-mini [32] correctness judgments into binary labels. **Agreement** is the proportion of tasks for which the two evaluators assign the same label. Table 10 summarizes these four categories. Figure 9 shows the corresponding confusion matrices.

Table 10: Dentist–GPT-5-mini consistency for the four ordinary task categories. Higher Agreement indicates closer alignment with dentist judgments. Treatment and follow-up planning are assessed separately with continuous rubric scores in Table 11.

	OralFlow-Cohort		OralFlow-Literature	
Task category	**# Samples**	**Agreement (%)**	**# Samples**	**Agreement (%)**
Modality perception	60	93.33	60	95.00
Longitudinal evidence recall	40	97.50	40	82.50
Cross-modal/temporal reasoning	40	100.00	40	87.50
Memory update and correction	40	95.00	40	92.50
Treatment/Follow-up	180	-	211	-
Overall	180	96.11	180	90.00

For decision tasks, the dentist and GPT-5-mini independently score the same frozen OralGPT-Omni-7B answers using the task-specific 100-point rubrics. Table 11 reports their mean normalized rubric scores for patient treatment planning, report treatment planning and follow-up management.

Across the 360 ordinary annotated tasks, GPT-5-mini agrees with the dentist on 93.06% of binary decisions. Treatment and follow-up planning are evaluated separately using continuous task-specific rubrics. These results support the use of GPT-5-mini as a scalable evaluator while indicating that expert review remains particularly important for decision-task judgments.



(a) Modality perception **(b)** Longitudinal evidence recall **(c)** Cross-modal/-temporal reasoning **(d)** Memory update and correction

Figure 9: Dentist–GPT-5-mini confusion matrices for the four ordinary task categories.

Table 11: Effectiveness analysis for rubric-based decision tasks using OralGPT-Omni-7B answers. Scores are mean normalized task-level rubric scores. Δ is the difference between dentist and GPT-5-mini scores, and lower values indicate stronger alignment with expert assessment.

Evaluator	Cohort T-Rubric	Literature T-Rubric	Literature F-Rubric	Overall
Dentist	40.99	24.21	52.81	39.34
GPT-5-mini	36.00	26.07	50.39	37.49
$\Delta(\downarrow)$	4.99	-1.86	2.42	1.85

E.2 Stability

We apply GPT-5-mini [32] to the same frozen Vector Memory answers from Qwen3.8-Max [34], MedGemma-4B [36], and OralGPT-Omni-7B [13] five times under an identical evaluation prompt. The evaluation covers the same 40 cases as Appendix E.1. For the four ordinary task categories with binary correctness labels, **exact agreement** is the proportion of tasks whose five labels are identical, and **pairwise agreement** is the fraction of matching labels across all 10 pairs of repeated runs. For rubric-based decision tasks, we compute the mean score, sample standard deviation, and coefficient of variation (CV) over the five run-level aggregate rubric scores. Higher binary agreement and lower rubric-score variability indicate greater stability. Table 12 reports the aggregate metrics, and Figure 10 breaks them down by task type.

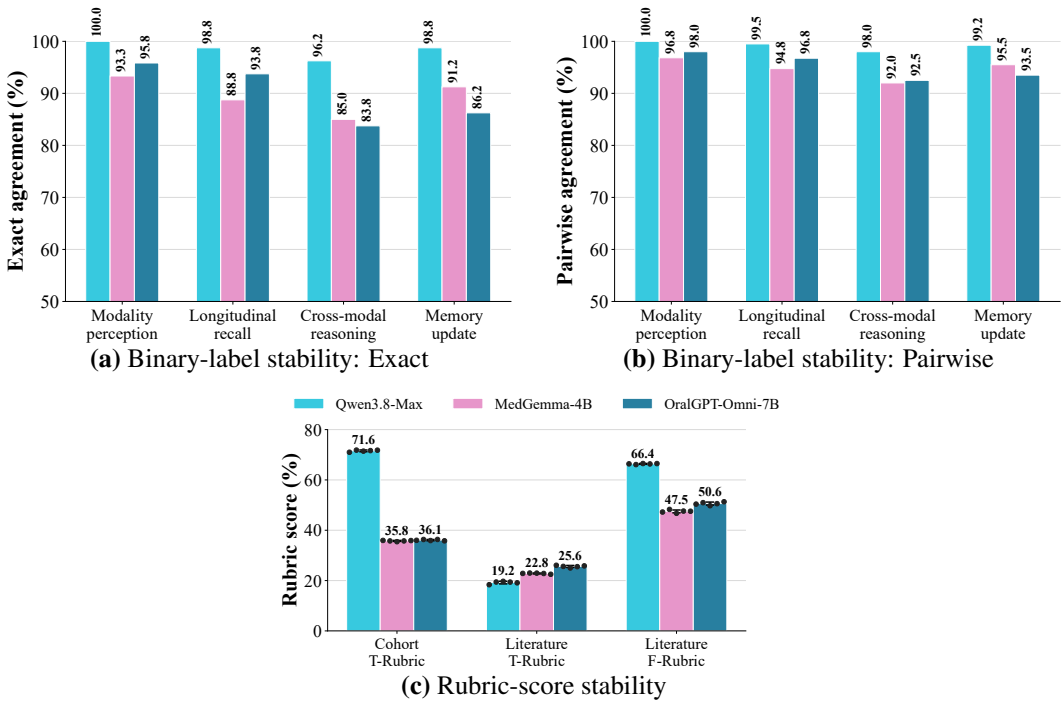
Table 12: **Stability metrics over five repeated GPT-5-mini evaluations.** Exact and pairwise agreement are computed over all 360 binary ordinary-task labels per model. Rubric statistics are computed from five run-level aggregate decision-task scores.

Model	Exact (%)	Pairwise (%)	Rubric Mean $\pm$ Std. (%)	CV (%)
Qwen3.8-Max [34]	98.61	99.28	50.78$\pm$0.32	0.63
MedGemma-4B [36]	90.00	95.00	32.69$\pm$0.15	0.45
OralGPT-Omni-7B [13]	90.56	95.50	34.39$\pm$0.25	0.73

The results indicate that GPT-5-mini provides stable evaluations for the models. Exact agreement ranges from 90.00% to 98.61%, and pairwise agreement ranges from 95.00% to 99.28% across binary ordinary tasks. Rubric-score variability is also low, with CV values below 1% for three models. These results support using GPT-5-mini as a reproducible evaluator for OralFlow-Bench, while the effectiveness analysis remains the complementary check of alignment with dentist judgments.

F Memory Efficiency and Operational Analysis

Beyond predictive quality, a longitudinal memory method must control the cost of incorporating new observations, retrieving task-relevant history, and maintaining a compact answer context. We therefore analyze the operational traces recorded for six memory methods.



(a) Binary-label stability: Exact

(b) Binary-label stability: Pairwise

(c) Rubric-score stability

Figure 10: **Stability over five repeated GPT-5-mini evaluations.** (a) Exact agreement and (b) pairwise agreement for each of the four binary ordinary task categories. (c) Mean rubric scores with standard-deviation error bars for Cohort T-Rubric, Literature T-Rubric, and Literature F-Rubric.

F.1 Measurement Protocol

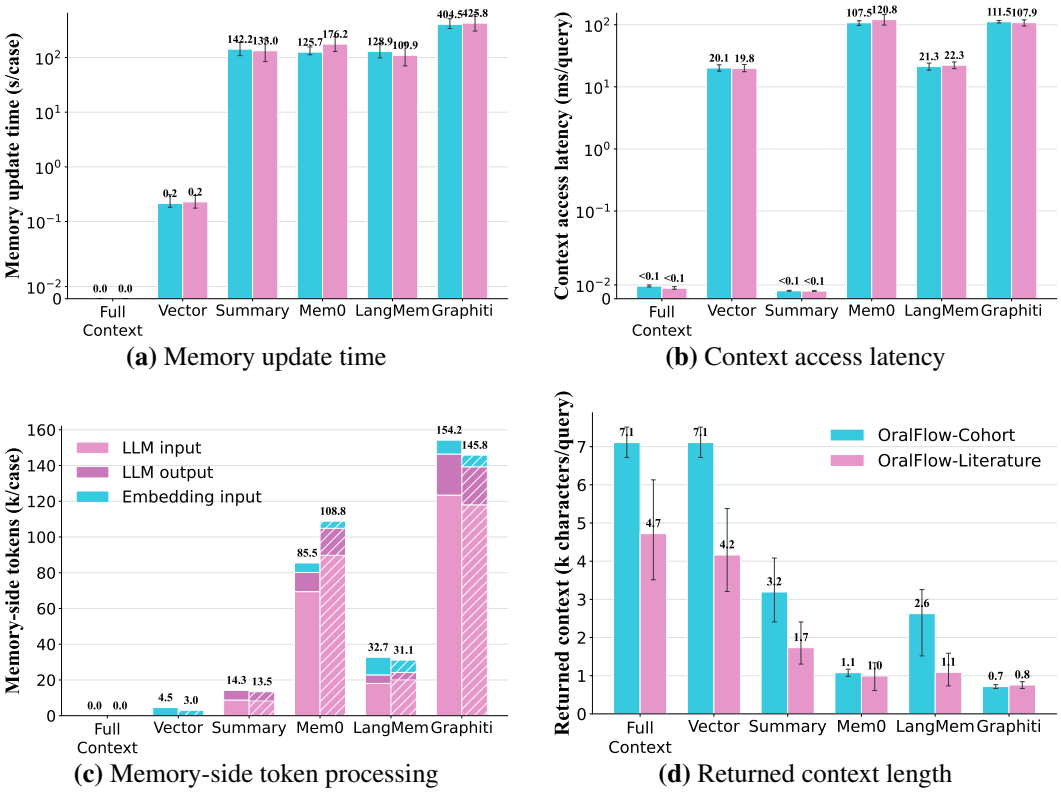
For every run, the evaluator records the number and wall-clock duration of memory writes and context accesses, memory-side LLM calls and tokens, embedding calls and tokens, and failed memory operations. We define per-case update time as the accumulated duration of all memory writes, and per-query access latency as total context-access time divided by the number of retrieval calls. Memory-side tokens comprise LLM input and output tokens used to construct or update memory plus embedding input tokens. They exclude the answer model and scoring judge. To quantify the downstream context burden without imposing a model-specific tokenizer, we compute the mean number of characters in the returned `memory_context` across all answered tasks in each run.

For the standard-trajectory analysis, we pool Qwen3.8-Max, MedGemma-4B, and OralGPT-Omni-7B, yielding 1,161 patient and 843 report runs per method. Wall-clock measurements include backend, database, API, retry, and queueing delays and should therefore be interpreted as operational measurements in our deployment rather than hardware-independent complexity constants.

F.2 Operational Profile

Figure 11 decomposes the same measurements into update time, context-access latency, memory-side token processing, and returned context size. The methods occupy sharply different operating regimes. Full Context performs only in-process concatenation and consequently has negligible measured memory-layer latency and no memory-side model calls, but it returns the longest context. Vector Memory adds only embedding and similarity-search overhead. Its update time is about 0.2 seconds per case and its access latency is approximately 20 milliseconds per query. In contrast, generative and structured memory methods move computation to ingestion. Summary, Mem0, LangMem, and Graphiti require roughly 110–426 seconds per case for memory updates, while Summary makes subsequent context access effectively free because it reads a maintained summary directly.

The token panel in Figure 11 explains the update-time gap. Graphiti processes the largest memory-side token volume, about 154.2k tokens per patient run and 145.8k per report run when LLM and



(a) Memory update time

(b) Context access latency

(c) Memory-side token processing

(d) Returned context length

Figure 11: **Operational profile of six memory methods.** Update time and context-access latency use symmetric-log axes. Solid and hatched bars denote OralFlow-Cohort and OralFlow-Literature.

embedding inputs and outputs are combined. This pattern is consistent with repeated graph extraction and temporal episode construction. Mem0 is the next most token-intensive memory method. Summary is substantially lighter, whereas LangMem occupies an intermediate regime. Vector requires only 4.5k and 3.0k embedding tokens and no memory LLM calls.

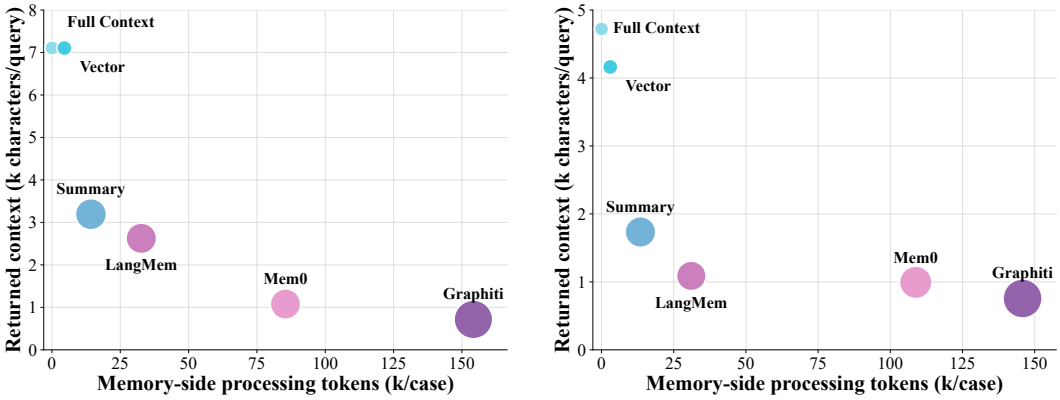
Returned context reveals a complementary trade-off. On patient trajectories, Full Context and Vector both return about 7.11k characters per query. Thus, with top-8 retrieval and relatively few released stage chunks, Vector reduces memory-method computation but does not necessarily reduce the text presented to the answer model. Summary approximately halves this burden to 3.19k characters, LangMem returns 2.63k, and Mem0 and Graphiti compress it to 1.08k and 0.72k. The same ordering broadly holds for reports. Compression is therefore not free: the shortest contexts are produced by the memory methods with the largest construction or retrieval overhead.

