İSTANBUL UNIVERSITY PRESS

Journal of Data Analytics and Artificial Intelligence Applications, D3AI

https://doi.org/10.26650/d3ai.1726054
Submitted: 24.06.2025
Revision Requested: 30.06.2025
Last Revision Received: 06.07.2025
Accepted: 06.07.2025
Published Online 21.07.2025

Journal of Data Analytics and Artificial Intelligence Applications

Research Article

Open Access

Evaluating the Utility of Generative AI Models for Synthetic Medical Image Generation: A Study on Panoramic Dental X-Rays

Elif Yilmaz [1], Abdullah Huzeyfe Köse [2,3] & Rahime Yilmaz [4]

[1] İstanbul University, Restorative Dentistry, Institute of Graduate Studies in Health Sciences, İstanbul, Türkiye
[2] İstanbul Technical University, Department of Electrical and Electronic Engineering, İstanbul, Türkiye
[3] Altınbaş University, Department of Electrical and Electronic Engineering, İstanbul, Türkiye
[4] İstanbul Aydin University, Department of Computer Engineering, İstanbul, Türkiye

Abstract

Integrating artificial intelligence (AI) into dental practice can streamline processes that are otherwise manual and time-consuming. However, implementing AI in dentistry poses challenges, primarily due to the limited availability of clinical data, especially for rare cases, compounded by ethical and regulatory constraints. To address these challenges, synthetic data generation has increasingly been recognised as a promising solution. In this study, we aimed to evaluate the clinical usability of AI-generated panoramic dental X-rays produced by large-scale multimodal generative AIs, specifically GPT o4 mini high and Copilot, using textual prompts. These synthetic images were generated across multiple categories, enabling a comprehensive evaluation of the models' capabilities. The primary objective was to determine whether AI-generated images, produced from structured textual prompts and real sample descriptions, could be used for learning-based model training. For this purpose, an expert dentist evaluated the 61 generated synthetic images from a clinical perspective, focusing on their visual realism, anatomical accuracy, clinical/educational utility, image clarity, and absence of artefacts. Both models generated images rated satisfactorily across all quality dimensions, with GPT o4 mini high achieving significantly superior anatomical fidelity and realism than Copilot. It was observed that neither system could consistently reproduce fine anatomical details with pixel-level accuracy, a limitation that may prevent its use in applications requiring exact replication. Despite these limitations, the high quality and realism of the synthetic images highlight the potential of multimodal generative AIs. Thus, the study's findings suggest that synthetic panoramic images can reduce the reliance on scarce patient data, thereby accelerating the development of AI applications in dentistry.

Keywords

Synthetic Data Generation · Dental Radiography · Multimodal Generative AI · Clinical Usability · Data Scarcity in Dentistry

Citation: Elif Yilmaz, Abdullah Huzeyfe Köse, and Rahime Yilmaz. 2025. Evaluating the utility of generative AI models for synthetic medical image generation: A study on panoramic dental X-Rays. *Journal of Data Analytics and Artificial Intelligence Applications*, Advance Online Publication. https://doi.org/10.26650/d3ai.1726054

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Yilmaz, E., Köse, A. H. & Yilmaz, R.

Corresponding author: Rahime Yilmaz rahime1yilmaz@gmail.com

1. Introduction

In recent years, synthetically generated medical images have gained significant attention as a potential remedy for the scarcity of annotated medical imaging data, which improves the understanding of healthcare professionals engaged in AI research [1, 2]. The limited availability of well-annotated and publicly accessible dental imaging datasets poses an essential challenge to the development of AI research and clinical applications in dentistry [3, 4]. Typically, dental datasets are derived from real patients, so their use is subject to ethics approval as well as institutional permission, which can significantly lengthen the research process. In particular, obtaining sufficient radiographic data for rare or specific dental conditions can be challenging because such data are rarely available in public datasets. Consequently, researchers aiming to create AI models for clinical decision support must often wait for years to collect sufficient relevant and authorised patient data. In response to this challenge, synthetically generated medical images have gained attention as a potential alternative for creating AI models [5, 6]. Recent advances in generative artificial intelligence have made it possible to create high-quality, realistic medical images based on textual prompts without requiring extensive real datasets [7-9]. Advances in generative artificial intelligence now enable the creation of high-quality, realistic medical images from textual prompts. Vision-capable generative models such as GPT o4 mini high (OpenAI) and Copilot (Microsoft) have demonstrated advanced capabilities in producing structured, clinically relevant images from detailed prompts [10-13]. We selected GPT o4 mini high and Copilot primarily due to their wide accessibility, popularity, and ease of use as image generation tools. The aim of this study was to explore whether dental professionals or experts without detailed knowledge of diffusion models or advanced AI techniques can readily access these tools to generate clinically useful synthetic images for research and educational purposes. While other AI tools were also tested, many remain underdeveloped or failed to produce images due to security restrictions. Ultimately, GPT o4 mini high and Copilot were chosen as the most suitable models, balancing ease of access and functionality, making them practical options for broader clinical and research adoption [14].

Building on this foundation, in this study, we explore whether AI models can generate clinically usable dental radiographs based solely on structured textual prompts describing specific oral and dental characteristics. To systematically investigate this, we designed a methodological prompting strategy consisting of seven tasks: General Prompt, which involves generating images without any specific details about the X-ray; Specific Prompt, focusing on generating images with localised pathological features; Similarity, guiding image-conditioned generation based on real X-rays; Blend Generation, combining multiple inputs; Adjustment, modifying specific parts of images such as root canals or decays while preserving the rest; Extra Category, addressing rare or specific clinical scenarios; and Detailed Prompt, used to evaluate whether prompt specificity affects the generated image quality. For each task, multiple images were generated per model and evaluated through a structured clinical protocol. The evaluation is primarily based on the assessment of an expert dentist, who ensures that the generated images fulfil five key criteria: visual realism, anatomical correctness, clinical and educational usability, image clarity, and absence of artefacts [1, 15]. The dentist was asked to evaluate each generated image using a scale ranging from very satisfying to very unsatisfying. These results were then converted to a Likert scale. Following this expert evaluation, statistical analysis was carried out to draw further conclusions from the data. In addition, the Blind or Referenceless Image Spatial Quality Evaluator (BRISQUE), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) metrics were employed as supportive quantitative measures to complement the dentist's assessments [16-18]. Although standardised tools for assessing synthetic image quality in dentistry

have not yet been established, these metrics provide valuable supplementary indicators [19, 20]. To assess the models' success, the dentist also compared and evaluated the image enhancements and adjustments. This comprehensive assessment is critical given the known challenges in AI image synthesis, such as hallucination artefacts and limited anatomical diversity, which can impact downstream applications in AI training and clinical education [2, 15]. Through this evaluation, we aimed to critically assess the current capabilities and limitations of these AI tools in generating synthetic dental radiographs, while exploring their potential to enhance dataset availability and support AI-driven research and education in dentistry. Based on these assessments, we developed the following research questions:

RQ1: How successful are large-scale multimodal generative AI models, specifically GPT o4 mini high and Microsoft Copilot, in generating clinically usable synthetic dental radiographers from structured textual prompts?

