Supplemental Notes

Note 1: Problem Formulation

Data modalities in DentVLM

Seven common imaging modalities in oral diagnosis and treatment are included in this study (Figure 1.a, Extended Table 1). Specifically, patients undergo lateral X-ray (LAT) and panoramic X-ray (PAN) by orthopantomograph, and take intraoral images by intraoral cameras or portable edge devices, such as intraoral left view (INL), intraoral right view (INR), intraoral front view (INF), upper dental arch view (UPP), and lower dental arch view (LOW). Collectively, these modalities not only capture dental and maxillofacial structural information at multiple levels, but also offer high clinical practicality and representativeness due to their rapid, convenient, and cost-effective accessibility.

Definition of the 36 Tasks

Then, based on the characteristics of different modalities in representing information, and in order to adapt DentVLM to broader scenarios such as health management and clinical applications, we defined a total of 36 tasks across two major categories (i.e. oral diseases and malocclusion). Meanwhile, to ensure clinical relevance, three professional dentists were invited to define an image-task mapping (Figure 1.a), avoiding situations where a given image could not be used for a specific diagnostic task. Furthermore, based on the label formats of these tasks, we categorized them into 17 general multi-class oral disease tasks (Extended Table 2), 17 multi-class malocclusion tasks (Extended Table 3), and 2 multi-label malocclusion tasks (Extended Table 4). And in the multi-label tasks, the model is required to perform a holistic analysis of the image and comprehensively generate a list of all potentially present diseases.

Outputs formats generated from DentVLM

Finally，for each task t, we created input pairs (I_t, Q_t) for DentVLM, where I_t indicates the specific image and Q_t is the corresponding question. And DentVLM will generate results including answer A_t, rationale R_t, and location L_t, where L_t is empty when t belonging to the malocclusion category. Specifically, the answer A is determined by the label set of each specific task (Extended Table 2-4). The rationale R represents a reasoning path for the current diagnostic result, which integrates prior knowledge and image-specific information. The location L refers to our predefined partitioning of the oral regions (Extended Table 5), because many diseases do not confine themselves on a single tooth but instead affect multiple teeth or broader areas such as the gingiva. Additionally, considering the significant differences among teeth within different horizontal regions, we summarized the following nine descriptions as the complete set of location information: 1) the right posterior region of both the upper and lower dentition, 2) the anterior region of both the upper and lower dentition, 3) the left posterior region of both the upper and lower dentition, 4) the right posterior region of the upper dentition, 5) the anterior region of the upper dentition, 6) the left posterior region of the upper dentition, 7) the right posterior region of the lower dentition, 8) the anterior region of the lower dentition, and 9) the left posterior region of the lower dentition.

Note 2: Data Statistics

Before conducting specific evaluations, we first performed a comprehensive statistical analysis of the constructed dataset to illustrate its characteristics and distribution (Figure 2.c). The complete dataset, including two training sets, test set and clinical study set, encompasses 23,597 patients and 115,784 images, with the proportion difference among the seven modalities being less than 4.6%, demonstrating the large scale and modality well-balance of our dataset.

Statistic of training and test sets

We presented the number of patients, images, and specific VQA pairs for both oral disease and malocclusion tasks across the 1st stage and 2nd stage training set, as well as the test set (Extended Figure 2.a–c). This conveys our design philosophy: during the 1st training stage, we constructed an ultra-large-scale simple VQA set D_1 to align vision and language, facilitating domain adaptation of the model; in the 2nd stage, we curated a high-quality rationale VQA set D_2 to guide diagnostic instruction tuning; finally, a carefully designed test set was constructed to validate the DentVLM and methodology. However, due to the absence of patient information in the oral diseases data, we only illustrated patient demographics for malocclusion tasks within the training and test sets (Extended Figure 2.d–f). Notably, patients under 25 years old constitute over 70% for each subset, and the proportion of female patients exceeds that of males, indicating that the population undergoing orthodontic treatment is primarily composed of adolescents and females. Furthermore, we detailed the label distribution for each multi-class task across these three datasets (Extended Figure 2.g–l), illustrating the comprehensive coverage of various conditions for each task. As a result, although imbalances in inter-task sample sizes and intra-task label distributions are present, the overall distributions closely reflect the real-world clinical scenarios.

Statistic of clinical study set

Then, we separately presented statistical information regarding the clinical study set. We subdivided the tasks into oral diseases (multi-class), malocclusion (multi-class), and malocclusion (multi-label) to show the distribution of patients, images, and data pairs (Extended Figure 3.a). Analogous to the previous patient information absence related to oral disease tasks, we only illustrated the age and gender distribution of patients associated with malocclusion tasks (Extended Figure 3.b). The patients' ages range from 8 to 56 years old. Specifically, children (≤12 years) constitute approximately 22.61%, adolescents (12-18 years) account for 24.04%, young adults (18-25 years) with incomplete skeletal maturity represent 30.74%, and adults (>25 years) comprise 22.61%, as well as the male-to-female ratio was approximately 38%:62%, which demonstrates a consistent data distribution with training and test set. Additionally, we analyzed the relationship between the dentists-annotated answer confidence and the complexity of cases during the construction of the clinical study set (Extended Figure 3.c). The proportion of cases with high-confidence answers decreased as complexity increased. Finally, label distributions for each multi-class task, overall answer confidence, and case complexity statistics were detailed (Extended Figure 3.d–i). These statistics demonstrated that the clinical study set comprehensively covers diverse tasks, with the majority of answers exhibiting high confidence, thereby providing a solid data foundation for rigorous clinical trials. Moreover, case complexities across all tasks adhere to the normal distribution, ensuring the representativeness of our dataset and the generalizability of experimental results.