F.3 Efficiency Frontier and Method-Specific Trade-offs

Figure 12 jointly shows memory-side token processing, returned context, and update time. No method dominates all three axes. The plot instead exposes a continuum between *deferred computation*, where history is passed onward with little preprocessing, and *eager structuring*, where substantial work is performed before answering to obtain a compact context.

Full Context. Full Context is deterministic and incurs almost no measured memory-layer computation, making it a strong systems baseline. Its apparent efficiency is incomplete, however: it shifts the uncompressed history to the answer model and returns the largest contexts, about 7.11k characters for patients and 4.72k for reports.

Vector Memory. Vector provides the lowest-overhead learned retrieval path: no memory LLM is required and context access remains near 20 ms/query. Its limitation in this setting is weak compression. The patient context is identical in size to Full Context. Top-8 retrieval is therefore attractive



(a) OralFlow-Cohort **(b)** OralFlow-Literature

Figure 12: **Memory efficiency frontier on standard trajectories.** The horizontal axis is memory-side processing tokens per case, including LLM input, LLM output, and embedding input. Lower-left points require less memory-side processing and return shorter memory contexts.

for latency and implementation simplicity, but its context benefit depends on trajectories containing more candidate chunks than the retrieval budget.

Summary Memory. Summary exhibits a write-heavy/read-light profile. It spends 133–142 seconds per case updating a generated state, but context access is an in-memory read and its LLM token use is the lowest among generative memory methods. The resulting context is 55% shorter than Full Context for patients and 63% shorter for reports. This design is suitable when one memory state serves many future queries, allowing ingestion cost to be amortized. Its structural limitation is irreversible compression: details omitted during an update cannot be recovered by later retrieval.

Mem0. Mem0 achieves compact contexts, approximately 1k characters per query, and lower update time than Graphiti, but it incurs high memory-LLM input volume and approximately 108–121 ms/query retrieval latency. Mem0 therefore offers aggressive compression with a relatively simple retrieved representation, but depends on multiple LLM, embedding, and vector-store components whose latency accumulates.

LangMem. LangMem provides the most balanced learned-memory profile in these experiments. Its access latency, 21–22 ms/query, is close to Vector, while returned context is materially smaller. Its update cost and token use exceed Summary but remain far below Graphiti. This middle position makes LangMem attractive when moderate semantic consolidation is preferred over either raw retrieval or maximal compression.

Graphiti. Graphiti returns the most compact context, 0.72k characters for patients and 0.76k for reports, but pays the largest ingestion cost, 405–426 seconds per case, the largest memory-side token volume, and roughly 108–111 ms/query access latency. Its temporal graph construction is thus an eager investment rather than a free retrieval improvement. Such a design may be justified when a persistent structured memory is reused across many interactions.

G Case Studies

We present four representative trajectories: (1) Section G.1 shows a severe skeletal Class III presentation. (2) Section G.2 follows a skeletal Class II presentation with mandibular retrognathia through TMJ stabilization and staged orthodontic-orthognathic rehabilitation. (3) Section G.3 follows an endodontic-periodontal lesion from diagnosis through regenerative surgery and recovery [54]. (4) Section G.4 traces implant placement, open healing, restoration, and long-term follow-up [55].

For clarity, **magenta** boxes show the standard trajectory VQA turns containing multimodal figures, whereas **blue** boxes show benchmark evaluation questions and their corresponding answers.

G.1 Case 1: Severe skeletal Class III with mandibular excess

Clinical synopsis

A 20-year-old woman with a long-standing mandibular-protrusion concern. Clinical occlusion and cephalometry converge on severe skeletal Class III driven mainly by mandibular excess.

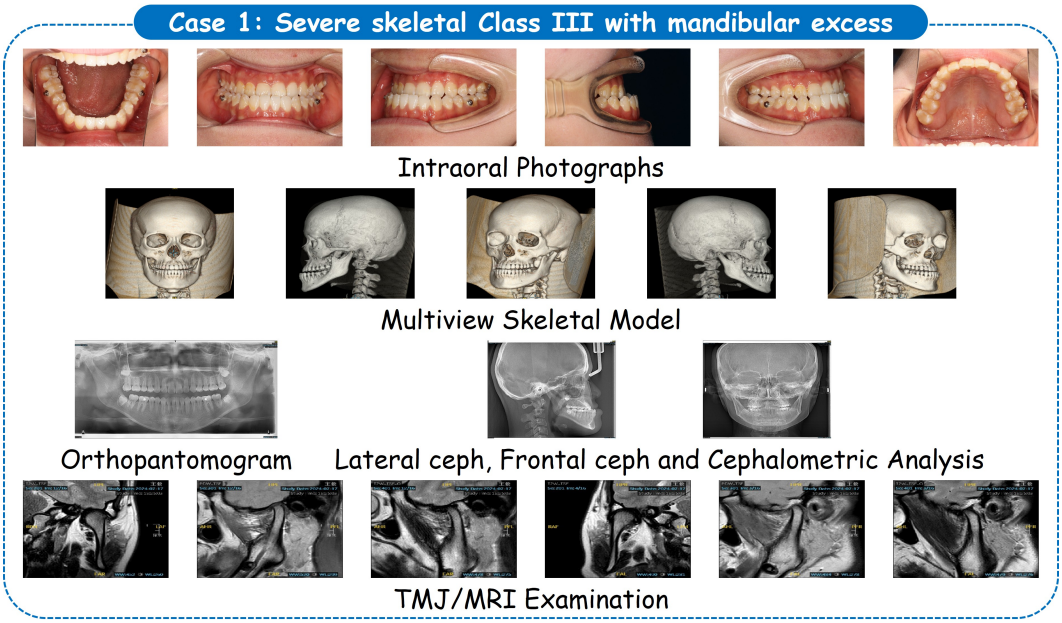


Figure 13: De-identified VQA image panel for Case 1. Facial photographs are intentionally omitted.

G.1.1 Standard VQA trajectory

Stage: S0_PROFILE (Clinical stage: perception)

Q1: Please provide your basic information, including age, gender, height, etc.

A1: She is a 20-year-old female.

Q2: What is the primary concern of the patient?

A2: The patient's primary concern is mandibular protrusion, which has been present for 10 years.

Q3: What is the patient's past medical or surgical history?

A3: The patient has a history of orthodontic treatment from August 2023 to April 2024.

Stage: S1_FP (Clinical stage: perception)

Q4: What objective findings can be observed from the facial photographs?

A4: The facial photographs show an asymmetric facial structure with the chin deviated 2 mm to the right, no upper incisor display at rest, no gummy smile, an upper lip length of 20 mm, a concave profile, and an obtuse nasolabial angle.

Stage: S2_DP (Clinical stage: perception)

Q5: What are the occlusal and dental findings based on the intraoral examination?

A5: Intraoral examination shows a Class III molar relationship bilaterally, with a reverse overjet of 6 mm and an anterior open bite of 1 mm. The upper dental midline is central, while the lower

dental midline is shifted 1 mm to the right. There is also occlusal canting, with the left side 2 mm higher.

Stage: S3_XR_XLA (Clinical stage: perception)

Q6: Based on the lateral and posteroanterior cephalometric radiographs and measurements, what are the key dento-skeletal characteristics of this patient?

A6: Key findings

Sagittal (AP)

- Severe skeletal Class III pattern: ANB $\approx -7^\circ$, Wits $\approx$ –14.8 mm.
- Maxilla positioned close to average relative to the cranial base (SNA $\approx 82^\circ$).
- Mandible markedly prognathic relative to the cranial base (SNB $\approx 89^\circ$; FH–NPo $\approx 97^\circ$) with a very prominent chin (NA–APo markedly negative).

Vertical

- Hyperdivergent tendency with increased lower anterior facial height: ANS–Me/N–Me $\approx$ 58% ($\uparrow$).
- Mandibular plane on the steeper side (MP–SN $\approx 36.6^\circ$, FMA $\approx 28^\circ$).
- Reduced posterior facial height/short ramus: S–Go/N–Me $\approx$ 59% ($\downarrow$), Co–Go $\approx$ 52 mm ($\downarrow$).

Dental (inferred from films/compensation)

- Class III dental relationship with an anterior crossbite/negative overjet.
- Typical dentoalveolar compensation: maxillary incisors proclined and mandibular incisors retroclined/uprighted.

Transverse (PA ceph)

- No gross skeletal cant; possible mild mandibular asymmetry/menton deviation (small).

Soft tissue profile

- Concave profile with a prominent lower lip/chin and relatively retrusive midface.

Overall impression: a patient with a severe skeletal Class III primarily due to mandibular excess, accompanied by a vertical (hyperdivergent) pattern with increased lower facial height and reduced posterior facial height, plus dentoalveolar compensations for the Class III discrepancy.

Q7: Please generate a comprehensive panoramic x-ray report based on the provided panoramic x-ray image.

A7: Panoramic X-ray Report:

Image quality:

- Diagnostic panoramic radiograph of acceptable quality for overall assessment.

Dentition:

- Permanent dentition is largely present.
- No primary teeth considered and likely non-detected/unerupted third molar-related regions corresponding to:
  - Upper right third molar (18) present/impacted appearance.
  - Upper left third molar (28) present/impacted appearance.
  - Lower right third molar (48) present/impacted appearance.
  - Lower left third molar (38) present, with posterior mandibular third molar region visualized.
- Overall pattern is consistent with a near-complete permanent dentition including third molars.

Impactions / eruption status:

- Impacted teeth detected in the third molar regions:
  - 18: impacted.
  - 28: impacted.
  - 48: impacted.
- These correspond well with the visible radiographic appearance of unerupted/partially erupted third molars.

Restorative findings:

- No obvious radiopaque restorations were identified.
- No implant detected.
- No crown or root canal treatment detected by the tool set. Periapical / endodontic findings.
- No periapical lesion detected on general condition analysis.
- No granuloma, cyst, or abscess subclass detected.
- No root piece identified.
- No definite carious lesion detected by the tool.
- Bone loss segmentation did not identify a clear generalized bone loss pattern.
- Alveolar bone levels appear grossly preserved on this panoramic image.

Jaw structures / surrounding anatomy:

- Mandibular canals visualized bilaterally.
- Maxillary sinuses visualized bilaterally.
- No obvious abnormality flagged in these structures by the analysis tools.

Impression:

- Panoramic radiograph demonstrates near-complete permanent dentition with impacted third molars, most evident at 18, 28, 38, and 48.
- No tool-detected evidence of periapical pathology, implants, root remnants, or significant bone loss.
- Bilateral mandibular canals and maxillary sinuses are identified without flagged abnormality.

Stage: S4_CT (Clinical stage: perception)

Q8: Please analyze the dentofacial deformity based on the provided 3D CT craniofacial reconstruction. You may internally utilize prior sagittal information (e.g., from cephalometric analysis) as a reference to guide your understanding of maxillomandibular relationships.

A8: 1. Frontal View Analysis:

- Facial symmetry: Mild-to-moderate lower facial asymmetry. The mandibular midline and chin are deviated toward the image-left (subject's right). Inferior mandibular border shows a slight roll/cant consistent with this deviation.
- Orbital region: Orbits appear level with no evident vertical dystopia; symmetric orbital rims.
- Zygomatic region: Zygomatic arches and malar buttresses show symmetric width and projection; no obvious malar deficiency or asymmetry.
- Maxilla: Anterior nasal spine and piriform aperture are near midline. Mild vertical cant of the maxillary dentoalveolar complex with the image-left side appearing slightly lower.
- Mandible: Chin deviated to image-left. Mandibular angles asymmetric; the image-left gonial region appears more inferior and laterally flared than the opposite side, suggesting ramus/body asymmetry.

2. Lateral View Analysis:

- Sagittal position of the maxilla: Retruded—reduced anterior projection of the maxillary alveolus and zygomaticomaxillary complex relative to the cranial base.
- Sagittal position of the mandible: Protruded—prominent mandibular body and pogonion advanced anterior to the maxilla.
- Maxillomandibular relationship: Skeletal Class III pattern, dominated by mandibular prognathism with contributory maxillary retrusion.
- Vertical facial proportions: High-angle/long-face (hyperdivergent) tendency with a steep mandibular plane, clockwise mandibular rotation, and increased lower anterior facial height.
- Chin morphology: Protruded symphysis/pogonion with a robust/broad chin. 3. Three-Dimensional Summary:
- Sagittal discrepancies: Marked negative maxillomandibular sagittal relationship (Class III). The dominant component is mandibular excess/prognathism; the maxilla is mildly to moderately retruded.

• Vertical discrepancies: Overall increased lower anterior facial height with a steep/high mandibular plane angle. Mild occlusal/maxillary cant down on image-left.

• Transverse asymmetry: Mandibular asymmetry with chin deviation to image-left and gonial/ramus asymmetry; maxilla largely centered with a slight transverse/vertical cant coordinated to the mandibular deviation. Orbits and zygomas remain symmetric.

