Ontology-Guided Prompting for Reasoning in Multimodal Vision-Language Models: An Application to Rare Dental Disease

Kareem Elgohary[1] , **Ali Ayadi**[1] , **KAWCZYNSKI Marzena**[2]
Agnes BLOCH-ZUPAN[2] and **Cédric Wemmert**[1]
[1] SDC team, ICube CNRS UMR7357, University of Strasbourg, Strasbourg, France
[2]Reference Center for Rare Oral and Dental Diseases (CRMR-O-Rares), Oral Medicine and Surgery Department, Hôpitaux Universitaires de Strasbourg (HUS), Strasbourg 67000, France
kareem.elgohary@etu.unistra.fr,marzena.kawczynski@chru-strasbourg.fr
{ali.ayadi, agnes.bloch-zupan, wemmert}@unistra.fr

Abstract

Vision-language models (VLMs) have demonstrated strong generalization across multimodal tasks, enabling applications in medical image interpretation, robotic perception, and education. However, their lack of grounding in domain specific knowledge often leads to hallucinations, especially when applied to out of distribution data where precision and explainability are critical. Prompt engineering provides a lightweight alternative to fine tuning for adapting VLMs to specialized tasks, but remains fragile and lacks guarantees for factual accuracy. Fine tuning, while more robust, is computationally expensive and often impractical in privacy sensitive environments. We focus on a high stakes application: symptom level reasoning in rare dental diseases, such as dental agenesis and enamel defects. These conditions present diagnostic challenges due to low prevalence, overlapping symptoms, and limited labeled data making them an ideal testbed for evaluating the adaptability of general purpose VLMs. We propose an ontology guided prompting framework that enables interpretable, step by step reasoning without model retraining. A domain specific ontology, created with clinical experts, models the rare disease domain of our dataset including disease symptom relationships and supports the generation of chain-of-thought (CoT) prompts. These prompts guide VLMs such as MiniGPT-4, LLaVA, and BLIP-2 to extract medically grounded reasoning from dental images. Our method leverages the models' latent medical knowledge through symbolic constraints and semantic filtering based on ontology terms. We evaluate three prompting strategies: zero-shot, human feedback, and ontology-guidance and assess reasoning quality using F1 score, ontology coverage, and hallucination rate. Results show that ontology-guided prompting significantly improves factual alignment and reduces hallucinations, supporting safe and explainable VLM deployment in clinical domains.

1 Introduction

Recent vision-language models (VLMs), such as MiniGPT-4 [Zhu *et al.*, 2023], LLaVA [Liu *et al.*, 2023], and BLIP-2 [Li *et al.*, 2023], have demonstrated impressive capabilities across multimodal tasks like image captioning, visual question answering, and scene understanding. These models enable general purpose reasoning by jointly processing visual and textual inputs, allowing rapid adaptation to a wide range of applications.

However, when deployed in specialized domains like healthcare, VLMs frequently hallucinate or produce clinically irrelevant outputs [Agarwal *et al.*, 2024]. This is often due to a lack of grounding in domain specific knowledge, which is typically absent from the large scale, general purpose corpora used during pretraining. This issue is particularly critical in high stakes settings like medical diagnosis, where untrustworthy reasoning can compromise safety and adoption.

Prompt engineering has emerged as a lightweight strategy to adapt VLMs to specific tasks without retraining. Techniques such as Few-shot Prompting [Brown *et al.*, 2020], Chain-of-Thought [Wei *et al.*, 2023], and Chain-of-Symbol (CoS) Prompting [Hu *et al.*, 2024] have been developed to enhance reasoning structure, improve consistency, or reduce hallucination. However, many of these methods remain sensitive to prompt phrasing and lack robustness under domain shift. Moreover, they rarely incorporate external knowledge sources. On the other hand, fine tuning offers stronger adaptation but introduces computational cost and raises privacy and reproducibility concerns in regulated fields [Ding *et al.*, 2025]. It also requires a large data regime to work effectively. In most real-world cases, obtaining such data is not feasible especially in domains such as rare diseases.