Note 3: Extended Ablation Studies

Ablation study on training strategy

We first compared the overall effect of the two-stage training pipeline (Extended Figure 4.a). The results demonstrate that performance improves progressively across the two stages and consistently outperforms only the second stage. Then, we investigated the effect of different freezing strategies for the vision encoder during the two-stage training (Extended Figure 4.b). Compared to DentVLM's chosen setting—where the vision encoder is unfrozen in the first stage and frozen in the second—we designed two alternative configurations: 1) unfreezing the vision encoder in both stages and 2) freezing it in both stages. We excluded the configuration that freezes in the first stage and unfreezes in the second, as it contradicts our design philosophy of first establishing robust vision-language alignment. According to the results, DentVLM's current freezing strategy achieved the best performance across all task categories, except for a slight 0.13% drop in the Malocclusion (EN) category compared to the fully unfrozen setting. This not only validates the effectiveness of our freezing strategy but also supports the soundness of our overall training framework and design philosophy. We also conducted an ablation study on the choice of epochs during the two-stage training process (Extended Figure 4.c). The results show that setting the number of epochs to 3 in the first stage yields the best average performance across all four major task categories. Given that the first-stage dataset contains approximately 2.4 million samples, this choice also offers a practical balance between performance and computational cost. For the second stage, performance began to plateau and fluctuate when the number of epochs exceeded 5, so we ultimately set the epoch count for this stage to 5. Finally, we compared full-parameter with LoRA-based fine-tuning across both training stages (Extended Figure 4.d), indicating that full-parameter fine-tuning remains the better choice under the current data scale.

Ablation study on training dataset

Apart from the impact of the different scale of dataset, we also validated the effectiveness of training on a dataset that comprehensively covers both oral disease and malocclusion tasks (Extended Figure 4.e). Compared to training on only a single task category, using the full dataset consistently led to superior performance, indicating that task diversity is crucial for the generalization ability and overall performance of DentVLM. Additionally, we evaluated the impact of the dataset's bilingual composition (Extended Figure 4.f). We found that even when trained on a single language, the model can achieve nearly 70% or even higher performance in the other one. And incorporating bilingual training leads to further improvements. This not only highlights the necessity of constructing multilingual datasets but also suggests that knowledge learned in one language can generalize effectively to another.

Ablation study on inference process

In addition to its performance, we further evaluated DentVLM's robustness in terms of its sensitivity to the input image and text. First, we evaluated its sensitivity to image pixels during inference (Extended Figure 4.h). Specifically, we resized input images under different upper bounds on resolution while maintaining their original aspect ratio. Although we set an upper resolution limit of 512×512 during the first-stage training to reduce computational cost, the model still demonstrated the ability to extract useful information from higher-resolution images to support its reasoning. Moreover, the observation that higher image resolutions yield better performance is consistent with conclusions drawn in the general domain. Interestingly, we also observed that DentVLM achieved its best performance on oral disease-related tasks when using images with a resolution upper bound of 1024×1024. We hypothesize that is because such tasks are relatively straightforward and likely well-represented in the pretraining corpus, making the model more inclined to commonly seen resolution settings such as 1024×1024. Next, we assessed DentVLM's robustness to variation in instruction phrasing (Extended Figure 4.i). Specifically, based on the 9 prompts for each task (Extended Table 6), we compared the model's performance given only one type or uniformly and randomly sampled these instructions. The results show that, across all task categories and training stages, the performance of DentVLM remained stable regardless of prompt variation, which demonstrates its strong robustness to instruction formulation in downstream tasks.

Note 4: Data Allocation and Consistency Evaluation in Clinical Study

To ensure the objectivity of the clinical study, we invited a total of 25 dentists—13 junior dentists and 12 senior dentists—to participate in the evaluation, and employed a carefully designed data allocation strategy to assess both self-consistency for each individual dentist and group-consistency within dentists of the same expertise level (Extended Figure 7.a–c).