Morphology-based clinical inference: The CT morphology shows a skeletal Class III deformity primarily due to mandibular prognathism, with secondary midface (maxillary) retrusion and mild transverse asymmetry/cant. Surgical correction would typically center on mandibular setback with maxillary advancement and de-rotation, including correction of the maxillary cant and mandibular asymmetry to recenter the chin. All conclusions are derived from the 3D bony morphology visible on the provided reconstruction.

Stage: S5_TMJ (Clinical stage: perception)

Q9: What are the temporomandibular joint (TMJ) findings on clinical examination?

A9: On clinical examination of the temporomandibular joints, the degree of mouth opening was 3.7 cm, the opening path was zigzag, and there were clicking sounds on the left during both opening and closing.

Treatment 1 of 9

Q1: What is the diagnosis based on the clinical, radiological, and functional findings?

A1: Part I: Dentofacial Deformity-Related Diagnosis

1. Skeletal Malocclusion (Orthognathic Perspective)

• Sagittal skeletal pattern: Skeletal Class III dentofacial deformity, predominantly due to mandibular prognathism/excess, with an additional component of mild-to-moderate maxillary retrusion/deficiency. Cephalometric severity is marked (ANB approximately –7°, Wits approximately –14.8 mm).

• Vertical pattern: High-angle/hyperdivergent vertical skeletal pattern.

• Transverse relationship: Mild mandibular asymmetry with menton/chin deviation to the patient's right and associated gonial/ramus asymmetry. Mild maxillary occlusal cant, with the left side lower on the image. No gross orbital or zygomatic asymmetry, no major yaw deformity demonstrated.

2. Dental Malocclusion (Orthodontic Perspective)

• Angle classification: Angle Class III malocclusion.

• Overjet / overbite condition: Anterior crossbite with negative overjet. Overbite is not assessable based on current data.

• Dental compensation: Class III dentoalveolar compensation with maxillary incisor proclination and mandibular incisor retroclination/uprighting.

• Alignment: Not assessable based on current data for crowding/spacing.

3. Craniofacial and Structural Conditions

• Congenital syndromes: No syndromic craniofacial abnormality identifiable on current imaging.

• Condylar pathology: No definite condylar hyperplasia, hypoplasia, osteochondroma, or other condylar osseous pathology identified on the provided data.

• Jaw defects: Maxillary sagittal deficiency and mandibular excess/prognathism as components of the dentofacial deformity, no post-traumatic, post-surgical, or segmental jaw defect identified.

• Airway-related conditions: Not assessable based on current data.

• Temporomandibular joint disorders: Not assessable based on current data, aside from absence of definite osseous condylar lesion on available imaging.

4. Integrated Dentofacial Diagnosis

• Severe skeletal Class III dentofacial deformity characterized predominantly by mandibular prognathism with contributory maxillary retrusion, associated Angle Class III malocclusion

with anterior crossbite/negative overjet and Class III incisor compensation, plus mild transverse mandibular asymmetry with chin deviation to the right and mild occlusal cant.

• Vertical pattern is consistently high-angle/hyperdivergent on 3D CT morphology and cephalometric measurements, with increased lower anterior facial height and clockwise mandibular rotation.

Part II: Dental and Systemic Disease Diagnosis

1. Dental Hard Tissue Diseases

• No definite caries identified on current imaging.
• Other tooth defects/non-carious lesions: Not assessable based on current data.

2. Periodontal Conditions

• No radiographic evidence of significant generalized periodontal bone loss on the panoramic image.
• Periodontal soft tissue status/gingivitis: Not assessable based on current data.

3. Dentition and Prosthodontic Status

• Permanent dentition essentially complete.
• Impacted third molars: 18, 28, 38, and 48.
• No implant, root remnant, crown, or radiographically evident endodontically treated tooth identified on current imaging.

4. Other Oral Pathologies

• No definite periapical radiolucency, odontogenic cyst, abscess, tumor, or other focal osseous jaw lesion identified on the provided imaging.

5. Systemic Conditions

• Not assessable based on current data.

Treatment 2 of 9

Q2: What treatment strategies are available for this patient, and how do they differ conceptually?

A2: One comprehensive treatment strategy is defined: staged orthodontic-orthognathic treatment with third-molar extractions, presurgical orthodontic decompensation and arch coordination, segmental Le Fort I osteotomy, BSSRO, paranasal bone grafting, and postsurgical orthodontic finishing. A clinically plausible alternative is non-surgical orthodontic camouflage, which would be less invasive but would not provide the same skeletal correction. Another possible alternative is deferring definitive treatment with monitoring until readiness for surgery is confirmed. The main conceptual difference is definitive skeletal correction versus dental compensation or delayed intervention.

Treatment 3 of 9

Q3: Is active treatment necessary, or would observation or conservative treatment be sufficient?

A3: Active treatment is indicated because a full orthognathic pathway has been planned, including presurgical orthodontics, jaw surgery, and postsurgical finishing. Observation alone would not accomplish the intended skeletal and occlusal correction. Whether a conservative camouflage approach is acceptable depends on deformity severity, functional impairment, and patient goals; this requires further clinical and imaging evaluation.

Treatment 4 of 9

Q4: What functional, skeletal, occlusal, or aesthetic improvements are expected from treatment?

A4: Expected benefits include improved maxillomandibular coordination, correction of skeletal jaw discrepancy, better dental arch coordination, improved incisor positioning through mandibular anterior decompensation, establishment of a more stable occlusion, and enhance-

ment of midfacial support from paranasal bone grafting. The exact magnitude of functional and facial aesthetic improvement requires further clinical and imaging evaluation.

Treatment 5 of 9

Q5: Can treatment start immediately, or is follow-up required before definitive intervention?

A5: Treatment should begin with the planned preparatory phase rather than immediate jaw surgery. Initial steps include extraction of teeth 38 and 48 and presurgical orthodontics with maintenance of the U1-SN angle, root separation of teeth 11 and 21, mandibular incisor decompensation, and arch coordination. Surgical intervention follows after orthodontic preparation. Final readiness for surgery requires further clinical and imaging evaluation.

Treatment 6 of 9

Q6: What is the main treatment pathway, and what are the principal components of the plan?

A6: The main pathway is comprehensive orthodontic-surgical treatment in sequence: 1) extraction of teeth 38 and 48; 2) presurgical orthodontics to maintain U1-SN, separate roots of 11 and 21, decompensate mandibular anterior teeth, and coordinate the arches; 3) surgery with extraction of 18 and 28, segmental Le Fort I osteotomy for coordination, BSSRO, and paranasal bone grafting; 4) postsurgical orthodontics to finalize and stabilize occlusion. The core objective is definitive skeletal and occlusal correction with stable postoperative intercuspation.

Treatment 7 of 9

Q7: Are there alternative treatment options, and how do they differ in invasiveness, comprehensiveness, and expected outcomes?

A7: Yes. The planned orthodontic-surgical approach is the most comprehensive and invasive option, with the greatest potential for skeletal, occlusal, and facial correction. A possible alternative is orthodontic camouflage without jaw surgery, which is less invasive but mainly provides dental compensation and may leave the skeletal discrepancy uncorrected. Another alternative is delaying definitive treatment with interim observation if surgical readiness, growth status, or patient preference is uncertain. Suitability of these alternatives requires further clinical and imaging evaluation.

Treatment 8 of 9

Q8: Which plan is more comprehensive, which is more limited, and what factors influence plan selection?

A8: The bimaxillary orthognathic plan with segmental Le Fort I osteotomy, BSSRO, paranasal grafting, and perioperative orthodontics is the more comprehensive plan. Orthodontic camouflage or delayed treatment are more limited approaches. Plan selection depends on the severity of skeletal discrepancy, occlusal goals, facial aesthetic objectives, functional concerns, periodontal and dental conditions, patient willingness to undergo surgery, and confirmation of surgical readiness. Some of these factors require further clinical and imaging evaluation.

Treatment 9 of 9

Q9: What is the recommended decision direction, what key factors affect plan selection, and what information still requires evaluation?

A9: The current decision direction favors proceeding with the planned comprehensive orthodontic-orthognathic pathway. Key factors affecting final selection include the degree of skeletal deformity, anticipated benefit from segmental maxillary surgery and mandibular repositioning, need for stable occlusion, facial soft-tissue goals, and patient preference regarding surgical invasiveness. Information still requiring evaluation includes complete clinical examination, cephalometric and 3D imaging analysis, occlusal records, periodontal and root status, and surgical readiness assessment.

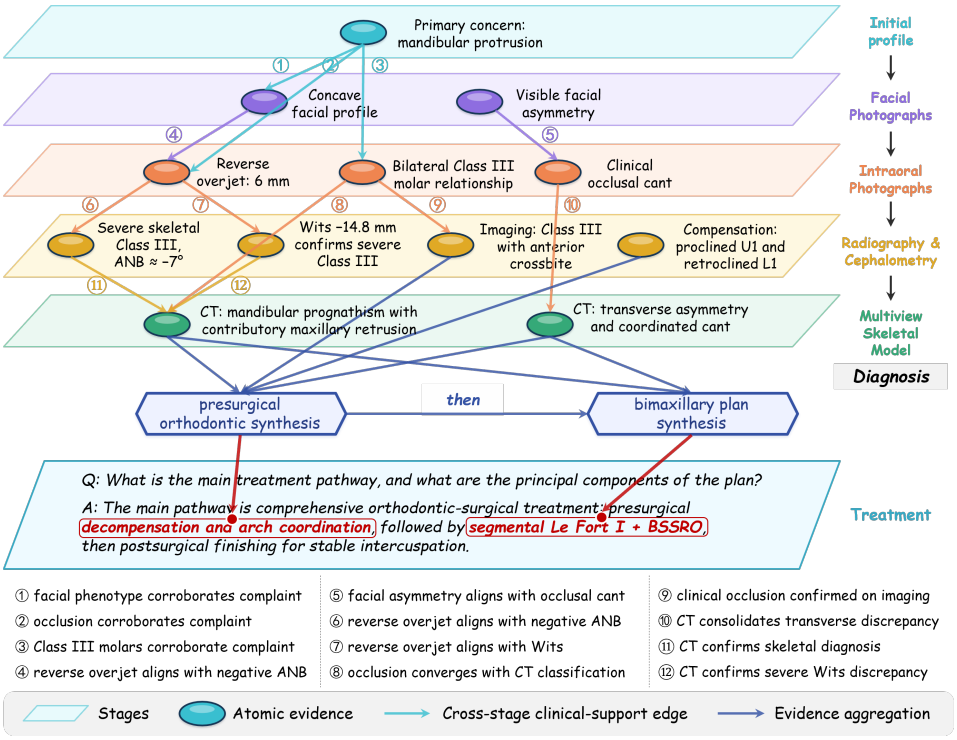


Figure 14: Evidence graph for Case 1.

G.1.2 Representative tasks

Modality perception

Q: Based on the facial photographs, what did you observe about this patient's facial symmetry, profile, chin position, nasolabial angle, upper lip length, and lip posture at rest?

A: The facial photographs show an asymmetric face with the chin deviated about 2 mm to the right and a concave profile with an obtuse nasolabial angle. Upper lip length measures 20 mm, and there is no upper incisor display or gummy smile visible at rest.

Longitudinal evidence recall

Q: Considering the facial photographs and cephalometric radiograph from earlier visits, how does the three-dimensional CT scan's assessment of chin/mandibular midline deviation compare with those earlier findings?

A: The facial photographs from the initial exam were the first to document a 2 mm chin deviation to the right. This was later corroborated by the cephalometric radiograph, which showed a mild menton deviation without a gross skeletal cant. The CT scan then confirmed and refined this finding, showing the mandibular midline and chin deviated to the patient's right, consistent with the earlier photographic and cephalometric observations.

Cross-modal reasoning

Q: Bringing together the patient's reported facial concerns, the intraoral bite findings, the earlier cephalometric measurements, and now the three-dimensional CT, how do these pieces together explain whether her Class III skeletal pattern is primarily driven by mandibular overgrowth, maxillary deficiency, or a combination of both?

A: The chief complaint of mandibular protrusion and a concave facial profile are corroborated dentally by bilateral Class III molars with 6 mm reverse overjet, and skeletally by a severely negative ANB (˜-7°) and Wits (˜-14.8 mm). CT confirms an overall skeletal Class III pattern

driven primarily by mandibular prognathism (protruded mandibular body/pogonion, SNB ˜89°) with only a mild contributory maxillary retrusion (SNA near-average at ˜82° but reduced anterior projection on CT). Together the findings indicate a mandible-dominant Class III skeletal discrepancy rather than a purely maxillary-deficiency pattern.

G.2 Case 2: Skeletal Class II with mandibular retrognathia and staged rehabilitation

Clinical synopsis

A 25-year-old woman with maxillary incisor protrusion and a gummy smile. Intraoral examination, cephalometry, and CT indicate skeletal Class II driven mainly by mandibular retrognathia, with increased overjet, lower-incisor compensation, and a mild hyperdivergent pattern. Bilateral TMJ clicking motivates stabilization before staged orthodontic-orthognathic treatment.