RQ2: How effective are the generative AI models in producing image-conditioned and blended synthetic dental radiographs, and to what extent can this approach provide medically valid and quantitatively reliable data to address the scarcity in dental imaging datasets?

RQ3: To what extent do prompt variations, such as general versus detailed prompts, image-conditioned generation, and blending or adjustment techniques, affect the quality, diversity, and clinical validity of the generated synthetic dental images, especially in rare or complex dental scenarios?

The system architecture shown in Figure 1 illustrates the overall workflow for prompt-based dental X-ray generation and assessment.

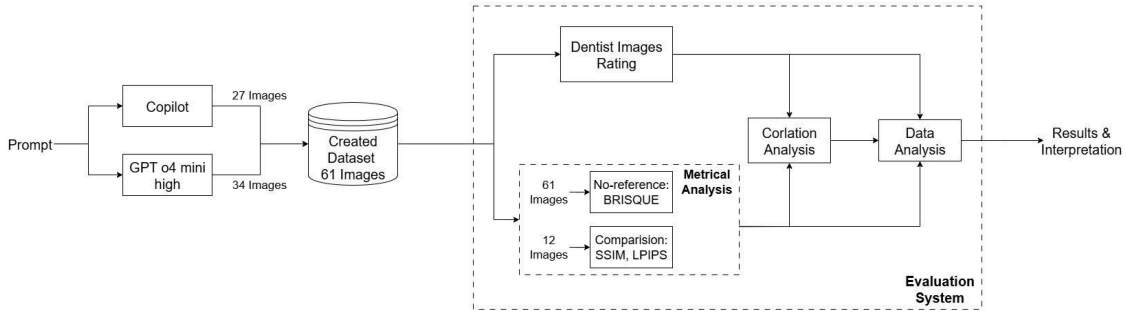


Figure 1. System Architecture for Prompt-Based Dental X-ray Generation and Assessment

The findings of this study suggest that synthetic dental radiographs generated via prompt-based AI tools are promising. As these technologies continue to evolve, they may significantly reduce the reliance on real patient data and accelerate the development of intelligent systems in the field of dentistry. In summary, our methodology integrates commonly used generative AI tools with a rigorous, clinically informed evaluation framework, aiming to demonstrate the feasibility of using accessible AI models to produce synthetic dental radiographs that can support research, education, and potentially augment clinical datasets.

The remainder of this paper is organised as follows. Section 2, Related Work, reviews prior studies relevant to synthetic medical image generation and AI applications in dentistry. Section 3, Methodology, details the study's approach, including the generative models employed (3.1), the image generation prompting strategy (3.2), and the clinical evaluation protocol (3.3), which encompasses both qualitative expert evaluations (3.3.1) and quantitative analysis (3.3.2). Section 4, Evaluation and Results, presents the dentist evaluation results (4.1), including descriptive (4.1.1) and statistical analyses (4.1.2), followed by objective metric evaluations such

as no-reference image evaluation (4.2.1), paired image evaluation (4.2.2), and expert side-by-side comparisons (4.2.3). The correlation between objective metrics and dentist evaluations is discussed in Section 4.3. Section 4.4 compares image adjustments, covering both the image adjustment evaluation (4.4.1) and the image enhancement evaluation (4.4.2). Section 5, Discussion, provides an in-depth interpretation of the findings. Section 6, Threats to Validity, addresses the potential limitations of the study. Finally, Section 7, Conclusion, summarises the key outcomes and suggests directions for future research.

2. Related Work

The generation of high-resolution, diagnostically accurate medical images is inherently complex due to the intricate anatomical structures and the necessity for clinical precision [21]. Dental panoramic radiographs, in particular, pose unique challenges because of their wide aspect ratio and the need to capture highly detailed features across a broad anatomical area, which can often result in blurring or deformation if not carefully managed [21]. To address these challenges, early research primarily focused on generative adversarial networks (GANs) trained on specific medical datasets. While GANs were capable of producing realistic images, they faced limitations such as mode collapse, which restricts diversity, and complex training dynamics [22]. More recently, diffusion models have emerged as a promising alternative, overcoming many shortcomings associated with GANs [23]. Both GANs [21] and diffusion models [24] have become important focal points in the synthesis of dental X-ray images, reflecting ongoing advancements in this specialised area of medical imaging.

A notable example of the GAN approach is the generation of realistic intraoral images using progressive GAN architectures. This study addresses the challenge of limited clinical intraoral image data for dental education and AI applications due to patient privacy concerns [25]. This study used a large dataset of paediatric intraoral images to train Progressive Growing GANs (PGGAN) to generate synthetic intraoral images at various resolutions. In this study, the generated images were evaluated both quantitatively and by paediatric dentists, who attempted to distinguish real from synthetic images. The study results showed that images at 512×512 pixels or lower were visually indistinguishable from real images. Although higher-resolution images (1024×1024) exhibited some detectable artefacts, the overall results of the study suggest that GAN-generated intraoral images can effectively supplement clinical data for dental education and AI model training, while avoiding privacy issues. Similar to this study, our research demonstrates the feasibility and clinical acceptability of AI-generated dental images to overcome data scarcity by generating synthetic panoramic dental X-rays using multimodal AI tools (GPT o4 mini high and Copilot). While this study focuses on intraoral photographic images generated via GANs, our study explores synthetic radiographs generated from textual prompts using large-scale multimodal models.

Another study using GANs presents a novel method for generating virtual tooth images to supplement restricted data for privacy protection in the dental field. The authors proposed a GAN-based image-to-image translation model, pix2pix, combined with a contextual reconstruction fill (CR-Fill) inpainting technique to generate realistic virtual tooth images from intraoral scanner data [26]. Pix2pix converts sketch outlines derived from real tooth images into virtual tooth images, enabling multiple variations from the same input, while CR-Fill refines these images by restoring missing tooth areas and preserving detailed features such as tooth fissures. Both studies address data scarcity in dental imaging via synthetic data generation but differ in imaging modalities, panoramic radiographs versus intraoral scans, and generative techniques, multimodal prompting of large AI models versus GAN-based image-to-image translation and inpainting. Our

study focuses on panoramic radiograph synthesis for diagnostic AI support, whereas the other studies target 3D intraoral scan-based tooth image generation and restoration of missing tooth areas for prosthetic and restorative applications.