To address these challenges, we propose an ontology prompting framework that guides pretrained VLMs using an expert curated ontology associated with chain-of-thought prompts. Rather than fine tuning model parameters, our method aligns model reasoning with structured domain knowledge using ontology guidance and an adaptive feedback loop. This allows us to extract interpretable outputs without requiring labeled data or retraining.

As a testbed, we apply our method to rare dental diseases

an underrepresented domain characterized by low data availability, overlapping symptoms, and complex diagnostic logic. We evaluate three pretrained VLMs: MiniGPT-4, LLaVA, and BLIP-2 on a private multimodal dataset of 1,280 annotated dental images spanning 22 diseases and 30 textual symptoms [de La Dure-Molla *et al.*, 2019]. While the domain is medical, our framework is general purpose and applicable to other verticals that require grounded reasoning.

In summary, our contributions are threefold: (1) we introduce a general purpose symbolic prompting framework that integrates ontology based guidance associated with chain-of-thought reasoning to guide vision-language models; (2) we design an adaptive feedback loop that extracts semantically aligned outputs using similarity based filtering and prompt refinement, without requiring fine tuning; and (3) we demonstrate the framework's effectiveness on a real world case study in rare dental disease diagnosis, showing improved reasoning quality and reduced hallucination across three VLMs: MiniGPT-4, LLaVA, and BLIP-2 in a zero-shot, privacy preserving setting.

2 Related Work

The emergence of powerful VLMs, such as MiniGPT-4 [Zhu *et al.*, 2023], LLaVA [Liu *et al.*, 2023], and BLIP-2 [Li *et al.*, 2023], has sparked increasing interest in applying these models to medical image interpretation [Chen *et al.*, 2023; Alayrac *et al.*, 2022]. Despite their strong performance on general purpose multimodal tasks such as visual question answering (VQA) and image captioning, recent studies [Liang *et al.*, 2023; Agarwal *et al.*, 2024] have shown that VLMs frequently hallucinate or generate clinically irrelevant outputs when applied to specialized domains like healthcare.

To improve domain alignment, several approaches have explored the integration of structured knowledge, such as ontologies and knowledge graphs, into large language models (LLMs). Methods include knowledge enhanced pretraining [Yao *et al.*, 2022], retrieval augmented generation [Li *et al.*, 2024], and graph constrained decoding [Luo *et al.*, 2024]. In the medical domain, symbolic knowledge has been used to improve the precision and alignment of factual information in tasks such as the generation of reports and the recognition of named entities [He and others, 2022; Chen and others, 2022]. However, most of these methods rely on fine tuning or require modifying model parameters, which is often impractical in privacy sensitive or compute-constrained environments.

Prompt engineering offers a lightweight alternative that allows model adaptation without retraining. CoT prompting [Wei *et al.*, 2023] has emerged as a prominent strategy for improving reasoning by encouraging step by step outputs. Extensions to knowledge grounded CoT have shown promise in math and scientific domains, but their use in multimodal medical tasks remains underexplored. Moreover, most multimodal evaluations focus on surface-level metrics (e.g., fluency or answer accuracy) rather than symbolic alignment with structured domain knowledge.

Our work bridges these gaps by introducing a symbolic prompting framework that operates entirely at inference time, combining CoT reasoning with an ontology driven feedback loop. To the best of our knowledge, this is the first approach to use ontology guiding prompting with general purpose VLMs for interpretable, symptom level reasoning in multimodal medical data. By avoiding model fine tuning and focusing on inference time alignment, our method offers a practical and scalable solution for trustworthy AI in low resource and clinical settings.