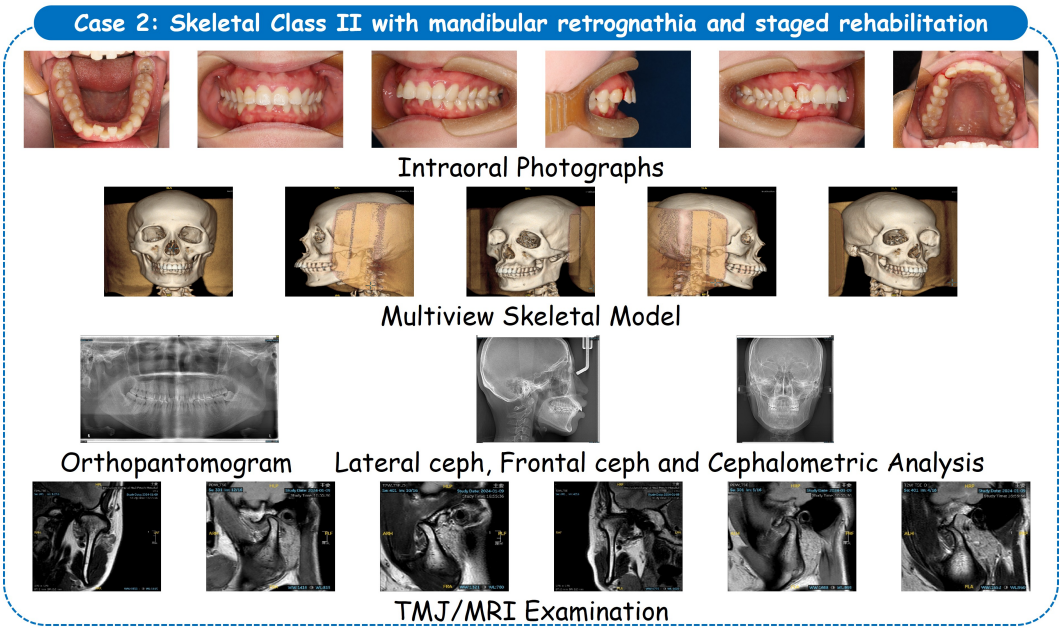


Figure 15: De-identified VQA image panel for Case 2. Facial photographs are intentionally omitted.

G.2.1 Standard VQA trajectory

Stage: S0_PROFILE (Clinical stage: perception)

Q1: Please provide your basic information, including age, gender, height, etc.

A1: She is a 25-year-old woman, 167 cm tall and weighing 53 kg.

Q2: What is the primary concern of the patient?

A2: The patient's primary concern is protrusion of the upper incisors with a gummy smile, present for 8 years.

Q3: What is the patient's past medical or surgical history?

A3: The patient has no relevant past dental, medical, or surgical history.

Stage: S1_FP (Clinical stage: perception)

Q4: What objective findings can be observed from the facial photographs?

A4: The facial photographs show a symmetric facial structure with 8 mm of maxillary incisor display at rest and a 4 mm gummy smile. The upper lip length is 20 mm. The profile is convex, and the nasolabial angle is acute.

Stage: S2_DP (Clinical stage: perception)

Q5: What are the occlusal and dental findings based on the intraoral examination?

A5: Intraoral examination showed a Class II molar relationship on the left and Class II on the right, with an overjet of 8 mm and an overbite of 7 mm. The upper dental midline was central, while the lower dental midline was shifted 2 mm to the left. There was no canting, and crowding was present from 32 to 42.

Stage: S3_XR_XLA (Clinical stage: perception)

Q6: Based on the lateral and posteroanterior cephalometric radiographs and measurements, what are the key dento-skeletal characteristics of this patient?

A6: Key dento-skeletal characteristics (from lateral and PA cephs and the measurements):

Skeletal pattern

- Skeletal Class II: ANB 7.6°, Wits +7.9 mm.
- Maxilla mildly protrusive/forward: SNA 83.6° (slightly ↑).
- Mandible retrusive and relatively small: SNB 76.0° (↓); short ramus/body values (e.g., Co–Go 66 mm; S-Vert–Co 111 mm, both ↓); chin retruded (soft-tissue convexity N′-Sn-Pg′ 23.2°).
- Vertical pattern: mild hyperdivergence/clockwise rotation of the mandible (SN-MP 36°, MP-FH 29°, Y-axis 66.6°). Overall anterior/posterior facial height proportions are near normal.

Dental pattern

- Upper incisors near normal to slightly proclined/forward (U1-NA 21°, 6.9 mm; U1-SN 104.6°); mild dentoalveolar extrusion (U1-PP 30.3 mm, U6-PP 23.8 mm).
- Lower incisors significantly proclined/forward (IMPA 100.5°, L1-NB 37.1°, 9.0 mm); reduced inter-incisal angle 114.4° → dental compensation for Class II.
- Overjet increased to 6.1 mm; overbite increased to 4.5 mm (deep bite tendency).
- Occlusal plane slightly steep (Mx occlusal plane-FH 10.7°).

Soft tissues

- Convex profile with retruded chin and lip protrusion relative to E-plane (UL-EP +5.5 mm, LL-EP +7.5 mm).

Transverse (PA ceph)

- No marked skeletal transverse asymmetry or cant is evident on the PA film.

Overall diagnosis: Skeletal Class II due mainly to mandibular retrognathia with a mild hyperdivergent tendency; dentoalveolar compensation characterized by proclined lower (and slightly upper) incisors, increased overjet and mild deep bite, and a convex soft-tissue profile with relatively protrusive lips.

Q7: Please generate a comprehensive panoramic X-ray report based on the provided panoramic X-ray image.

A7: Panoramic X-ray Report:

Image quality:

- Diagnostic panoramic radiograph obtained with mild central superimposition/ghosting, but overall interpretable.

Dentition:

- Partial permanent dentition is present.
- Automated tooth identification detected most posterior and mandibular teeth, anterior maxillary teeth were only partially identified, likely due to overlap and image quality.

- Bilateral mandibular third molars are present.

Key findings:

- Impacted teeth:
  - 38: Impacted mandibular left third molar.
  - 18, 28, and 48: Erupted third molars.
  - Tool-based condition detection and tooth localization support impaction of the mandibular left third molar (38).
- Periapical pathology:
  - No periapical lesion detected.
  - No granuloma, cyst, or abscess identified by lesion subclass analysis.
- Caries / restorations / endodontics:
  - No definite caries, crowns, fillings, implants, root canal treatment, or root pieces detected on automated analysis.
- Bone support:
  - No definite generalized bone loss identified by automated bone-loss segmentation.
- Jaw structures:
  - Bilateral maxillary sinuses are visualized.
  - Bilateral mandibular canals are identified.

Impression:

- 18, 28, and 48: Erupted third molars; 38: impacted mandibular left third molar.
- No definite radiographic evidence of periapical inflammatory lesion.
- No obvious restorations, implants, or endodontically treated teeth detected.
- No definite generalized periodontal bone loss identified on this panoramic image.

Stage: S4_CT (Clinical stage: perception)

Q8: Please analyze the dentofacial deformity based on the provided 3D CT craniofacial reconstruction. You may internally utilize prior sagittal information (e.g., from cephalometric analysis) as a reference to guide your understanding of maxillomandibular relationships.

A8: 1. Frontal View Analysis:

- Facial symmetry: Overall near-symmetric craniofacial skeleton; no appreciable menton (chin) deviation from the midsagittal plane on the 3D CT.
- Orbital region: Bilaterally level orbital rims with symmetric vertical position; no vertical dystopia evident.
- Zygomatic region: Symmetric zygomatic width and malar projection; no unilateral flattening.
- Maxilla: No obvious maxillary cant or vertical asymmetry of the alveolar crest or piriform rims; midline appears centered.
- Mandible: Mandibular angles appear bilaterally similar in height and contour; no conspicuous asymmetry of ramus height; chin appears centrally aligned.

2. Lateral View Analysis:

- Sagittal position of the maxilla: Approximately normal to mildly protrusive relative to the cranial base; zygomatic–maxillary buttress and ANS position are not posteriorly placed.
- Sagittal position of the mandible: Retruded/deficient. The mandibular body and pogonion are positioned posteriorly with an obtuse gonial angle and a steep mandibular plane suggesting clockwise rotation.
- Maxillomandibular relationship: Skeletal Class II pattern, dominated by mandibular retrognathia. Overjet on the CT appears increased, consistent with the mandibular setback relative to the maxilla.
- Vertical facial proportions: Long-face tendency. Features include increased lower anterior facial height, steep mandibular plane, and clockwise rotation of the mandible.
- Chin morphology: Retruded chin (pogonion posterior to a nasion vertical reference and to subnasale-based facial line), small-to-moderate symphyseal projection.

3. Three-Dimensional Summary:

• Sagittal discrepancies: Clear skeletal Class II relationship primarily due to mandibular retrusion; maxilla is normal to slightly protrusive but not the dominant factor.
• Vertical discrepancies: Long-face pattern with increased lower facial height and obtuse gonial angle; mandibular clockwise rotation. Possible mild anterior vertical maxillary excess cannot be excluded, but the dominant vertical finding is mandibular plane steepness/rotation.
• Transverse asymmetry: No significant transverse skeletal asymmetry detected. Orbits and zygomas are symmetric; maxillary base shows no obvious cant; mandibular rami and angles are bilaterally similar with the menton near midline.

Stage: S5_TMJ (Clinical stage: perception)

Q9: What are the temporomandibular joint (TMJ) findings on clinical examination?

A9: On clinical examination of the temporomandibular joints, mouth opening is 4 cm, the mandible shifts to the right on opening, and there are clicking sounds on both opening and closing bilaterally.

Treatment 1 of 9

Q1: What is the diagnosis based on the clinical, radiological, and functional findings?

A1: Part I: Dentofacial Deformity-Related Diagnosis

1. Skeletal Malocclusion (Orthognathic Perspective)

• Sagittal skeletal pattern: Skeletal Class II pattern, dominated by mandibular retrognathia. Approximately normal to mildly protrusive maxilla relative to the cranial base; zygomatic-maxillary buttress and ANS position are not posteriorly placed. Mandible is retruded/deficient.
• Vertical pattern: Long-face tendency. Features include increased lower anterior facial height, steep mandibular plane, and clockwise rotation of the mandible.
• Transverse relationship: The facial examination shows a symmetric facial structure with 8 mm of maxillary incisor display at rest and a 4 mm gummy smile. Occlusal examination shows no canting, and crowding is present from 32 to 42.

2. Dental Malocclusion (Orthodontic Perspective)

• Angle classification: Class II molar relationship on the left and right.
• Overjet / overbite: Overjet of 8 mm and overbite of 7 mm.
• Dental compensation: Dentoalveolar compensation characterized by marked mandibular incisor proclination/protrusion and slight maxillary incisor proclination.
• Alignment: No canting, crowding present from 32 to 42.

3. Craniofacial and Structural Conditions

• Congenital syndromes: No syndromic craniofacial anomaly identified on current imaging.
• Condylar pathology: No definite condylar hyperplasia, hypoplasia, osteochondroma, or other osseous condylar deformity identified.
• Jaw defects: Maxilla approximately normal to mildly protrusive; mandible retruded/deficient.
• Airway-related conditions: Not assessable based on current data.
• Temporomandibular joint disorders: Temporomandibular disorder (TMD), based on the provided clinical/functional diagnosis; specific structural TMJ abnormality is not definitively characterized on the current imaging dataset.

4. Integrated Dentofacial Diagnosis

• Skeletal Class II pattern dominated by mandibular retrognathia, with an approximately normal to mildly protrusive maxilla.
• Long-face tendency with increased lower anterior facial height, steep mandibular plane, and clockwise mandibular rotation.
• Symmetric facial structure with 8 mm maxillary incisor display at rest and a 4 mm gummy smile.
• Class II molar relationship on the left and right, with an overjet of 8 mm and overbite of 7 mm.

• Dentoalveolar compensation characterized by marked mandibular incisor proclination/protrusion and slight maxillary incisor proclination.

Part II: Dental and Systemic Disease Diagnosis

1. Dental Hard Tissue Diseases

• No definite radiographic evidence of caries, non-carious hard tissue defects, or periapical inflammatory disease on the current panoramic imaging.

2. Periodontal Conditions

• No definite generalized periodontal bone loss identified radiographically.
• Clinical periodontal status is not assessable based on current data.

3. Dentition and Prosthodontic Status

• Third molars: 18, 28, 38, 48, per provided diagnosis sheet content; panoramic imaging clearly supports impaction of 38.
• Restorative/prosthodontic status: No definite crowns, fixed prostheses, implants, or endodontically treated teeth identified.
• Other missing teeth or agenesis: Not assessable based on current data.

4. Other Oral Pathologies

• No definite odontogenic cyst, tumor, abscess, granuloma, or other focal osseous lesion identified.

5. Systemic Conditions

• Not assessable based on current data.

Treatment 2 of 9

Q2: What treatment strategies are available for this patient, and how do they differ conceptually?

A2: The primary available strategy is a comprehensive multidisciplinary approach: TMJ stabilization with rehabilitation therapy, presurgical orthodontic decompensation with planned extractions, orthognathic surgery with segmental Le Fort I osteotomy and BSSRO, and postsurgical orthodontic finishing. A clinically reasonable alternative is a modified comprehensive surgical plan using the same orthodontic-surgical pathway but omitting optional adjuncts such as PAOO or subapical osteotomy unless intra-treatment findings indicate need. A non-surgical or observation-based option would be limited and would not address the planned skeletal correction; its suitability requires further clinical and imaging evaluation.

Treatment 3 of 9

Q3: Is active treatment necessary, or would observation or conservative treatment alone be sufficient?