One of the most recent GAN studies has also explored the generation of synthetic dental images, focusing on improving training stability, image quality, and clinical relevance to address data scarcity and privacy concerns in dental AI applications [27]. In this study [27], a deep convolutional GAN (DCGAN) with a Wasserstein loss and gradient penalty (WGAN-GP) was trained on a dataset of 2,322 radiographs of varying quality. The authors implemented extensive preprocessing, including cropping and denoising, to standardise the inputs while preserving anatomical variability. Similar to our evaluation strategy, clinical experts assessed the generated images for anatomical visibility and realism. The findings indicated that while many synthetic radiographs moderately depicted key structures such as the mandibular canal and trabecular bone, artefacts remained a limitation in contrast to our study. This study highlights a trade-off between models optimised for finer anatomical details and those producing clearer, sharper overall images. However, as an exploratory effort, the authors suggest that further refinement and evaluation are needed before clinical deployment and that the use of more advanced generative techniques, such as diffusion models for high-resolution image generation, should be considered.

The successful implementation of generative AI in dental X-ray synthesis is not a straightforward application of general image generation models. It necessitates a deep understanding of the unique dental anatomy, the physics of dental imaging, and modality-specific challenges. Following the development of GANs, numerous studies have investigated the use of advanced models, such as diffusion models, to create realistic synthetic dental images, aiming to overcome data limitations and enhance the training of AI systems in dentistry. Diffusion models (DMs) have demonstrated superior performance compared to traditional generative methods such as GANs in terms of both image fidelity and comprehensive distribution coverage [28]. DMs for dental imaging have evolved with specialised architectures and training protocols designed to prioritise anatomical accuracy and diagnostic relevance rather than mere photorealistic appearance [29]. These models employ iterative denoising processes and domain-specific optimizations to faithfully replicate fine dental structures and pathological features, ensuring clinical utility and precision in synthetic radiographs [30, 31]. In the literature, the transition from GAN-based models to diffusion models has been noted as a significant advancement in the generation of synthetic medical images. Although diffusion models have demonstrated superior stability and image quality in various medical imaging tasks, their application specifically in dental imaging has been relatively limited. Therefore, to explore potential approaches, we reviewed several recent studies employing diffusion models for medical image synthesis in related fields such as brain MRI, leukaemia cell imaging, and chest CT scans. These studies highlight the potential of diffusion models to generate realistic synthetic data that can effectively augment limited datasets.

Khader et al. [32] presented a novel approach for generating high-quality synthetic 3D medical images using denoising probabilistic diffusion models (DDPMs) [32]. In their study, similar to our evaluation strategy, two radiologists quantitatively rated the quality of the synthesised images based on “realistic image appearance,” “anatomical correctness,” and “consistency between slices.” The authors demonstrate that their latent diffusion model effectively generates diverse and anatomically precise 3D images, enabling the efficient and realistic synthesis of volumetric data from MRI and CT scans across multiple anatomical regions (brain, chest, breast, and knee). While Khader et al. advanced the field through technical innovations in 3D image synthesis and demonstrated clear benefits for segmentation tasks, our study investigates the

accessibility and clinical applicability of AI-generated 2D dental images for diagnostic support. Importantly, this study [32] shows that pre-training segmentation models with synthetic data generated by their approach significantly improves performance in low-data settings, highlighting the value of diffusion models for data augmentation and privacy-preserving AI development in medical imaging. The model in this study [32] also outperforms traditional GANs in terms of image diversity and robustness. In another study, Hung et al. [33] introduced Med-cDiff, a conditional denoising diffusion probabilistic model (cDDPM) designed for medical image generation tasks such as super-resolution, denoising, and inpainting [33]. The authors demonstrate that diffusion models, which have recently achieved state-of-the-art results in natural image generation, can also outperform traditional generative models such as GANs and VAEs in medical imaging. In their study, Med-cDiff was evaluated across multiple datasets, including MRI and X-ray images, and was shown to produce high-fidelity, realistic synthetic images under various conditional settings. While this study and our study address the challenge of limited access to clinical imaging data and explore synthetic data generation as a solution, they differ in their methodological focus and application domain. Hung et al.[33] leveraged conditional diffusion models for tasks such as MRI super-resolution and X-ray denoising, emphasising the technical superiority of diffusion models over GANs and VAEs in general medical imaging. In contrast, our study evaluated the clinical usability of synthetic panoramic dental X-ray images generated by large multimodal AI models (GPT o4 mini high and Copilot) using textual prompts, with a specific focus on dental diagnostics.

Lastly, a survey study by Shi et al. [34] provides a comprehensive review of the application and development of diffusion models in medical image computing [34]. The survey explains the core principles of diffusion models, including DDPMs, score-based generative models (SGMs), and stochastic differential equations (SDEs). It details the forward process of progressively adding noise to the data and the reverse process of denoising to generate new samples, highlighting the mathematical formulations and neural network training objectives involved. The study categorises the use of diffusion models into five main medical image computing tasks: image generation, modality conversion, image segmentation, image denoising, and anomaly detection, discussing their theoretical foundations and practical applications. The authors also emphasise the clinical significance of each task while reviewing recent advances and challenges in the field. Similar to work employing generative adversarial networks (GANs) and diffusion models for synthetic dental radiograph generation, our approach focuses on producing anatomically accurate and diagnostically relevant images through detailed textual prompts, which are then clinically assessed by experienced dentists for realism and applicability. By fine-tuning a large-scale diffusion model and conducting comparative experiments, Shi et al. demonstrated the promising potential of diffusion models to outperform traditional generative approaches in several medical image computing tasks, particularly in image generation and denoising, while also identifying areas requiring further research, such as segmentation and anomaly detection [34].

While prior studies, such as those using GANs or diffusion models, emphasise architectural innovations and training strategies to enhance image quality and anatomical detail, our study contributes by exploring the usability of large language model-driven image generation tools adopted in dental radiography, thus expanding the range of generative AI methodologies applied in this domain [35, 36]. Moreover, our clinical evaluation highlights the potential of these tools to reduce reliance on real patient data, a challenge also noted in the literature due to ethical and privacy constraints [35, 37].

3. Methodology

3.1. Selected Generative Models

In this study, two commonly used, vision-capable generative models were selected for evaluation on the basis of (i) public accessibility, (ii) documented performance in prompt-based image synthesis, and (iii) the ability to render medical radiographs. Preliminary screening of several alternative multimodal systems revealed either institutional content filters that blocked medical imagery or unacceptably poor output quality; therefore, these candidates were excluded.