3 Methodology

3.1 Ontology Construction

Following the Ontology Development 101 guidelines [Noy *et al.*, 2001], we constructed a modular, domain specific ontology (Figure 1) to formally encode structured clinical knowledge for rare dental diseases. We defined core biomedical classes such as `Patient`, `Disease`, `Symptom`, `Image`, `Gene`, and `Region`, etc, along with domain specific subclasses including `XRay`, `RGB`, `Occlusion`, `Texture`, etc.To capture semantic relationships between clinical entities, we introduced object properties such as `has_symptom`, `shows`, `symptom_of`, which link patients, images, regions, and diagnoses. To enable visual grounding and structured reasoning, we modeled each image as composed of anatomical regions, which may contain localized symptoms. These regions are further described by fine grained features such color, and transparency capturing visual cues relevant to clinical interpretation. This structure allows textual symptom evidence to be linked hierarchically, from pixel-level cues to diagnostic categories. We populated the ontology using a private dataset of 1,280 annotated dental images spanning 22 rare conditions and 30 textual symptoms [de La Dure-Molla *et al.*, 2019]. The ontology was serialized in OWL format. Its modular design supports future extensions and integration with external ontologies.

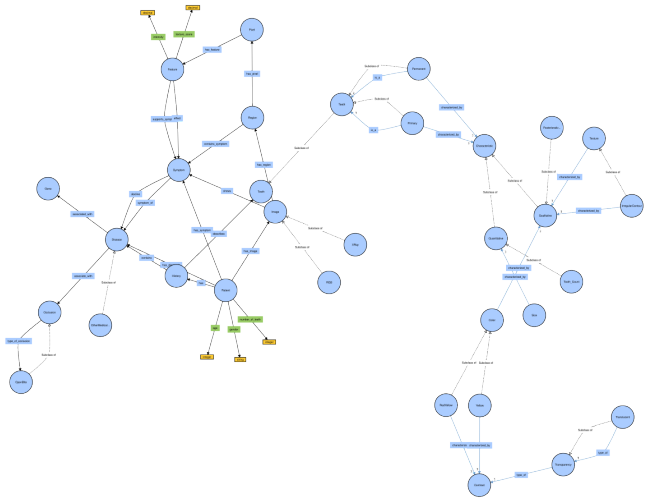

Figure 1: The dental rare disease ontology.

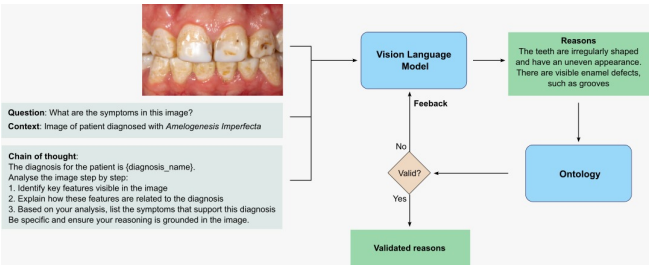

Figure 2: Ontology-constrained prompting and feedback loop. Each reasoning output is filtered, refined, and selected based on semantic alignment to symbolic constraints.

3.2 Ontology-Guided Prompting and Feedback Loop

We introduce a symbolic chain-of-thought prompting framework that leverages structured clinical knowledge from our ontology to guide VLMs toward interpretable, symptom level reasoning. Each prompt follows a structured format mimicking diagnostic logic: it asks the model to describe visual features, identify observable symptoms, and explain how they relate to the given diagnosis.

Our ontology-guided prompting method dynamically injects symbolic information such as symptom terms and their relationships from the ontology into the prompt. These constraints explicitly steer the model's reasoning during inference, grounding it in clinically relevant semantics and reducing reliance on hallucinated associations.

To operationalize this strategy, we implement an adaptive feedback loop (Algorithm 1). For each image diagnosis pair, the model generates up to five reasoning attempts. Each output is cleaned, segmented, and filtered using semantic similarity via Sentence-BERT [Reimers and Gurevych, 2019] against ontology defined symptoms. If the extracted reasoning fails to meet a similarity threshold, the prompt is automatically refined by reinforcing the symbolic constraints from the ontology. The response with the highest alignment score is selected as the final output.

Figure 2 illustrates this iterative refinement process, where model outputs are validated and improved through tight coupling with domain specific symbolic knowledge.