A3: Active treatment appears necessary because the plan includes presurgical orthodontics, multiple extractions, and orthognathic surgery, indicating a deformity not intended for observation alone. Conservative management by itself may help TMJ stabilization but would not substitute for the planned orthodontic-surgical correction. Whether a purely conservative pathway is acceptable requires further clinical and imaging evaluation.

Treatment 4 of 9

Q4: What functional, skeletal, occlusal, or aesthetic improvements are expected from treatment?

A4: Expected benefits include TMJ stabilization, dental leveling and alignment, orthodontic decompensation, correction of skeletal and occlusal discrepancies through maxillary and mandibular surgery, and improvement of final occlusal stability after postsurgical orthodontics. Aesthetic improvement is also expected from orthognathic correction, but the specific magnitude and facial goals require further clinical and imaging evaluation.

Treatment 5 of 9

Q5: Can treatment start immediately, or is follow-up required before definitive intervention?

A5: Treatment can begin with rehabilitation measures such as an occlusal splint to stabilize the TMJ, followed by presurgical orthodontic preparation. Definitive surgery should follow completion of orthodontic leveling, alignment, and decompensation. The need for adjunctive PAOO or subapical osteotomy should be determined during ongoing evaluation and treatment progression.

Treatment 6 of 9

Q6: What is the main treatment pathway, and what are the principal components of the plan?

A6: The main pathway is: 1) TMJ rehabilitation with occlusal splint therapy, 2) presurgical orthodontics with extraction of 34, 44, 38, and 48 for leveling, alignment, and decompensation, with PAOO in the 13-23 region if necessary, 3) orthognathic surgery with extraction of 14, 24, 18, and 28 combined with segmental Le Fort I osteotomy and BSSRO, with subapical osteotomy if necessary, and 4) postsurgical orthodontics to establish stable occlusion. The core objective is comprehensive skeletal and occlusal rehabilitation.

Treatment 7 of 9

Q7: Are there alternative strategies, and how do they differ in invasiveness, comprehensiveness, and expected outcomes?

A7: An alternative within standard practice is a comprehensive orthognathic plan without optional adjunctive PAOO or subapical osteotomy unless treatment response shows they are required; this is somewhat less invasive but still comprehensive. Another theoretical alternative is conservative management with splint therapy and orthodontic camouflage, but this would be less comprehensive and would not predictably achieve the same skeletal correction as bimaxillary orthognathic surgery. Selection among these options requires further clinical and imaging evaluation.

Treatment 8 of 9

Q8: Which plan is more comprehensive, which is more limited, and what factors influence plan selection?

A8: The most comprehensive plan is the full staged surgical-orthodontic pathway including TMJ rehabilitation, presurgical orthodontics, extractions, segmental Le Fort I osteotomy, BSSRO, and postsurgical orthodontics, with PAOO and subapical osteotomy used if indicated. A more limited plan would omit optional adjunctive procedures or rely on conservative measures, but this may reduce the extent of correction. Plan selection depends on TMJ status, severity of skeletal and dental discrepancy, orthodontic decompensation needs, surgical goals, and whether adjunctive procedures are necessary during treatment.

Treatment 9 of 9

Q9: What is the recommended decision direction, what key factors affect final plan selection, and what still needs evaluation?

A9: The recommended direction is to proceed with staged comprehensive treatment beginning with TMJ stabilization and presurgical orthodontic preparation, followed by definitive orthognathic surgery and postsurgical finishing. Key factors affecting final selection include response of TMJ symptoms to rehabilitation, readiness for orthodontic decompensation, and intra-treatment need for PAOO or subapical osteotomy. Exact surgical design, necessity of all adjunctive procedures, and appropriateness of any less invasive alternative require further clinical and imaging evaluation.

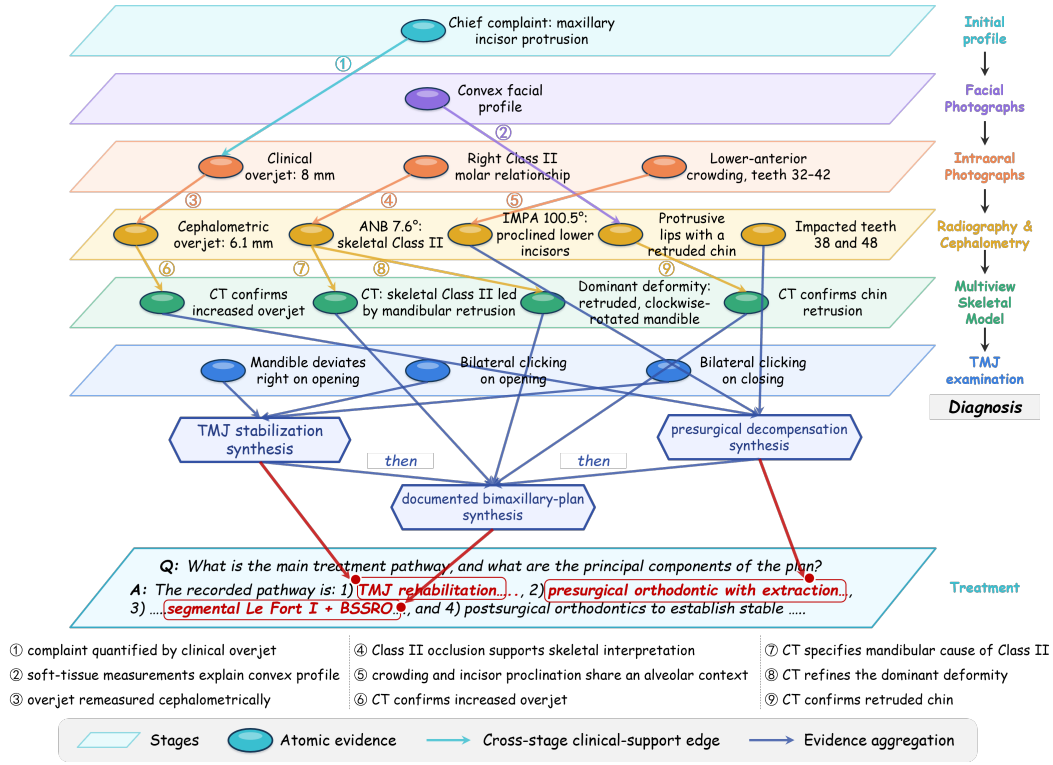


Figure 16: Evidence graph for Case 2.

G.2.2 Representative tasks

Modality perception

Q: Based on the intraoral dental photographs, what were the molar relationships on each side, and what were the overjet and overbite measurements, dental midline alignment, presence of canting, and any lower anterior crowding?

A: The intraoral photographs show a Class II molar relationship on the left and right, with 8 mm of overjet and 7 mm of overbite. The upper dental midline is centered while the lower midline is shifted 2 mm to the left, no occlusal canting is seen, and crowding is present from tooth 32 to tooth 42.

Longitudinal evidence recall

Q: How was the skeletal Class II relationship in this patient first identified on earlier radiographic and dental photograph evaluation, and how did the later three-dimensional CT scan confirm and refine the understanding of which jaw was primarily responsible for this pattern?

A: The skeletal Class II pattern was first established on cephalometric radiographic assessment, with an ANB of 7.6°, a slightly increased SNA (83.6°), and a decreased SNB (76.0°) indicating a mildly protrusive maxilla combined with a retrusive mandible; this was supported by a Class II molar relationship on the right side seen on the intraoral dental photographs. The later three-dimensional CT confirmed this skeletal Class II relationship and specified that mandibular retrognathia, rather than maxillary excess, is the dominant deformity, with the maxilla approximately normal to mildly protrusive and the mandible retruded/deficient.

Cross-modal reasoning

Q: Given that her facial photographs, PA cephalometric radiograph, and 3D CT all showed a symmetric craniofacial skeleton without menton deviation or ramus asymmetry, how should we

interpret the mandibular shift and bilateral clicking observed during the TMJ examination—does this indicate a true skeletal asymmetry, and if not, what is the more likely explanation?

A: Facial photographs, the PA cephalometric radiograph, and 3D CT all show a symmetric craniofacial skeleton with no menton deviation, no maxillary cant, and bilaterally similar mandibular rami, ruling out structural skeletal asymmetry. However, TMJ examination shows the mandible shifts to the right on opening with bilateral clicking on opening and closing, indicating the asymmetry present is a dynamic/functional finding related to joint dysfunction rather than a skeletal deformity.

Memory update / conflict correction

Q: The cephalometric radiographs showed both a mild increase in SNA and a decrease in SNB contributing to the skeletal Class II relationship — based on the three-dimensional CT findings, which jaw is now identified as the primary driver of this Class II pattern, and how does that change the surgical planning focus compared to the earlier cephalometric interpretation?

A: Cephalometric radiographs established a skeletal Class II relationship (ANB 7.6°) with both a slightly increased SNA (83.6°, mild maxillary protrusion) and a decreased SNB (76.0°, mandibular retrusion), leaving it unclear which jaw was the primary driver. The 3D CT clarified this by showing the maxilla is only approximately normal to mildly protrusive while the mandible is truly retruded/deficient with a posteriorly positioned body and pogonion, establishing that the Class II pattern is dominated by mandibular retrognathia (with a clockwise rotational component) rather than by maxillary excess. This shifts surgical planning emphasis toward correcting mandibular deficiency rather than treating the malocclusion as a combined or maxilla-driven problem.

G.3 Case 3: Endodontic-periodontal lesion with regenerative surgery

Clinical synopsis

A #30-31 lesion followed from pulpal necrosis and persistent sinus tract through CBCT reassessment, apical surgery with guided tissue regeneration, early wound complications, and one-year clinical and radiographic recovery.

G.3.1 Standard VQA trajectory

Stage: T0_perception_00 (Preoperative / initial diagnostic workup: perception)

Q1: What was visible in the preoperative clinical and radiographic images of #31?

A1: An abscess was present on the buccal gingiva interproximally between #30 and #31. The preoperative radiograph demonstrated a periapical lesion extending from the distal root of #30 to the mesial and distal roots of #31, and tracing showed the abscess to the interproximal area between #30 and #31.

Q2: What diagnostic findings were recorded for #30 and #31 at the initial workup?

A2: #31 had no response to electric pulp testing and cold testing. 9–13 mm probing depths were present interproximally, with normal probing depths present in all other areas. Teeth #30 and #31 were previously restored with full coverage gold crowns with no clinical signs of recurrent caries. Tooth #30 was diagnosed as normal pulp and normal apical tissues, and #31 was diagnosed as necrotic pulp with acute apical abscess.

Stage: T1_treatment_00 (After initial diagnosis: treatment)

Q3: What initial endodontic treatment was performed for #31?

A3: #31 received NSRCT by an endodontist.

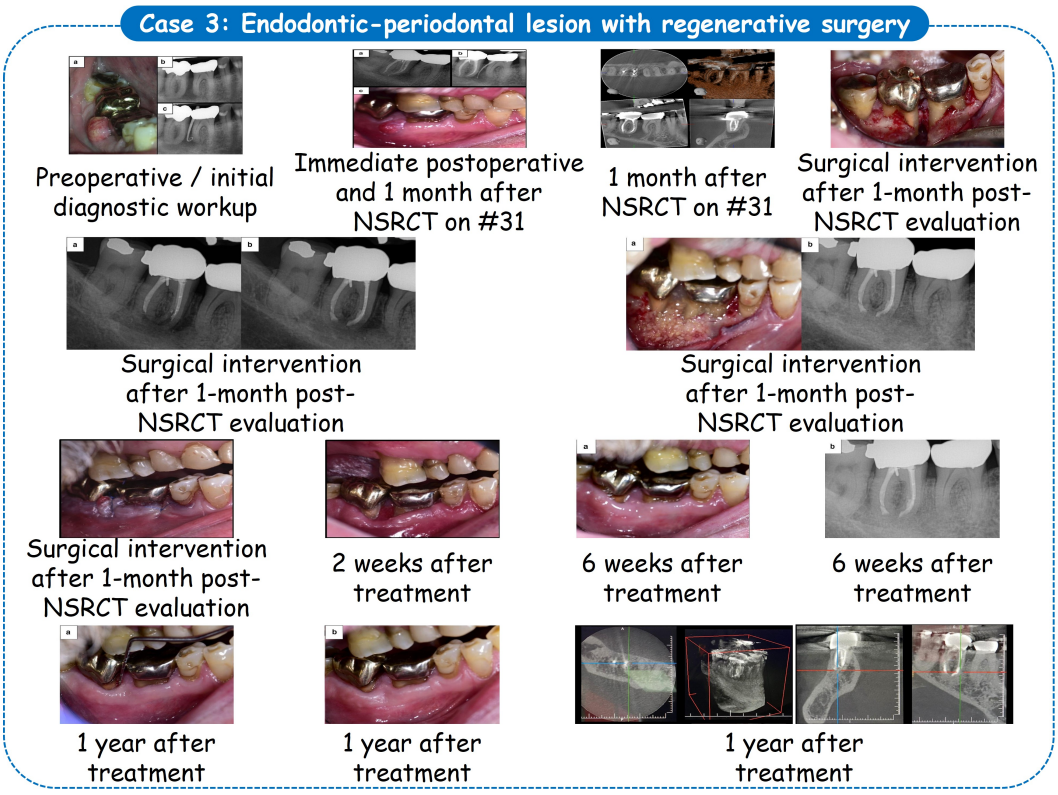


Figure 17: Representative source figures distributed across the complete timeline for Case 3.