- GPT o4 mini high (OpenAI): A high-capacity text-to-image model available through the OpenAI "Plus" subscription tier and widely regarded as a benchmark system in open-domain image generation [38, 39].
- Copilot (Microsoft): A freely accessible multimodal model integrated into the Windows 11 interface that supports both text-to-image and image-to-image generation workflows [40, 41].

To ensure methodological consistency, both models were tested under identical experimental conditions and received the same prompt inputs. Seven prompt categories were defined: General, Specific, Similarity, Blend Generation, Adjustment, Extra, and Detailed, producing a total of sixty-one prompts. Each system completed 27 of these prompts. Copilot could not execute the Blend Generation tasks that require multiple input images and declined the Detailed Prompts due to privacy constraints. As a result, the remaining seven prompts, four from Blend Generation and three from Detailed Prompts, were processed only by GPT o4 mini high. This varied prompt set provided a comprehensive and clinically relevant basis for comparison while preserving uniform testing conditions. Table 1 reports the number of images generated by Copilot and GPT o4 mini high across the seven prompt categories, totaling 61 images.

Table 1. Distribution of Generated Panoramic X-ray Images by Model and Prompt Category

Prompt Category	Copilot	GPT o4 mini high	Total
General Prompt	6	6	12
Specific Prompt	8	8	16
Similarity	6	6	12
Blend Generation	0	4	4
Adjustment	2	2	4
Extra Category	5	5	10
Detailed Prompt	0	3	3
Sum	27	34	61

3.2. Image Generation Prompting Strategy

The primary objective of this stage was to synthetically generate panoramic dental X-ray images using only descriptive textual prompts to evaluate the generative models' capabilities in synthesising panoramic dental X-ray images. The prompts varied in complexity, clinical specificity, and conditioning, covering various realistic scenarios encountered in dental radiography. The prompts were developed in consultation with a specialist dentist and designed by considering common clinical scenarios frequently encountered in dental practice. A list of the main prompt categories is provided below, along with important characteristics and examples:

Task 1: General Prompt (General Text Prompts):

These prompts requested the generation of panoramic dental X-rays without specific clinical conditions or references. The goal was to assess the models' baseline ability to produce anatomically plausible and visually coherent images. Prompt 1 presents a representative example from this category.

Prompt 1:

> "A high-resolution panoramic dental X-ray with clear tooth edges, uniform contrast, and no pathological artefacts."

Task 2: Specific Prompt: (Prompting)

These prompts included detailed clinical descriptions requiring the models to generate images reflecting particular patient characteristics or dental conditions, such as missing deciduous teeth in children or visible bone structures. Prompt 2 presents a representative example from this category.

Prompt 2:

> "Generate a high-resolution panoramic dental X-ray of a 5-year-old child with one or two missing deciduous teeth, showing clear bone structure."

Task 3: Similarity (Image to Image Conditional Generation)

For these prompts, a real panoramic dental X-ray image obtained from the public dataset [42] was provided as a reference. Models were instructed to generate a new synthetic image that closely matched the anatomy, tooth positioning, contrast, and visual style of the reference, without duplicating it. This task evaluated the models' ability to perform conditional image synthesis with high structural fidelity. Prompt 3 presents a representative example from this category.

Prompt 3:

> "Please look at the attached panoramic dental X-ray image. Generate a new, high-resolution panoramic dental X-ray that matches its overall anatomy, tooth position, contrast, and detail level as closely as possible, but is a distinct, newly synthesised image. Keep the same style and visual characteristics (bone texture, tooth shadows, greyscale contrast) without simply duplicating the original. The output should look like a realistic dental X-ray 'twin' of the provided image."

Task 4: Blend Generation:

These prompts, used only with GPT o4 mini high, involved providing two reference images from the dataset [42] and asking the model to generate new images that visually blend the characteristics of both, such as greyscale contrast, tooth positioning, jaw shape, and bone texture. The aim was to test the model's ability to synthesise novel images plausibly belonging to the same imaging dataset. Prompt 4 presents a representative example from this category.

Prompt 4:

> "Please examine the two attached panoramic dental X-ray images. Based solely on their overall visual characteristics—such as greyscale contrast, tooth positioning, jaw shape, and bone texture, generate

new, high-resolution panoramic dental X-rays that could plausibly come from the same imaging dataset."

Task 5: Adjustment (Enhancement)

These prompts instructed the models to produce modified versions of the given images by adding specific clinical features, such as root canal fillings or carious lesions, while keeping all other anatomical details unchanged. This test tested the models' ability to perform targeted, clinically meaningful image augmentations. Prompt 5 presents a representative example from this category.

Prompt 5:

"Please look at the attached panoramic dental X-ray (original.png). Produce a new X-ray that is identical except

Add a filled root canal in the upper right second molar (radiopaque filling in the canal).

Add an occlusal carious lesion in the lower left first molar (radiolucent area on the biting surface).

Keep everything else the same."

Task 6: Extra Category (Rare Conditions Prompting)

These prompts required the generation of images depicting less common clinical scenarios, such as adolescents with fixed orthodontic appliances, to evaluate the models' ability to synthesise complex, detailed features. Prompt 6 presents a representative example from this category.

Prompt 6:

"Generate a panoramic dental X-ray of an adolescent with fixed orthodontic brackets and wires on all teeth, showing bracket placements clearly."

Task 7: Detailed Prompt (Detailed Clinical Prompting)

These specific prompts demanded photorealistic images incorporating multiple pathological features and anatomical details, such as multiple carious lesions, missing teeth, and signs of periodontal disease, while maintaining the natural anatomy and realistic radiographic characteristics. Prompt 7 presents a representative example from this category.

Prompt 7:

"Generate a high-resolution, photorealistic panoramic dental X-ray of an adult patient. Include:

Three distinct carious lesions (e.g., one on a molar, one on a premolar, and one on an incisor), each with a realistic radiolucent appearance.

A missing lower first molar on the right side, showing the empty socket and adjacent bone.

Mild signs of periodontal disease around several posterior teeth (slight bone loss and calculus shadows near the gingival margin).

Keep the overall anatomy natural, show well-defined enamel and dentine layers, visible pulp chambers, realistic curved roots, and subtle bone trabeculation. No implants or restorations. Ensure that the final image looks like a genuine clinical panoramic X-ray."

Figure 2 presents representative dataset samples of each panoramic X-ray alongside its corresponding prompt in English for both models, illustrating how the output varies by category and clinical context.