We treat the number of reasoning attempts N in the feedback loop as a tunable hyperparameter. In our experiments, we set $N = 5$ based on empirical stability, where most prompts converged within 2–3 iterations. The loop stops when a high confidence symptom set is extracted or the maximum attempts are exhausted.

4 Experiments

We evaluate our symbolic prompting framework using three pretrained (VLMs): *MiniGPT-4*, *LLaVA (v1.6 Mistral)*, and *BLIP-2*. All models are used in their released, pretrained form without any fine tuning, allowing us to test their zero-shot generalization to an underrepresented and privacy sensitive domain such as rare dental diseases. To ensure lightweight deployment, all models were quantized to 4-bit precision using the BitsAndBytes library and executed on a single

Algorithm 1 Ontology-Guided Prompting and Feedback Loop

Input: Image I, Diagnosis D, Ontology symptoms S_{gt}, Attempts N, VLM
Output: Best reasoning response r^*, predicted symptoms s^*

Initialize: prompt $\leftarrow$ generate_prompt(D)
for $i = 1$ to N **do**
  $r_i \leftarrow \text{VLM.generate}(I, \text{prompt})$
  $f_i \leftarrow$ filter and clean(r_i)
  $s_i \leftarrow$ extract_symptoms(f_i, S_{gt})
  $score_i \leftarrow$ cosine_similarity(s_i, S_{gt})
  if $score_i <$ threshold **then**
    prompt $\leftarrow$ refine_prompt(prompt, s_i, S_{gt})
  else
    break
  end if
end for
Select best: $\langle r^*, s^* \rangle \leftarrow \arg\max_i \ score_i$
return r^*, s^*

NVIDIA GeForce RTX 3060 GPU (12GB). This setup reflects real world constraints in clinical environments where computational resources and patient data availability are limited.

We compare three prompting strategies: *(1) Base prompting*, a generic instruction format with no medical specific adaptation; *(2) Human feedback prompting*, where templates were refined through iterative development with a clinical experts to improve clarity and specificity (though no human involvement occurred during final evaluation); and *(3) Ontology-guided prompting*, where disease specific symptom terms from the ontology were dynamically injected into the prompt. Each model received up to five reasoning attempts per image diagnosis pair, with responses filtered using semantic similarity to ontology terms via Sentence-BERT. The best scoring output was selected for evaluation. This inference pipeline was kept consistent across all models and strategies to ensure fair comparison.

5 Results and Discussion

We report the performance of MiniGPT-4, LLaVA, and BLIP-2 under three prompting strategies: base (zero-shot), human feedback, and ontology-guided.

5.1 Evaluation Metrics

We evaluated model performance on symptom identification using four clinically relevant metrics: *Precision* measures the proportion of correctly predicted symptoms out of all extracted ones; *F1 score* captures the harmonic mean of precision and recall, reflecting a balanced view of accuracy; *Ontology coverage* quantifies the proportion of relevant ontology-defined symptoms recovered by the model; *Hallucination rate* denotes the proportion of symptoms generated by the model that do not appear in the ontology. All metrics were computed using set based comparisons. Sentence-BERT was used to compute cosine similarity between predicted and

ground truth terms, and natural language processing (NLP) preprocessing was handled via SpaCy. We also report *inference time* in milliseconds to highlight the practical feasibility of our lightweight pipeline.