Stage: T2_followup_00 (Immediate postoperative following NSRCT on #31: followup)

Q4: What immediate postoperative radiographic documentation was reported after NSRCT on #31?

A4: An immediate postoperative radiograph following NSRCT on #31 was reported.

Stage: T3_followup_01 (1 month after NSRCT on #31: followup)

Q5: What was visible on the postoperative CBCT scan obtained 1 month after NSRCT on #31?

A5: Cross sections showed an extensive radiolucent lesion extending from the distal root of #30 to the distal root of #31, with loss of buccal plate present interproximally between #30 and #31, leading to a 1 wall periodontal defect.

Q6: What clinical and imaging findings were reported 1 month after NSRCT on #31?

A6: One-month post-treatment, a persistent sinus tract was present. The postoperative radiograph 1 month after treatment demonstrated a sinus tract traced to the interproximal area between #30 and #31, and clinical findings showed a persistent sinus tract present 1 month after NSRCT interproximally between #30 and #31. A CBCT scan was obtained which revealed complete loss of the buccal and interproximal bone between #30 and #31, leading to a 1-wall defect.

Q7: What diagnosis and prognosis were assigned after the persistent lesion was evaluated?

A7: Prognosis for #31 was deemed questionable; however the patient wished to save the tooth. A diagnosis of an endodontic-periodontal lesion without root damage in a non-periodontitis patient, Grade 3 was made.

Q8: What combined management approach was selected for the persistent lesion?

A8: A collaborative surgery was conducted to treat the lesion through apicoectomy of #31 with GTR.

Stage: T4_followup_02 (Surgical intervention after 1-month post-NSRCT: followup)

Q9: What was visible after surgical access and removal of the lesion for biopsy?

A9: After removal of the lesion for biopsy, loss of buccal plate and a 1-wall defect could be visualized between #30 and #31.

Q10: What was visible in the apicoectomy images of #31?

A10: The mesial and distal root apices of #31 had been resected, and the apicoectomy was shown completed with placement of a retrograde filling.

Q11: What was visible in the guided tissue regeneration images of the defect between #30 and #31?

A11: Mineralized freeze-dried bone allograft was visible in the osseous defect combined with EMD, and the immediate postoperative radiograph demonstrated placement of allograft in the defect.

Q12: What was visible after the flap was replaced at the surgical site?

A12: The flap was replaced and sutured with 6–0 polypropylene sutures.

Q13: What anesthesia and flap approach were used for the collaborative surgery?

A13: After obtaining written informed consent, the surgical area was anesthetized with 3.4 mL of 2% lidocaine with 1:100,000 epinephrine and 1.7 mL of 2% lidocaine with 1: 50,000 epinephrine. A full-thickness mucoperiosteal flap was elevated from #29 to #32 on the buccal and lingual to expose the lesion.

Q14: How was the lesion handled once surgically exposed?

A14: Once exposed, the entire lesion was curetted and placed in 10% formalin for biopsy.

Q15: How was the apicoectomy of #31 performed?

A15: Mesial and distal roots of #31 were resected 3 mm from the root apices, prepared, and filled with retrograde fillings.

Q16: What regenerative procedures were performed at #30 and #31 during the surgery?

A16: Thorough root planing was completed on #30 and #31, followed by root conditioning with 24% EDTA for 2 minutes. The site was irrigated, and enamel matrix derivative (EMD) was applied. The site was filled with mineralized freeze-dried cortico-cancellous allograft and covered by a resorbable membrane.

Q17: What closure and postoperative care were prescribed after the surgery?

A17: The flap was repositioned and sutured with 6–0 polypropylene sutures. Postoperative care included amoxicillin 500 mg three times a day for 7 days, 0.12% chlorhexidine rinse twice daily for 14 days, and ibuprofen 800 mg as needed.

Stage: T5_followup_03 (2 weeks after treatment: followup)

Q18: What was visible in the 2-week postoperative photo of the surgical area?

A18: The 2-week postoperative photo showed gingival recession with the interproximal area granulating in between #30 and #31 following food impaction in the surgical area during the healing phase.

Q19: What clinical outcome and biopsy result were reported at 2 weeks?

A19: At 2 weeks, delayed healing with recession was observed due to food impaction between #30 and #31, which was removed. The biopsy report confirmed a definitive diagnosis of periapical cyst.

Q20: What management was performed for the healing complication noted at 2 weeks?

A20: Food impaction between #30 and #31 was removed.

Stage: T6_followup_04 (6 weeks after treatment: followup)

Q21: What was visible in the 6-week postoperative clinical and radiographic images?

A21: The 6-week postoperative photo showed slight recession on #30 and #31 with minimal inflammation present and absence of a sinus tract. The 6-week postoperative radiograph demonstrated resolution of the periapical lesion and maintenance of the graft.

Q22: What healing status was reported at 6 weeks?

A22: At 6 weeks, the site healed uneventfully with slight recession and absence of a sinus tract.

Stage: T7_followup_05 (1 year after treatment: followup)

Q23: What was visible in the 1-year postoperative clinical images?

A23: The 1-year postoperative photo showed healthy periodontium at #30 and #31, and the patient showed no signs of infection.

Q24: What was visible on the CBCT scan at the 1-year postoperative visit?

A24: The CBCT scan showed substantial bone fill with graft integration at sites #30 and #31, and the periapical lesion had resolved.

Q25: What clinical and CBCT outcomes were reported at 1 year?

A25: At 1 year, the patient was asymptomatic, with probing depths ranging 2–4 mm. The CBCT showed excellent bone fill, integration of the graft, and resolution of the periapical lesion.

G.3.2 Representative benchmark tasks

Modality perception

Q: What did follow-up CBCT imaging show regarding sites #30 and #31 and the periapical lesion?

A: At the 1-year postoperative visit, CBCT showed substantial bone fill and graft integration at sites #30 and #31. The periapical lesion had resolved.

Longitudinal evidence recall

Q: Before the 6-week clinical examination documented that no sinus tract was present, what had the 1-month post-NSRCT clinical examination shown about the sinus tract and its location?

A: Before the 6-week finding that a sinus tract was absent, the relevant earlier context was the 1-month post-NSRCT visit, when a persistent sinus tract was present interproximally between #30 and #31.

Cross-temporal reasoning

Q: In the surgical report for the defect between teeth #30 and #31, what material was placed there, and what did the 1-year postoperative CBCT show about the repair and lesion status?

A: At surgery, the defect between #30 and #31 was filled with mineralized freeze-dried bone allograft combined with EMD. By 1 year, CBCT showed substantial bone fill, graft integration, and resolution of the periapical lesion.

Memory update / conflict correction

Q: At the later follow-up, what was the sinus-tract status compared with the earlier 1-month post-NSRCT finding?

A: Revise the stored sinus-tract status from a persistent sinus tract interproximally between #30 and #31 at 1 month after NSRCT to absence of a sinus tract at 6 weeks.

Representative follow-up task

Q: What management was performed for the healing complication noted at 2 weeks?

A: Food impaction between #30 and #31 was removed.

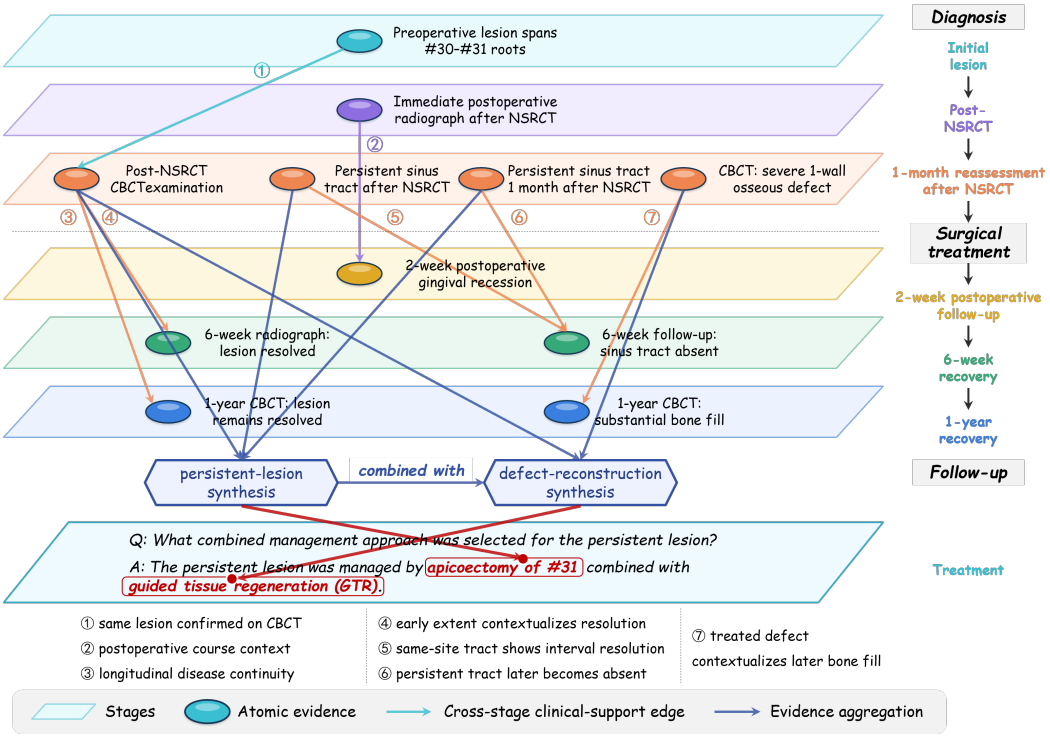


Figure 18: Evidence graph for Case 3.

G.4 Case 4: Immediate implant placement with open healing

Clinical synopsis

A molar extraction socket managed with flapless immediate implant placement and no graft or membrane, followed by serial soft-tissue migration, implant uncovery, restoration, and a stable 26-month peri-implant outcome.

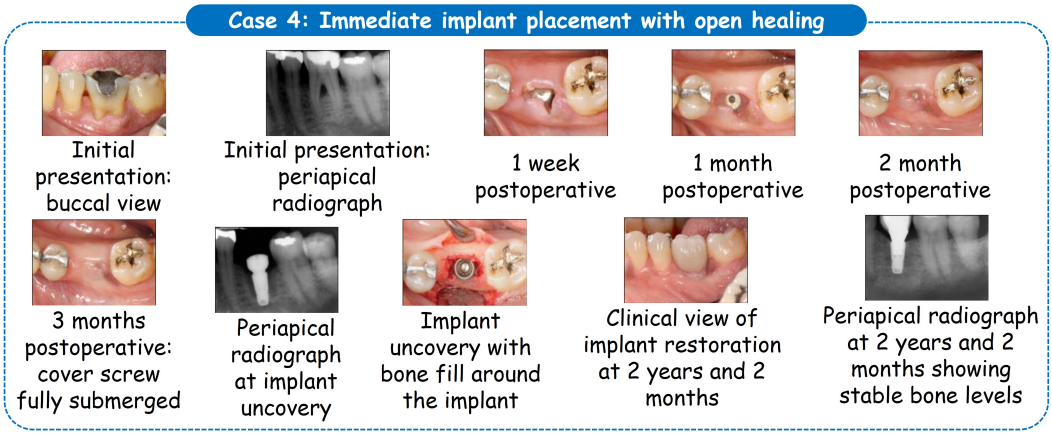


Figure 19: Representative source figures distributed across the complete timeline for Case 4.

G.4.1 STANDARD VQA TRAJECTORY

Stage: T0_perception_00 (Initial presentation: perception)

Q1: What is visible in Figure 1 at the initial presentation?

A1: The mandibular left first molar presented with 5 mm of buccal recession and Class II furcation invasion with a Class B vertical component.

Q2: What radiographic appearance is shown in Figure 2 at the initial presentation?

A2: Extensive furcation involvement was suggested radiographically.

Q3: What tooth and additional buccal gingival finding were present at presentation?

A3: A 55-year-old man presented for extraction of the mandibular left first molar. A 2 $\times$ 4-mm white streak on the buccal marginal gingiva did not wipe off with gauze.

Stage: T1_treatment_00 (Initial procedure: extraction and socket assessment: treatment)

Q4: What initial treatment was planned for the mandibular left first molar?

A4: The tooth was planned for extraction and immediate implant placement.

Q5: What medication was given before the procedure?

A5: The patient premedicated with 2 g amoxicillin prior to the procedure.

Q6: How was the extraction performed and what precautions were taken?

A6: Following administration of local anesthesia, atraumatic extraction was performed by sectioning the tooth through the furcation and elevating both roots, taking care to avoid damage to the socket walls.

Q7: How was the extraction socket prepared after extraction?

A7: The socket was carefully debrided with a surgical curette to remove all remnants of soft tissue, and the socket was irrigated with saline solution.

Q8: What was done with the buccal gingival streak?

A8: The streak of leukoplakia on the buccal gingiva was excised and stored in formalin solution for histologic analysis.

Q9: What was determined by periodontal probing of the socket?

A9: All socket walls were intact with no fenestration or dehiscence present. There was a height discrepancy between the buccal and lingual bone crest; the lingual bone crest was more coronal than the buccal bone crest, consistent with the initial presentation of buccal recession. The socket was classified as Type B according to Smith and Tarnow.

Q10: What septal bone finding supported implant stabilization?

A10: Enough septal bone was observed to stabilize but not completely surround the implant, which would leave gaps between one or more of the surfaces of the implant and the socket walls.