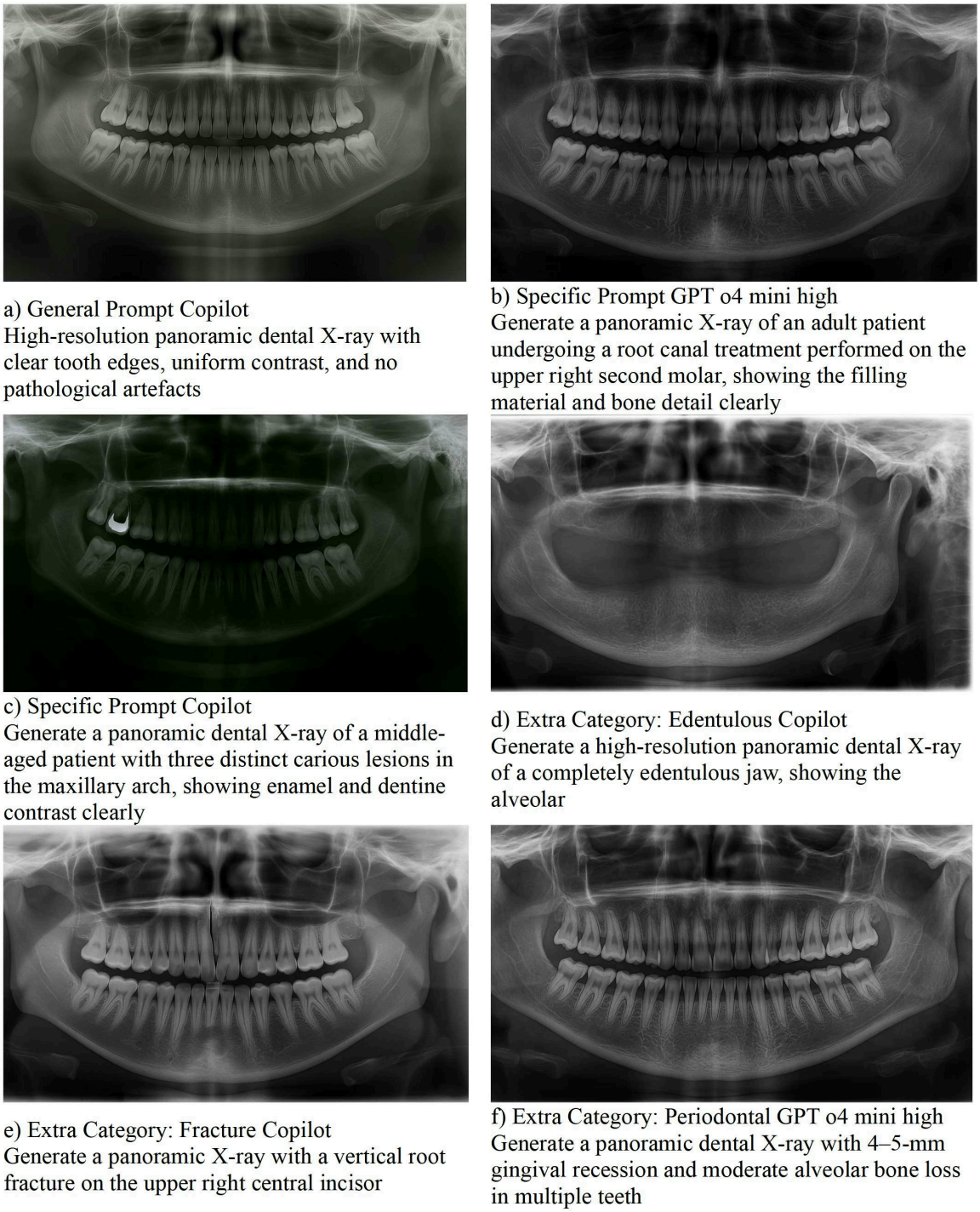


Figure 2. Sample prompts and corresponding panoramic X-ray outputs by model and category

For all scenarios, technically detailed prompts were constructed using anatomical and dental terminology to guide the models in producing clinically relevant synthetic dental X-ray images. By employing a diverse prompting strategy, the models were systematically evaluated across a broad spectrum of clinical

and imaging conditions, including general anatomy, complex pathology, and artefact scenarios. With the combination of unconditional text-based prompts, conditional image-to-image generation, and targeted enhancement requests, the approach enabled an assessment of each model's ability to generate both creative and clinically precise augmentations.

3.3. Evaluation Protocol

A structured two-step protocol was designed to carefully evaluate the synthetic panoramic dental radiographs. The primary analysis relied on a practising dentist who delivered detailed, case-level judgments of each image's realism, anatomical fidelity, clinical usefulness, clarity, and freedom from artefacts. The images were also processed with three commonly used quality indicators, BRISQUE, SSIM, and LPIPS, to provide an additional numerical perspective. These metrics were treated as supportive rather than definitive, serving only to complement and contextualise the expert's findings in the absence of a standardised scoring tool for synthetic dental imagery. Together, the expert-centred review and the auxiliary metrics offered a balanced evaluation of the generation performance of the GPT o4 mini high and Copilot.

3.3.1. Qualitative Expert Evaluations

First, all sixty-one synthetic panoramic images produced by the two models were evaluated by the expert dentist. Each image received an independent rating on five criteria: visual realism, anatomical accuracy, clinical and educational utility, image clarity, and artefact-free using a scale from "very dissatisfied" to "very satisfied." Furthermore, two representative image pairs from the similarity task were examined in depth to evaluate how well each model reproduced the source radiographs. Additionally, during the adjustment task, the evaluator compared every synthetic image with its corresponding reference to assess the loyalty of the inserted pathologies and the effectiveness of the enhancements.

Visual Realism: Scoring focused on how convincingly the image depicts dental structures as they appear in actual panoramic radiographs. Examiners assessed the shape of the crowns and roots, the contour of the pulp chamber, the texture of the hard tissues, and the number and positions of the teeth. For restorative or surgical elements, such as fillings, root-canal obturations, or implants, realism was judged by the accuracy of their location, shape, size, colour, and surface texture.

Anatomical Accuracy (Pathological Validity): This criterion assessed whether the key maxillofacial structures were depicted correctly. Raters inspected the form, size, and borders of the maxilla and mandible; the crown-and-root morphology of individual teeth; and the fidelity of the surrounding anatomical landmarks, including the maxillary sinuses, mandibular canal, and nasal cavities.

Clinical/Educational Utility: Here, the evaluators considered how well each generated image could serve as a teaching aid or clinical reference. Scores reflected both the overall resemblance to real radiographs and the degree to which anatomical details were rendered accurately enough for instruction or diagnostic simulation.