5.2 Results

Table 1: Model performance across prompting strategies. OC = Ontology Coverage, HR = Hallucination Rate, T = Inference Time (ms)

Model	Condition	Precision	F1	OC	HR	T
LLaVA	Base	70%	72%	0%	95%	4
	Ontology-Guided	100%	**94%**	90%	0%	6
	Human Feedback	89%	80%	0%	60%	6
BLIP-2	Base	59%	62%	0%	97%	5
	Ontology-Guided	80%	**85%**	80%	64%	7
	Human Feedback	75%	70%	0%	67%	9
MiniGPT-4	Base	49%	47%	0%	95%	10
	Ontology-Guided	60%	**62%**	45%	46%	12
	Human Feedback	59%	58%	0%	40%	12

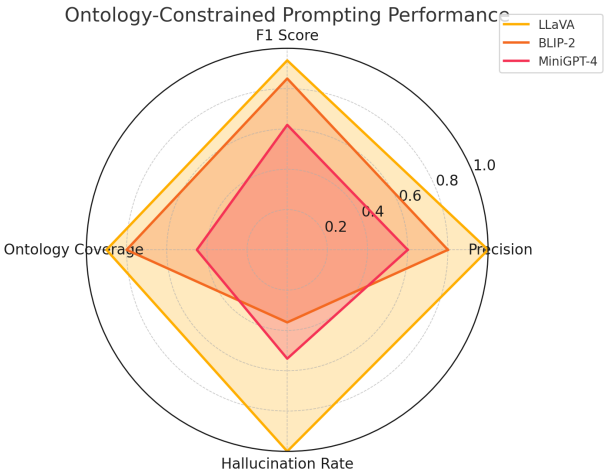

Figure 3: Radar plot comparing VLMs under ontology-constrained prompting. Metrics are normalized; hallucination rate is inverted for alignment.

Table 1 summarizes model outputs across four evaluation metrics: precision, F1 score, ontology coverage, and hallucination rate alongside inference time. Across all models, ontology-guided prompting consistently yielded the best performance. LLaVA achieved the strongest results, with 100% precision, 94% F1 score, and 90% ontology coverage demonstrating the effectiveness of strict symbolic filtering in eliminating hallucinated outputs. Notably, while this precision reflects perfect alignment with the ontology, it may omit undocumented but clinically valid symptoms.

BLIP-2 also showed substantial improvements with ontology-guided prompts, increasing F1 from 62% to 85% and reducing hallucination rate from 97% to 64%. MiniGPT-4, although showing the least improvement, still saw hallucinations reduced by half and F1 boosted to 62%.

Base prompts exhibited the highest hallucination rates (up to 97%) and near-zero ontology coverage, confirming the unreliability of unguided reasoning in specialized domains. Human feedback improved consistency but did not match the precision or interpretability offered by ontology-guided prompting. we visualize their normalized metric scores using a radar plot (Figure 3). LLaVA demonstrates strong and balanced performance across all criteria, followed closely by BLIP-2. MiniGPT-4 shows lower coverage and F1 but benefits from symbolic constraints in reducing hallucinations. This visualization highlights the overall robustness of ontology-guided prompting in maintaining factual consistency and structured reasoning across models.

To evaluate the stability and effectiveness of the adaptive feedback loop, we conducted a convergence analysis on 594 samples. We observed that the loop consistently converged within the maximum budget of five iterations. The mean number of iterations to convergence was 5.0, with all samples stabilizing before reaching the cap. After convergence, the hallucination rate dropped to 0%, confirming that the loop effectively filtered spurious outputs. The mean F1 score was 71.3% (±36.1%), and average ontology coverage reached 64.9%, demonstrating improved semantic alignment. These results provide quantitative evidence that the feedback loop not only stabilizes prediction quality but also boosts factual consistency via symbolic filtering.

5.3 Discussion

These results confirm that symbolic prompting enhances reasoning performance in clinical image analysis without requiring fine tuning. The gains were consistent across models and metrics, showing that pretrained VLMs can be steered to produce domain relevant outputs using structured prompts and ontology based guidance.

The pipeline is also efficient supporting real time inference with 4-bit quantized models on a single 12GB GPU making it practical for deployment in low resource or privacy sensitive environments. The symbolic feedback loop offers transparency and interpretability, allowing clinicians to inspect intermediate reasoning steps.

However, the method's success is contingent on the completeness of the ontology. High precision may hide missed symptoms if the ontology lacks coverage. Addressing this requires dynamic ontology expansion, either through experts annotation or automatic concept mining from generated outputs.