Q11: What implant placement approach was selected?

A11: A flapless implant approach was employed.

Stage: T2_treatment_01 (Initial procedure: implant placement: treatment)

Q12: How was the implant osteotomy prepared?

A12: The implant osteotomy was initiated in the septal bone with a round bur, and twist drills were used to progressively enlarge the osteotomy as per the manufacturer's recommendations.

Q13: What happened to the mesial and distal portions of the septum during osteotomy preparation?

A13: The coronal two-thirds of the mesial and distal aspect of the septum was obliterated.

Q14: What stability check was performed during progressive osteotomy preparation?

A14: Careful attention was paid to make sure that after each successive drill there was sufficient resistance to movement of the drill and guide pin within the osteotomy.

Q15: What implant was placed and with what insertion torque?

A15: A 4.5 × 9-mm TX Astra Tech implant fixture was placed with more than 35 Ncm of torque.

Q16: How was the implant collar positioned relative to the buccal and lingual bone crests?

A16: The collar of the implant was placed at the level of septal bone that was coincident with the buccal bone crest and 2 mm apical to the lingual bone crest.

Q17: What was the initial bone-to-implant contact and where were the remaining implant surfaces exposed?

A17: The bone-to-implant contact was 360 degrees in the apical third of the implant and extended coronally on the buccal and lingual implant surfaces. The coronal two-thirds of the proximal surfaces of the implant were not in contact with bone and were exposed to the residual socket space.

Q18: What was the distance from the implant to the socket walls?

A18: The distance from the mesial and distal border of the implant to their respective socket walls was 4 to 5 mm.

Q19: What was done after implant placement, and what grafting or flap procedures were omitted?

A19: A cover screw was placed over the implant, and the implant was left to heal within the extraction socket without the use of a bone graft, barrier membrane, or coronal flap advancement.

Stage: T3_followup_00 (1 week postoperative: followup)

Q20: What soft-tissue and socket appearance is visible at the 1-week postoperative visit in Figure 7?

A20: The surrounding soft tissue had collapsed toward the center of the socket, and a fibrin clot occupied the gap distance.

Q21: How was the site healing at 1 week postoperatively?

A21: The site was healing within normal limits. No signs of delayed healing or infection were present.

Q22: What covered the gap between the implant and the mesial and distal socket walls?

A22: There was a white layer overlying the gap between the implant and the mesial and distal socket walls, which indicated the presence of a fibrin clot superficially.

Q23: What was noted about the wound edge at the excisional biopsy area and the cover screw?

A23: The wound edge in the area of the excisional biopsy was further away from the center of the socket than the other wound edges and was expected to be the last area to reach the center of the socket during healing. The marginal soft tissue appeared to have collapsed toward the center of the socket and was slightly covering the lingual side of the cover screw.

Q24: What did histologic analysis show for the gingival lesion?

A24: The biopsy results showed that the histology of the lesion was consistent with hyperkeratosis.

Q25: What further management was indicated for the gingival lesion?

A25: No further treatment was indicated.

Stage: T4_followup_01 (1-month postoperative visit: followup)

Q26: What tissue appearance is visible at the 1-month postoperative visit in Figure 8?

A26: Young migrating tissue occupied the previous gap distance. Most of the cover screw and a portion of the implant were still exposed.

Q27: Where had the marginal soft tissue migrated at 1 month?

A27: The marginal soft tissue had migrated from the wound edges and was now contacting the implant and cover screw.

Q28: How did the new tissue differ from the original tissue surrounding the socket?

A28: The new tissue had a thinner epithelial layer with visible red streaks, indicative of blood vessels, and it did not appear completely keratinized at this time. The younger tissue was located more apically in the socket than the surrounding older tissue.

Q29: What was noted about the cover screw and plaque at 1 month?

A29: A portion of the cover screw was still not covered with soft tissue, but moderate plaque buildup was noted at the site.

Q30: What oral hygiene management was prescribed at 1 month?

A30: The patient was instructed to apply chlorhexidine gluconate 0.12% to this area twice a day.

Stage: T5_followup_02 (2 months postoperative: followup)

Q31: What tissue migration and cover-screw appearance are visible at 2 months in Figure 9?

A31: The tissue in the center of the socket had migrated coronally. A portion of the cover screw was still exposed.

Q32: How had the soft tissue changed at 2 months?

A32: The soft tissue had thickened and appeared keratinized; blood vessels were no longer visible to the naked eye. The newly formed soft tissue was more continuous with the surrounding tissue, indicating that it had proliferated coronally.

Q33: What part of the cover screw remained visible at 2 months?

A33: The gingiva had migrated onto the area of the cover screw that was previously exposed, but a small area on the distal aspect of the cover screw was still visible.

Q34: What cleaning instruction was continued at 2 months?

A34: The patient was instructed to continue to clean this area with chlorhexidine gluconate 0.12% until the next visit.

Stage: T6_followup_03 (3 months postoperative; implant uncovery and restoration: followup)

Q35: What soft-tissue appearance is visible at 3 months in Figure 10?

A35: The soft tissue had completely submerged the implant cover screw.

Q36: What radiographic appearance is shown at implant uncovery in Figure 11?

A36: There was suggestion of bone fill in the previous socket spaces.

Q37: What bone appearance is visible at implant uncovery in Figure 12?

A37: Bone fill was present around the implant.

Q38: What radiographic timing was reported for mineralized bone fill?

A38: Radiographic evidence of mineralized bone fill was not expected until 6 months.

Q39: What was found clinically after the implant was exposed?

A39: The buccal crest height and contour were precisely continuous with the buccal surface of the implant, whereas the lingual crest was thicker and was above the implant platform. Sounding with a probe revealed hard tissue around the implant except for a 1- to 2-mm area on the mesial and distal sides immediately adjacent to the first few threads of the implant that felt softer. No mobility of the implant was detected.

Q40: What procedure was performed for implant uncovery?

A40: A crestal incision was made and a full thickness flap was elevated to expose the implant and the cover screw.

Q41: What did the absence of implant mobility indicate?

A41: The absence of implant mobility indicated that the implant had achieved secondary stability.

Q42: What was placed after implant exposure and how were the soft tissues managed?

A42: A healing abutment was secured to the implant, and the soft tissues were sutured and adapted to the healing abutment.

Q43: How was the implant restored?

A43: The patient then returned to the restorative dentist and the implant was restored with a screw-retained porcelain-fused-to-metal crown.

Stage: T7_followup_04 (2 years and 2 months postoperatively: followup)

Q44: What is visible in the clinical view of the implant restoration at 2 years and 2 months postoperatively in Figure 13?

A44: The implant restoration was clinically present at 2 years and 2 months postoperatively.

Q45: What radiographic bone-level appearance is shown at 2 years and 2 months postoperatively in Figure 14?

A45: Crestal bone levels were up to the implant platform.

Q46: What were the peri-implant soft-tissue findings at the 2-year-and-2-month periodontal maintenance visit?

A46: The peri-implant soft tissue appeared healthy, with a wide zone of attached keratinized tissue on the buccal and lingual surfaces. Probing depths ranged from 2 to 3 mm with no bleeding on probing.

Q47: What did the periapical radiograph demonstrate at that visit?

A47: The periapical radiograph demonstrated crestal bone levels up to the implant platform.

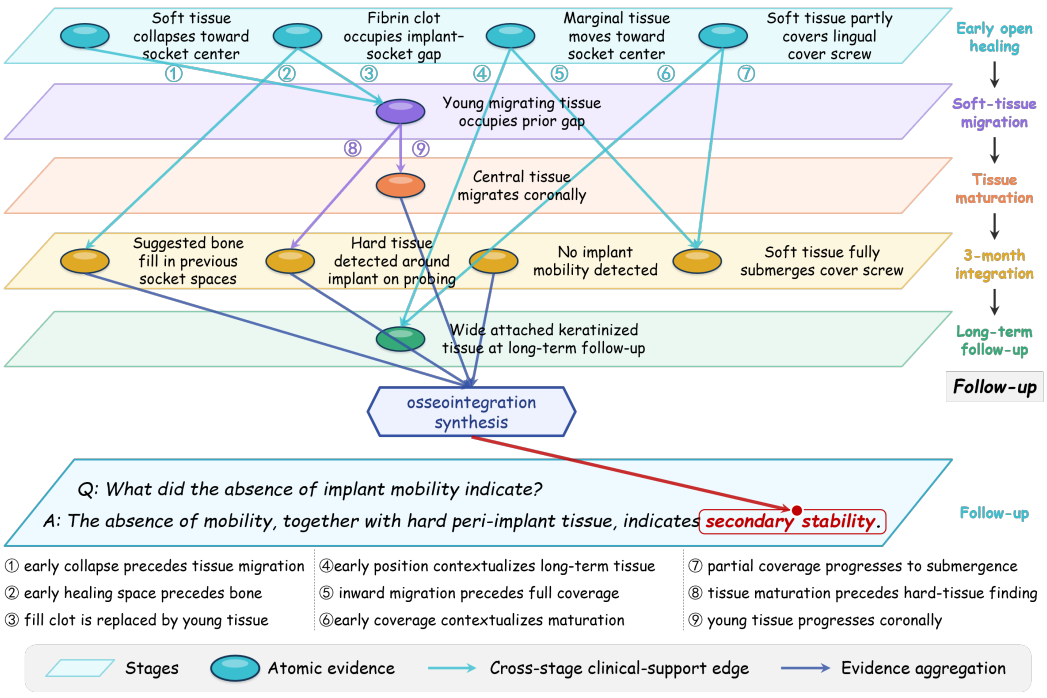


Figure 20: Evidence graph for Case 4.

G.4.2 REPRESENTATIVE BENCHMARK TASKS

Modality perception

Q: What was observed on the buccal marginal gingiva at that examination, including its size and whether it could be wiped off with gauze?

A: A 2 $\times$ 4 mm white streak was present on the buccal marginal gingiva. It did not wipe off with gauze.

Longitudinal evidence recall

Q: What did follow-up show about the gap between the implant and the socket walls at 1 week, and what was observed around the implant at uncovery 3 months later?

A: At the 1-week postoperative visit, a fibrin clot occupied the gap between the implant and the mesial and distal socket walls, with a white layer overlying the gap. At implant uncovery 3 months postoperatively, bone fill was present around the implant.

Cross-temporal reasoning

Q: What were the size, wipeability, and later biopsy histology findings of the white streak on the buccal marginal gingiva?

A: At presentation, the buccal marginal gingiva had a 2 $\times$ 4 mm white streak that did not wipe off with gauze. Histology from the gingival lesion biopsy was consistent with hyperkeratosis.

Memory update / conflict correction

Q: What did the later biopsy and histologic examination show about the non-wipeable white streak on the buccal marginal gingiva?

A: Biopsy histology of the gingival lesion was consistent with hyperkeratosis.

Representative follow-up task

Q: What procedure was performed for implant uncovery?

A: A crestal incision was made and a full thickness flap was elevated to expose the implant and the cover screw.

H PROMPT TEMPLATES

The prompt templates used throughout the benchmark are shown in Figures 22–30.

MULTIMODAL PERCEPTION SYSTEM PROMPT

```
You are a multimodal dental and maxillofacial clinical observer
    .

Use the attached current images as the primary evidence. Use
    the patient profile and available textual clinical records
    only as context.

Requirements:
1. Answer directly within the question's original scope. Do not
     repeat the question, output "<image>", or mention prompts,
     hidden references, evaluation, or ground truth.
2. Report only findings supported by the current images or
    explicit clinical context. Do not add unrequested diagnoses
    , treatment plans, later-stage conclusions, stereotyped
    findings, or broad negative claims.
3. Separate direct visual observations from clinical inference.
     If a requested item is not reliably assessable, say so
    instead of guessing.
4. Preserve supported laterality, tooth numbers, signs, units,
    and qualitative categories. Respect modality boundaries and
     do not infer clinical-exam findings such as mouth opening,
     pain, tenderness, or joint sounds unless provided.
5. When sources conflict, preserve measurement-based
    conclusions, state the disagreement, and keep appropriate
    uncertainty.
6. Use concise natural language or bullets for multiple
    findings, and avoid repetition.
```

Figure 21: System prompt templates for multimodal perception.

MULTIMODAL PERCEPTION USER PROMPT

```
Patient profile and initial context:
{profile}

Textual clinical records available before the current image
    question:
{text_memory}

Current stage: {stage_id}
Modality: {modality}

Answer the following clinical question in its original scope:
{question}

The current-stage images are attached. Begin directly with the
    answer. Base it on the current images and use the profile
    and textual clinical records as context.
```

Figure 22: User prompt templates for multimodal perception.

PERCEPTION VERIFIER PROMPT

```
You are a clinical perception reviewer evaluating one
   multimodal dental and maxillofacial answer.

Compare the MODEL ANSWER with the QUESTION, GOLD ANSWER, and
   the ATOMIC EVIDENCE for this exact source turn.

Judge atomic clinical claims, not wording. Match synonyms,
   tolerate small numeric differences, but require correct
   category, unit, sign, tooth number, and laterality when
   those attributes are present. A claim is supported only
   when it is supported by the provided atomic evidence and is
    within the question scope. Mark a claim as a contradiction
    when it conflicts with the evidence. Claims not covered by
    the curated evidence are unsupported, but do not
   automatically call them hallucinations if they are
   plausible and not contradicted. Mark hallucination true for
    fabricated, contradicted, or unjustifiably specific
   clinical claims, do not mark a justified uncertainty or
   limitation statement as hallucination.

Return only valid JSON with this structure:
{{
  "matched_evidence_ids": ["evidence_id"],
  "predicted_claims": [
    {{
      "claim": "atomic claim from the model answer",
      "matched_evidence_ids": ["evidence_id"],
      "supported": true,
      "contradiction": false,
      "hallucination": false
    }}
  ]
}}

QUESTION:
{question}

GOLD ANSWER:
{gold_answer}

ATOMIC EVIDENCE:
{evidence}

MODEL ANSWER:
{model_answer}
```

Figure 23: Prompt template for perception verification.