Image Clarity: Clarity was rated according to the distinguishability of fine dental and skeletal details: crown-root interfaces, enamel–dentine transitions, trabecular bone pattern, and the sharpness of structures such as the maxillary sinus, mandibular canal, and fossae. The easier it was to recognise these features, the higher the score.

Artefact Detection: This dimension captured the absence of image distortions that might hinder interpretation. Evaluators looked for blur, indeterminate regions, overlapping artefacts, or any other defects that could compromise the diagnostic quality, deducting points when such issues were present.

Qualitative judgments were then translated into quantitative data by assigning each expert rating to a five-point Likert scale, with 1 signifying "very dissatisfied" and 5 signifying "very satisfied." These standardised scores establish an expert-derived measure of image quality and provide the foundation for the next section, which adds metric-based quantitative analysis, statistical testing, and further subjective evaluation findings.

3.3.2. Quantitative Analysis:

In addition to the expert's qualitative appraisal, an objective image-quality analysis was carried out using well-recognised similarity metrics. This quantitative layer provided supportive numerical evidence rather than definitive proof of the image fidelity.

- Blind or Referenceless Image Spatial Quality Evaluator (BRISQUE): BRISQUE extracts point-wise statistics from locally normalised luminance signals and gauges deviations from natural-image statistics [43]. Lower values denote better perceptual quality. Because it does not require a reference image, BRISQUE was applied to the full set of synthetic panoramic images.
- Structural Similarity Index Measure (SSIM): SSIM quantifies likeness in luminance, contrast, and structural detail between two images, yielding a higher score for closer correspondence [44]. It was used in the image-conditioned generation task to compare each synthetic radiograph with its real counterpart.
- Learned Perceptual Image Patch Similarity (LPIPS): LPIPS employs deep neural features to measure perceptual similarity [45], with lower scores indicating superior quality. Similar to SSIM, it was calculated for the image-conditioned task.

By coupling the dentist's primary, clinically grounded evaluation with these supportive metrics, the study delivers a balanced comparison of GPT o4 mini high and Copilot, even in the absence of specialised dental-imaging quality standards. BRISQUE, SSIM, and LPIPS were employed as supportive quantitative metrics to underpin our findings. Despite the absence of standardised tools for assessing synthetic image quality in dentistry, these metrics provide valuable supplementary information.

4. Evaluation And Results

Following image generation, both dentist evaluations and quantitative assessments were conducted to provide a comprehensive, objective analysis of the results. The subsequent analytical procedures were as follows:

- The dentists' ratings were subjected to descriptive, statistical, and correlation analyses to identify overall trends and assess significance.
- BRISQUE scores were computed for all generated images to obtain no-reference quality assessments.
- LPIPS and SSIM values were calculated for pairwise comparisons of both models' outputs.
- Correlation analyses between pairwise quantitative metrics and expert evaluations were performed to determine the degree of agreement.
- Example pairwise comparisons of outputs were conducted by dentists to qualitatively interpret each model's strengths and weaknesses.
- Both models' outputs for enhancement and adjustment tasks were evaluated by dentists against clinical expectations to assess performance for given specific tasks.

In the following sections, we first examine dentists' ratings as our primary quality measure, then present the objective metric results, perform correlation analyses between those ratings and the quantitative outcomes to validate the expert evaluations, and finally explore dentists' interpretations of the image adjustment category to assess its effectiveness.

4.1. Dentist Evaluation Results

As our main evaluation method, the dentists' ratings were initially examined using descriptive statistics to reveal underlying data patterns and improve interpretability. Subsequently, statistical tests were applied to assess significance, followed by correlation analyses across the five evaluation criteria to examine their interrelationships.

4.1.1. Descriptive Analysis of the Results

An initial descriptive analysis, encompassing both category-based and model-based approaches, was performed for each criterion. Table 2 summarises the descriptive statistics of the dentist's ratings (converted to a 5-point scale), reporting the mean, median, interquartile range (IQR), standard deviation results for each evaluation criterion and the overall score.

Table 2. Descriptive Statistics of Dentist Ratings

	Visual Realism	Anatomical Accuracy	Clinical/educational Utility	Image Clarity	Artefact-Free	Overall
Median	4,00	4,00	4,00	4,00	4,00	4,00
IQR	2,00	1,00	2,00	1,00	0,00	2,00

As shown in Table 2, all criteria achieved a median rating of 4.00, indicating consistently positive evaluations, and the overall median of 4.00 (IQR = 2.00) reflects that, even prior to conversion into a 5-point Likert scale, the dentist's raw evaluations were predominantly positive. Although the IQR for Artefact-Free is 0, reflecting unanimous agreement, larger IQRs for Visual Realism and Clinical/Educational Utility (2.00) reveal more dispersed opinions, yet all IQRs remain within acceptable bounds. Collectively, these statistics underscore that the generated images met the dentist's expectations and were deemed above average in quality.

Table 3 reports the median and interquartile range (IQR) of the dentist's ratings for each prompt category across all five evaluation criteria, along with the number of images (N) assessed in each category.

Table 3. Median ± IQR Of The Dentist's 5-Point Likert Ratings by Prompt category

	N	Visual Realism	Anatomical Accuracy	Clinical/educational Utility	Image Clarity	Artefact-Free	Overall
General text	12	5±1.75	4±0.75	5±1.75	5±1	4±0.75	4±1
Specific clinical	16	3±2	4±1	3±2	4±0.75	4±1	4±1
Similarity	12	4±0.75	4±0	4±1	4±0	4±0	4±0
Image-to-image Blend	4	4±0.75	4±0	4±0.75	4±0	4.5±1	4±0
Refinement prompts	4	3.5±1.75	3±0.75	3±1.5	4±0.75	3.5±1	3±1
Rare-case scenarios	10	3.5±1.25	4±1.25	3.5±2	4.5±2	4.5±1	4±2
Detailed Prompt	3	4±2	4±0	4±2	4±1	4±1	4±1

Table 3 shows that general text prompts achieved the highest median Visual Realism (5 ± 1.75) and Clinical/Educational Utility (5 ± 1.75) scores but also showed the greatest variability, while specific clinical scenarios

exhibited lower realism (3 ± 2) yet more consistent anatomical accuracy (4 ± 1). Notably, both image-to-image similarity (N = 12) and blend (N = 4) categories yielded medians of 4 with narrow IQRs (≤ 0.75), reflecting stable performance, whereas rare-case and refinement prompts produced medians around 3.5–4 with wider IQRs, indicating more diverse expert opinions in these challenging contexts.

Table 4 summarises the median and interquartile range for the five quality criteria as well as the overall score, separately for GPT o4 mini high and Copilot.