The completeness of the ontology directly affects the system's recall and interpretability. To mitigate this bottleneck, we plan to explore ontology expansion mechanisms, including expert in the loop annotation, active learning, and concept mining from large medical corpora or model generated outputs.

While this work focused on symptom extraction, future extensions could incorporate probabilistic reasoning or structured diagnosis generation. The framework is generalizable and could be applied to other domains e.g., dermatology or radiology where symbolic constraints and interpretability are equally critical.

While our quantitative metrics capture overall performance, we also observed qualitative failure modes. In some

cases, the model generated plausible but out of ontology symptoms, which were filtered out, lowering recall. Future work will include detailed error analysis and case studies to better understand such failures and improve interpretability.

While our method shows strong performance across three VLMs, we acknowledge that the dataset is relatively small and imbalanced, with certain rare diseases represented by only one or two images. This reflects the inherent scarcity of annotated multimodal data in rare disease contexts. Although this setup mirrors real world clinical challenges, it may limit the generalizability of our findings. Future work will explore dataset expansion and evaluation on additional public or cross domain benchmarks to better assess robustness and transferability

6 Conclusion

We presented a symbolic prompting framework that guides general purpose VLMs to perform interpretable, symptom level reasoning in rare dental disease cases. By integrating an expert curated ontology with structured chain-of-thought prompts, we enabled models such as LLaVA, BLIP-2, and MiniGPT-4 to generate clinically relevant outputs without any fine-tuning.

Our results show that ontology-guided prompting significantly improves factual alignment, reduces hallucinations, and achieves high precision across models, all within a lightweight, low resource setup. This highlights the practical value of combining symbolic medical knowledge with large scale multimodal models.

The framework offers a scalable and interpretable approach for deploying VLMs in vertical medical domains where data is scarce and explainability is critical. While the current study is limited to symptom extraction, future extensions will explore diagnosis level reasoning, dynamic ontology expansion, and expert in the loop validation. This work contributes toward safe, deployable, and knowledge grounded artificial intelligence systems for clinical decision support.

Acknowledgment

This work was supported by the Interdisciplinary Thematic Institute HealthTech, as part of the ITI 2021–2028 program of the University of Strasbourg, CNRS and Inserm, with funding from IdEx Unistra (ANR-10-IDEX-0002) and SFRI (STRAT'US project, ANR-20-SFRI-0012) under the French Investments for the Future Program.

We thank the Reference Center for Rare Oral and Dental Diseases (CRMR-O-Rares), Oral Medicine and Surgery Department, Hôpitaux Universitaires de Strasbourg (HUS), for providing the annotated dataset and clinical expertise. This data contribution was supported through collaborative clinical research and structuring initiatives, including: e-GenoDENT project (Fonds d'Intervention Régionale FIR, ARS Grand Est, 2019–2022); Fondation Maladies Rares e-health co-design workshop (2019); AMI Economie numérique Grand-Est i-Dent (2020–2021); Impulsion recherche Filière TETECOU (2020, 2022); Bpifrance (as part of the digital health acceleration strategy of DNS and SGPI, Health Data Hub, "France 2030" strategy, 2021–2023); Fondation Force DIAGNODENT (2023–2026); ANR 3DBioDENT (ANR-23-CE17-0048-01, 2023–2026); and the MIG F04 for CRMRs, DGOS, French Ministry of Health and Prevention.

References

[Agarwal *et al.*, 2024] Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. Medhalu: Hallucinations in responses to healthcare queries by large language models, 2024.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[Brown *et al.*, 2020] 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, Clemens Winter, Christopher Hesse, Mark Chen, 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, 2020.

[Chen and others, 2022] Shaoxi Chen et al. Knowledge-aware prompt tuning for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

[Chen *et al.*, 2023] Yuwei Chen, Yixuan Gao, Zhongyu Yang, et al. Medblip: Medical bootstrapped language-image pretraining for bi-modal understanding and generation. *arXiv preprint arXiv:2309.02285*, 2023.