Clinical study data allocation

In the specific implementation, according to the clinical study set construction process (Extended Figure 1.b), we divided it into five subsets: D_{idp} and $D_{gv\{i\}}(i = 1, 2, 3, 4)$. This division serves two purposes: firstly, to reduce the burden on each dentist by evenly distributing the workload in D_{idp}, and secondly, to verify both self-consistency (SC) on different finish timeline and group-consistency (GC) within various expertise dentists through repeated data assignments. Specifically, these five subsets are combined then allocated across four experiments, and dataset assigned to each experiment can be expressed as $D_{exi} = \{D_{idp}, D_{idpr}, D_{gv\{i\%4\}}, D_{gv\{(i+1)\%4\}}\}$, where i denotes the clinical experiment index, and D_{idpr} is a replica of D_{idp} (Extended Figure 7.a). To illustrate the allocation strategy, we use 12 senior dentists and D_{ex1} as an example (Extended Figure 7.b). First, D_{idp} is evenly divided into 12 slices corresponding to the number of dentists, with each slice assigned to a different person. Then, each dentist is assigned an additional 15% of the data from D_{idpr} based on their previously allocated slice to verify self-consistency. Finally, dentists are divided into two groups based on the parity of their IDs and each group is assigned a replicated subset derived from the same D_{gv} to validate group-consistency (Extended Figure 7.c, Supplementary Note 3). Meanwhile, considering that each dentist participates in all four experiments under the same ID, different $D_{gv\{i\%4\}}$ subsets are assigned to the respective groups in each experiment to ensure the reliability of consistency assessments. Additionally, D_{idp} within each $D_{ex\{i\}}$ is shuffled to minimize the likelihood that the same data items are repeatedly assigned to the same dentist across experiments, thus reducing potential bias.

For junior dentists, the clinical study is conducted following the same strategy using identical data, but in a completely separate environment from senior dentists to avoid any interference.

Self-consistency for each individual dentist

Specifically, self-consistency was calculated as the probability that a dentist produced consistent diagnostic results for the same data item across different timelines. For dentist k, a dedicated dataset D_{idprk} was used for this evaluation, wherein each data item $i \in D_{idprk}$ had a uniquely paired counterpart $j \in D_{idpk}$ and $j = hash(i)$, representing a repeated diagnostic instance of the same case. Thus, the self-consistency for a dentist can be formally expressed as:

$$SC_k = \frac{\sum_{i\in D_{idprk},\, j\in D_{idpk},\, j=hash(i)} 1_{pred_i=pred_j}}{|D_{idprk}|}$$

where $pred_i$ indicates the prediction of i-th data item by dentist k, and $|D_{idprk}|$ refers the size of D_{idprk}.

Group-consistency within dentists of the same expertise level

Group-consistency is defined as the probability of the most frequently selected diagnosis for a data item among the group of dentists. For a group k consisting of n dentists, let the dataset used for evaluating group-consistency be D_{gvk}. For each data item $i \in D_{gvk}$, let the corresponding diagnostic results from all dentists be $\{pred_i^1, pred_i^2, ..., pred_i^n\}$. Then, the group-consistency of this group can be formally defined as:

$$GC_k = \frac{1}{|D_{gvk}|}\sum_{i\in D_{gvk}} \frac{argmax_{pred}|\{j|pred_i^j=pred\}_{j\in\{1,2,...,n\}}|}{n}$$

where $|D_{gvk}|$ indicates the number of data items of D_{gvk}, and $|\{j|pred_i^j = pred\}_{j\in\{1,2,...,n\}}|$ represents the number of dentists with diagnosis $pred$.

Results of consistency evaluation

We reported the self-consistency and group-consistency results for both junior and senior dentists across three diagnostic settings: Dentist only, Dentist+DentVLM-A, and Dentist+DentVLM-R (Extended Figure 7.d). The average median consistency value across all scenarios exceeds 83.21%, indicating that the results are not heavily influenced by individual subjectivity and are reliable. Furthermore, senior dentists consistently demonstrated higher consistency compared to junior dentists, which aligns with the intuitive expectation that senior clinicians, having more experience and standardized diagnostic practices, produce more stable and reliable outcomes. We also presented the evaluation statistics for each dimension assessed in the response evaluation experiment (Extended Figure 7.e). The median consistency for all dimensions exceeds 60%, with accuracy reaching nearly 80%. This is largely because accuracy involves an objective judgment of the answer’s correctness, whereas the other dimensions—focused on the rationale—are more susceptible to variability due to subjective interpretation. Nonetheless, achieving over 60% consistency across all dimensions suggests that the evaluations are generally aligned and reflect a consensus among most dentists.

Note 5: Voting Strategies in Practical Deployment

In health management at home or intelligent diagnosis at hospital, DentVLM could generate task-specific results for each image. Then, we introduce two voting strategies to aggregate the final result across different imaging modalities, which determines a list of potential diseases. Specifically, for each task i and its related modality list M_i, majority voting consolidates the outputs across different modalities into a single voted decision $mjv_i = argmax_c \sum_{m_i \in M_i} 1_{c_{m_i} = c}$, where c_{m_i} indicates the result of task i in modality m_i. Then we define the score of current task i by majority voting as: $s_i^{mjv} = 1_{mjv_i = gt_i}$, where gt_i denotes the ground truth for task i. In contrast, for each task i, matching voting presents results from all modalities simultaneously, i.e. $mcv_i = \{c_{m_i} | m_i \in M_i\}$. And we define the score of task i by matching voting as: $s_i^{mcv} = 1_{gt_i \in mcv_i}$. Therefore, unlike majority voting, matching voting considers a prediction as successful if any single modality produces the correct result, which establishes the upper bound of patient-level disease list prediction.