Table 4. Median ± IQR Of The Dentist's 5-Point Likert Ratings by Model

Model/Criteria	N	Visual Realism	Anatomical Accuracy	Clinical/educational Utility	Image Clarity	Artefact-Free	Overall
GPT o4 mini high	34	4±1	4±0	4.5±1	4±1	4±1	4±1
Copilot	27	3±2	3±1	3±2	4±1	4±1	4±1

GPT o4 mini high outperformed Copilot in Visual Realism (4 ± 1 vs. 3 ± 2), Anatomical Accuracy (4 ± 0 vs. 3 ± 1), and Clinical/Educational Utility (4.5 ± 1 vs. 3 ± 2), indicating superior fidelity and clinical usefulness. Both models achieved identical median scores for the Image Clarity and Artefact-Free criteria (4 ± 1), demonstrating that Copilot can produce comparably clear and artefact-free images. Overall median ratings were equivalent (4 ± 1), but the higher scores in key diagnostic dimensions for GPT o4 mini high underscore its advantage in generating more realistic and anatomically accurate X-ray images.

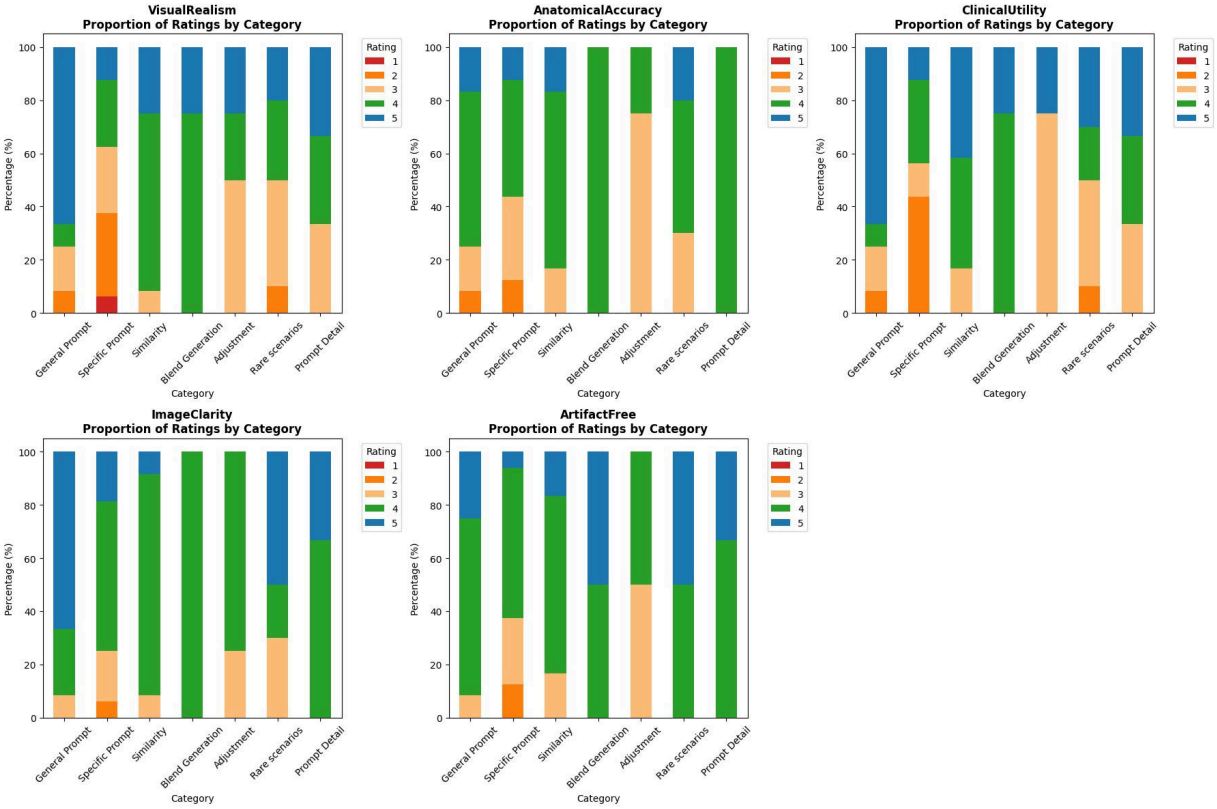


Figure 3. Percentage distribution of dentist ratings (1 = lowest to 5 = highest) across seven prompt categories for each of the five evaluation metrics: (a) Visual Realism, (b) Anatomical Accuracy, (c) Clinical Utility, (d) Image Clarity, and (e) Artifact-Free Quality.

A visual representation of the proportional breakdown of ratings by prompt category is presented in Figure 3.

4.1.2. Statistical Analysis

After descriptive analysis, the internal consistency of the dentist's ratings across the five image-quality criteria was assessed using Cronbach's α, yielding α = 0.894. An α of 0.894 indicates high internal consistency, which suggests that these five criteria reliably hang together as a single "image-quality" scale.

The Shapiro–Wilk test was then applied to evaluate normality within each prompt category, revealing that several groups departed from a normal distribution and therefore required nonparametric methods [46]. Accordingly, a Kruskal–Wallis H-test was conducted to determine whether significant differences existed among the seven prompt categories [47] the results of this analysis are presented in Table 5 with the H statistic, effect size (η^2), and p-values for each evaluation criterion.

Table 5. Kruskal–Wallis H-Test Results for Differences Across Seven Prompt Categories on Five Image-Quality Criteria.

Criterion	H statistic	η^2	p-value
Visual Realism	12.748	0.125	0.047
Anatomical Accuracy	6.130	0.002	0.409
Clinical/Educational Utility	11.944	0.110	0.063
Image Clarity	10.027	0.074	0.124
Artefact-Free	15.926	0.183	0.014

As shown in Table 5, Visual Realism exhibited a borderline significant difference across prompt categories (H = 12.748, η^2 = 0.125, p = 0.047), and Artefact-Free scores differed significantly (H = 15.926, η^2 = 0.183, p = 0.014), whereas the remaining criteria did not reach conventional significance (all p > 0.05).

To identify specific pairwise differences among the prompt categories, Dunn's post hoc test with Bonferroni correction was performed [48]. The results indicated that General-Prompt images were rated significantly more realistic than Specific-Prompt images (Bonferroni-corrected p = 0.035, r_n = 0.58), and Extra-Category scenarios were rated significantly more artefact-free than Specific-Prompt images (Bonferroni-corrected p = 0.031, r_n = 0.63). No other pairwise comparisons reached statistical significance.