[de La Dure-Molla *et al.*, 2019] Muriel de La Dure-Molla, Benjamin Philippe Fournier, Maria Cristina Manzanares, Ana Carolina Acevedo, Raoul C Hennekam, Lisa Friedlander, Marie-Laure Boy-Lefèvre, Stephane Kerner, Steve Toupenay, Pascal Garrec, Brigite Vi-Fane, Rufino Felizardo, Marie-Violaine Berteretche, Laurence Jordan, François Ferré, François Clauss, Sophie Jung, Myriam de Chalendar, Sebastien Troester, Marzena Kawczynski, Jessica Chaloyard, International Group of Dental Nomenclature, Marie Cécile Manière, Ariane Berdal, and Agnès Bloch-Zupan. Elements of morphology: Standard terminology for the teeth and classifying genetic dental disorders. *Am. J. Med. Genet. A*, 179(10):1913–1981, October 2019.

[Ding *et al.*, 2025] Yi Ding, Lijun Li, Bing Cao, and Jing Shao. Rethinking bottlenecks in safety fine-tuning of vision language models, 2025.

[He and others, 2022] Jingqing He et al. Infusing knowledge into pretrained language models: A survey. *arXiv preprint arXiv:2212.03551*, 2022.

[Hu *et al.*, 2024] Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. Chain-of-symbol prompting elicits planning in large langauge models, 2024.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[Li *et al.*, 2024] Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Hui Xue, and Jun Huang. On the role of long-tail knowledge in retrieval augmented large language models, 2024.

[Liang *et al.*, 2023] Paul Pu Liang, Zhirui Wu, et al. Vila: Improving vision-language alignment with synthetic instruction tuning. *arXiv preprint arXiv:2306.17107*, 2023.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[Luo *et al.*, 2024] Linhao Luo, Zicheng Zhao, Chen Gong, Gholamreza Haffari, and Shirui Pan. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models, 2024.

[Noy *et al.*, 2001] Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[Wei *et al.*, 2023] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[Yao *et al.*, 2022] Yuan Yao, Haoran Wang, et al. Kalm: Knowledge-augmented language model for long document summarization. *arXiv preprint arXiv:2202.04092*, 2022.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

7 Appendix

7.1 Model Descriptions

We used three publicly available vision-language models in this study: MiniGPT-4, LLaVA (v1.6 Mistral), and BLIP-2. All models were accessed through Hugging Face and used in their pretrained form without modification. MiniGPT-4 integrates a CLIP visual encoder with the LLaMA language model and supports conversational, multi-turn visual reasoning. LLaVA (Large Language and Vision Assistant) follows a similar architecture, aligning vision features with the Mistral-based LLaMA-2 model for grounded image understanding. BLIP-2 employs a lightweight vision-to-language bridging module, optimized for direct image-to-text generation. Our goal was not to benchmark model internals, but to compare their reasoning behavior under symbolic prompting. All models were quantized for efficient inference using 4-bit precision

7.2 Dataset

Dental anomalies are key indicators in diagnosing rare genetic disorders [de La Dure-Molla *et al.*, 2019]. The dataset used in this study is derived from this work, which provides comprehensive phenotypic characterization of rare dental anomalies. This dataset was chosen for its rich representation of rare oral diseases and its multimodal nature, making it suitable for integrating medical imaging with domain specific knowledge.

Dataset Composition

The dataset size 1280 images comprises:

- **Images:** High-resolution X-ray and RGB dental images capturing anomalies such as dental agenesis, supernumerary teeth, and enamel defects.
- **Textual Information:** Disease names, textual symptom lists, and associated genetic markers for each image.
- **Metadata:** Patient-specific attributes including age and gender.

Dataset Statistics

The dataset includes 114 patients who underwent examinations with X-ray and RGB imaging over different time spans. It comprises 22 diseases and 30 unique symptoms. The patient age range is 7 to 45 years. Some diseases have only a single image, illustrating dataset sparsity and annotation complexity.