ATOMIC EVIDENCE EXTRACTION PROMPT

```
You are a clinical information extraction specialist preparing
   atomic evidence for longitudinal oral and maxillofacial
   reasoning tasks.

Extract atomic clinical evidence from the current stage only.
```

```
Rules:
1. Extract only facts explicitly stated in this source QA pair,
     do not infer diagnosis or treatment. Prioritize explicitly
     stated diagnoses, classifications, stages, laterality, and
     negative conclusions, and never omit an explicit diagnosis
    .
2. Split compound statements into minimal atomic facts and
    preserve numbers, units, tooth numbers, laterality, and
    negation.
3. Set modality from the source stage. For profile/history
    stages use TEXT_QA, never use DP for patient history or
    chief complaints. Use only modalities listed for the stage.
4. Use concise English and return strict JSON only. Extract at
    most 12 important facts.

Return schema:
{
  "atomic_evidence": [
    {
      "source_turn_id": 5,
      "modality": ["DP"],
      "fact_text": "...",
      "fact_type": "profile|chief_complaint|history|
    facial_finding|occlusal_finding|cephalometric_measurement|
    panoramic_finding|ct_finding|tmj_finding|other",
      "clinical_dimension": "profile|sagittal_relationship|
    vertical_pattern|facial_asymmetry|dental_status|tmj_status|
    treatment_risk|other",
      "normalized": {"field": "...", "value": "...", "unit":
    null, "side": null, "tooth": null}
    }
  ]
}

Patient: $patient_id
Stage: $stage_id
Modalities: $modalities

Source QA pair:
$qa_text
```

Figure 24: Prompt template for atomic evidence extraction.

CROSS-STAGE EVIDENCE-EDGE PROPOSAL PROMPT

```
You are proposing candidate cross-stage clinical evidence edges
    .

Propose candidate pairs where an earlier fact may help
    interpret a later fact through confirmation, remeasurement,
     refinement, conflict, a supported clinical-state update,
    or clinically meaningful corroboration. Related findings
    across modalities or fields may be linked when they jointly
     support the same specific clinical interpretation. A
    different value alone does not prove an update. Do not
```

```
    propose pairs based only on an unrelated shared category or
     simple coexistence.

Rules:
1. Each pair goes from an earlier node to a later node.
2. Preserve entity, anatomical level, scope, direction,
    polarity, and uncertainty.
3. Use only provided ids. Do not infer facts beyond the nodes.
4. Give each pair a fact-specific English reason of at most 15
    words.
5. Return at most $max_edges candidates.

Return strict JSON only:
{
  "candidates": [
    {"source": "<node_id>", "target": "<node_id>", "reason":
    "..."}
  ]
}

Patient: $patient_id

Stage order:
$stage_order

Evidence nodes:
$nodes
```

Figure 25: Prompt template for cross-stage evidence-edge proposal.

CROSS-STAGE EVIDENCE-EDGE REVIEW PROMPT

```
You are independently reviewing candidate cross-stage clinical
    evidence edges.

Classify each candidate once:
- clinical_support: the earlier fact confirms, refines, updates
    , or meaningfully contributes to interpreting the later
    fact. This may include complementary findings from
    different modalities or fields that support the same
    specific clinical interpretation.
- context_consistency: the facts share a clinically relevant,
    fact-specific theme that helps contextualize the later fact
     but is not needed for reasoning.
- reject: the facts are unrelated, incompatible, or connected
    only by a broad category or simple coexistence.

Preserve entity, anatomical level, scope, direction, polarity,
    and uncertainty. Prefer context_consistency when a broad
    finding is related to, but does not specifically support, a
     localized finding. For the same attribute, entity, and
    scope, set relation to exactly one of: confirms for the
    same finding, remeasures for a repeated observation whose
    different value does not establish change, refines for
    compatible added precision, conflicts for comparable
    disagreement, or updates_clinical_state only for a
```

```
    supported real later change. Do not infer an update merely
    because one source appears later.

Return strict JSON only:
{
  "clinical_support": [
    {"source": "<node_id>", "target": "<node_id>", "relation":
    "<optional relation>", "reason": "..."}
  ],
  "context_consistency": [
    {"source": "<node_id>", "target": "<node_id>", "reason":
    "..."}
  ]
}

Candidates:
$candidates

Evidence nodes:
$nodes

Maximum accepted links per category: $max_edges
```

Figure 26: Prompt template for cross-stage evidence-edge review.

ANSWER-GENERATION PROMPT

```
You are a clinical assistant answering a question about ONE
    patient during a staged oral & maxillofacial diagnostic
    workup.

Base your answer on the MEMORY below and any attached images.
    Attached images are observed evidence from the visible
    stages. You MAY reason: using standard clinical knowledge,
    infer diagnoses, treatment strategies, comparisons, and
    expected outcomes that logically follow from the recorded
    findings, even when they are not stated verbatim. This kind
     of clinical reasoning/inference is expected and is NOT
    considered fabrication.

Hard limits:
- Do NOT invent specific facts or numeric values (measurements,
     angles, tooth numbers, dates, named findings) that are not
     present in memory or clearly visible in the attached
    images.
- If the memory and images do not contain enough findings to
    support a defensible answer, do NOT force one: answer with
    whatever IS supported and clearly state what cannot be
    determined and why.

How to answer:
- Answer the QUESTION directly and stay within what it actually
     asks (its modality/stage/topic).
- Make the clinical BASIS explicit. Whenever you give a
    diagnosis, a treatment plan, a surgical plan, or an
    expected outcome, also state the specific findings/evidence
     in memory it is based on --- the measurements, imaging
```

```
    findings, occlusal/skeletal/dental observations across the
    relevant stages and modalities that justify each conclusion
     or plan step (the "because"), and the clinical reasoning
    that connects them.
- For treatment/surgery, give a concise but complete reasoning
    chain: state the treatment goal, the clinically necessary
    sequence, and the key finding that justifies each major
    step. Include only patient-relevant steps, do not enumerate
     generic options, repeat the same evidence, or expand minor
     details that do not change the plan.
- Be concise and information-dense. State each conclusion and
    its strongest supporting evidence once. When the findings
    support a definitive clinical judgement, commit to it
    rather than staying vague.
- Give a complete but concise reasoning chain in no more than
    800 words. Do not use headings, numbered lists, or bullet
    points.

MEMORY (everything the model has observed so far):
$memory

QUESTION:
$question

Write the answer as plain English prose --- no JSON, no
    markdown, no headings. State both the conclusion and the
    supporting findings/evidence and reasoning that justify it.
```

Figure 27: Prompt template for clinical answer generation.

BASE ANSWER-JUDGING PROMPT

```
You are a clinical answer evaluator grading a model's response
    to a memory and reasoning question.

Judge two things:
1. Whether the answer is overall correct for the QUESTION,
    using the REFERENCE ANSWER.
2. For each item in SELECTED EVIDENCE, whether the answer
    correctly covers that clinical fact.

Evidence scope:
- SELECTED EVIDENCE was filtered during task construction and
    contains only the evidence required to answer this question
    . Every listed item is part of the fixed gold evidence set
    and must be judged. Do not introduce an in-scope/out-of-
    scope distinction during evaluation.

Clinical semantic equivalence:
- Accept clinically equivalent or synonymous phrasing as
    covering the same fact, do not require
identical terminology. Examples of equivalent expressions:
* "mandibular excess" / "protruded chin" / "mandibular
    prognathism" ≈ "mandibular protrusion"
* "skeletal Class III" ≈ "bilateral Class III molar
    relationship" (when describing the same relationship)
* "concave facial profile" ≈ "prognathic mandibular profile"
```

```
- Judge on clinical meaning, not surface wording.

Numeric tolerance (apply per fact type):
- Quantitative LINEAR measurements (mm, e.g. chin/menton
    deviation, overjet, overbite, midline shift): the DIRECTION
     / sign / laterality (left vs right, positive vs negative/
    reverse) MUST match, the magnitude is acceptable if it is
    within ±2 mm OR ±20% of the reference value, whichever is
    larger. Do NOT fail an item solely for a small magnitude
    difference within this tolerance.
- ANGULAR measurements (degrees, e.g. SNA/SNB/ANB, nasolabial
    angle): acceptable within ±10°, and any qualitative
    descriptor (acute/obtuse/normal, high-angle/low-angle) MUST
     match.
- CATEGORICAL / CLASSIFICATION findings (Angle or skeletal
    Class I/II/III, presence vs absence of a sign, left vs
    right laterality, crossbite yes/no): must match EXACTLY. A
    wrong class or wrong direction is incorrect / not covered
    --- no tolerance here.

Other rules:
- A selected evidence item is covered only if the model answer
    states the same clinical fact (per the equivalence and
    tolerance rules above) without contradiction.
- Missing a selected evidence item means it is not covered.
- Extra unrelated but non-contradicting content does not by
    itself make the answer wrong.
- Set overall "correct" to true only if the answer captures the
     essential factual content required by the question.

QUESTION:
$question

REFERENCE ANSWER:
$gold

MODEL ANSWER:
$answer

SELECTED EVIDENCE (JSON list):
$selected_evidence

Return strict JSON only, no markdown, no commentary:
{
  "correct": true,
  "reason": "<one short clause>",
  "covered_evidence_count": 0,
  "evidence": [
    {"evidence_id": "<id>", "covered": true, "reason": "<short
    >"}
  ]
}
```

Figure 28: Prompt template for base answer judging.

EVIDENCE-RECALL JUDGING PROMPT

```
You are measuring EVIDENCE RECALL for a model's ANSWER to a
    clinical question.

For EACH item in SELECTED EVIDENCE, decide whether the MODEL
    ANSWER covers that clinical fact (states the same fact,
    synonymous clinical phrasing is fine), using the REFERENCE
    ANSWER as ground truth. Every listed evidence item was
    preselected during task construction as required for this
    task and MUST be judged for coverage. Do not add a scope
    judgment or exclude any item. Return ONLY the ids of the
    covered items, no explanations.

Clinical semantic equivalence:
- Accept clinically equivalent or synonymous phrasing as
    covering the same fact, do not require
identical terminology. Examples of equivalent expressions:
* "mandibular excess" / "protruded chin" / "mandibular
    prognathism" ≈ "mandibular protrusion"
* "skeletal Class III" ≈ "bilateral Class III molar
    relationship" (when describing the same relationship)
* "concave facial profile" ≈ "prognathic mandibular profile"
- Judge on clinical meaning, not surface wording.

Numeric tolerance:
- Linear measurements (mm): direction/sign/laterality must
    match, magnitude acceptable within ±2 mm OR ±20% of the
    reference (whichever is larger).
- Angles (degrees): within ±10° and matching qualitative
    descriptor (acute/obtuse/normal, high/low-angle).
- Categorical / classification (Class I/II/III, left/right,
    presence/absence): must match EXACTLY.

QUESTION:
$question

REFERENCE ANSWER:
$gold

MODEL ANSWER:
$answer

SELECTED EVIDENCE (JSON list):
$selected_evidence

Return strict JSON only, no markdown, no commentary:
{"covered_evidence_ids": ["<id>"]}
```

Figure 29: Prompt template for evidence-recall judging.

RUBRIC JUDGING PROMPT

```
You are scoring a model's ANSWER to a clinical treatment/
    diagnosis question against a scoring RUBRIC checklist.
```

```
For EACH criterion, award points between 0 and its "max_score",
    based on how fully and correctly the model answer
   satisfies the criterion "description". Partial credit is
   allowed. Never exceed a criterion's max_score. Judge only
   against the rubric and the reference answer.

Scoring guidance:
- Primary / preferred plan: if the model correctly identifies
   the patient's MAIN or preferred diagnosis/treatment pathway
   , award high credit on the criteria that cover it, even if
   it omits alternative or edge-case options that may not
   apply to this patient. Only the criteria that explicitly
   target those alternatives should be reduced for their
   absence.
- Reasoning over restatement: reward answers that show clinical
    REASONING --- a causal/inference chain, the sequencing/
   workflow of the plan (e.g. presurgical orthodontics ->
   orthognathic surgery -> postsurgical orthodontics -> dental
    disease management), or why a diagnosis/plan follows from
   the findings. Do not give full credit to a mere restatement
   /summary of raw findings when the criterion calls for
   reasoning or an integrated plan.
- Clinical equivalence: accept synonymous clinical phrasing, do
    not require identical terminology.

QUESTION:
$question

REFERENCE (gold) ANSWER:
$gold

MODEL ANSWER:
$answer

RUBRIC CRITERIA (JSON list, each has name / max_score /
   description):
$criteria

Return strict JSON only, no markdown, no commentary:
{"criteria": [{"name": "<criterion name>", "awarded": 0, "
   reason": "<short>"}], "total_awarded": 0}
```

Figure 30: Prompt template for rubric judging.