Models' Performance Considerations

To control for model-level differences, a Wilcoxon signed-rank test was conducted on each dimension [49] median differences, p-values, and effect sizes are reported in Table 6.

Table 6. Model based Wilcoxon signed-rank test results.

Criterion	GPT o4 mini high Median	Copilot Median	Median Difference	p-value	Effect Size
Visual Realism	4.00	3.00	1.00	0.0002	−0.514
Anatomical Accuracy	4.00	3.00	1.00	0.0002	−0.480
Clinical/ Educational Utility	4.50	3.00	1.50	0.0002	−0.516
Image Clarity	4.00	4.00	0.00	0.0032	−0.370
Artefact-Free	4.00	4.00	0.00	0.0000	−0.542

GPT o4 mini high outperformed Copilot on Visual Realism (median difference = 1.00, p = 0.0002, r = 0.51), Anatomical Accuracy (1.00, p = 0.0002, r = 0.48), and Clinical Utility (1.50, p = 0.0002, r = 0.52), corresponding to medium-to-large effects. Although the medians for Image Clarity and Artefact-Free were identical (0.00),

both comparisons reached significance ($p = 0.0032$, $r = 0.37$; $p < 0.0001$, $r = 0.54$), indicating that Copilot produced slightly clearer and more artefact-free outputs. Overall, GPT o4 mini high led on three of the five criteria, while Copilot showed modest advantages on the remaining two.

4.1.3. Correlation Results

Finally, Spearman's rank correlations were computed to assess the interrelationships among the five evaluation criteria across all 61 synthetic images [50]. The resulting correlation matrix is presented in Figure 4.

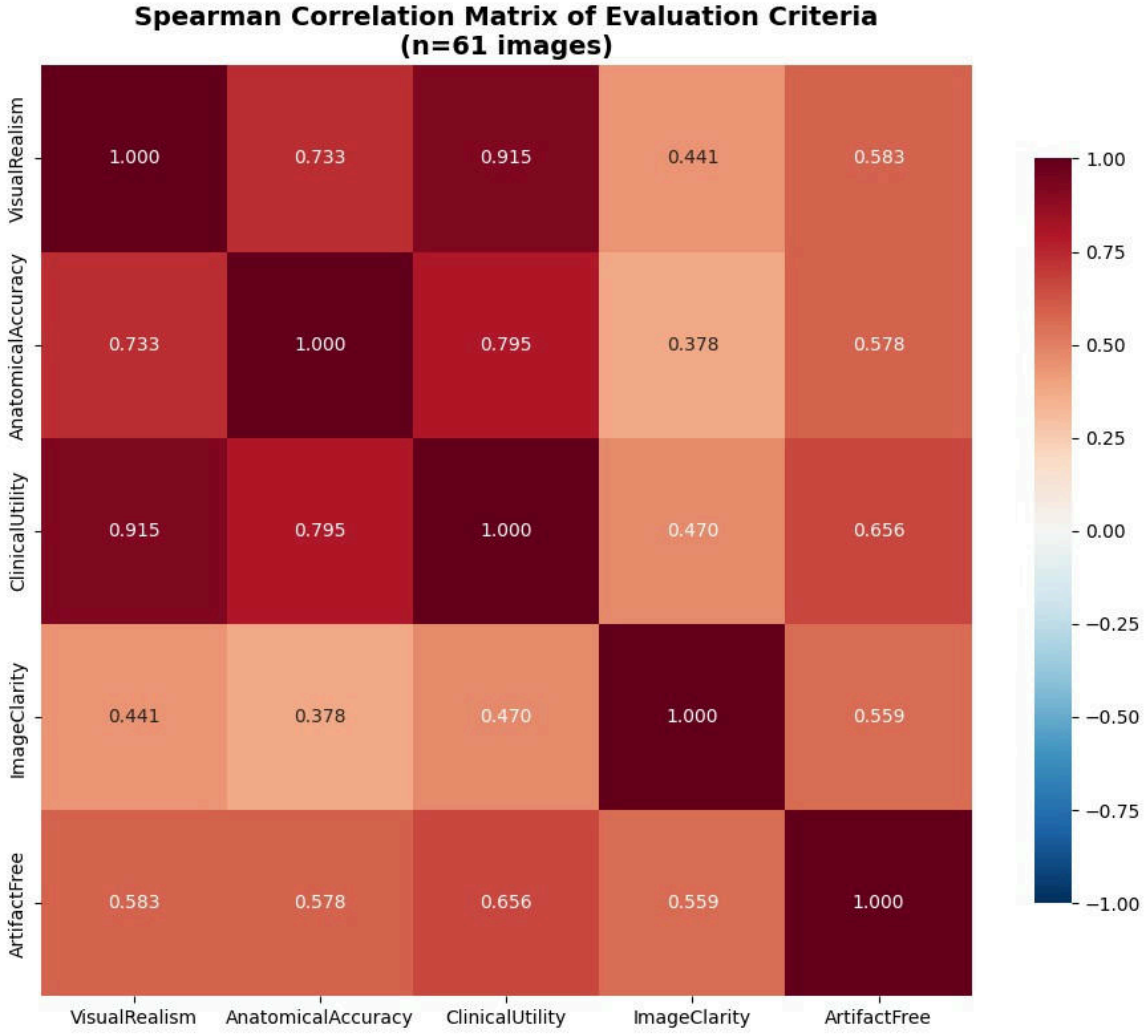


Figure 4. Spearman correlation matrix of the five evaluation criteria across 61 synthetic panoramic dental X-ray images.

Visual Realism correlates very strongly with Clinical Utility ($\rho = 0.915$) and strongly with Anatomical Accuracy ($\rho = 0.733$), suggesting that perceptions of realism are closely tied to both usefulness and anatomical fidelity. Image Clarity showed moderate correlations with the other criteria ($\rho = 0.378$–0.470), indicating that the clarity judgments were relatively independent. Artefact-free quality correlates moderately to strongly with all metrics ($\rho = 0.559$–0.656), implying that reduced artefacts contribute significantly to overall image quality, although they do not entirely determine other attributes.

4.2. Objective Metric Evaluation

Following the analysis of the dentist's ratings, we applied objective image-quality metrics to further quantify and validate our findings. Three recognised evaluation methods were employed; BRISQUE was calculated for all 61 generated images, as it provides no-reference quality scores. LPIPS and SSIM were computed for the 12 generated and original image pairs to assess perceptual similarity and structural fidelity. Furthermore, the dental expert side-by-side comparisons of two random samples generated an image with its corresponding original to provide qualitative insight. In the absence of standardised synthetic image quality metrics for dental applications, these no-reference and pairwise comparison techniques function as supportive indicators rather than as definitive measures of image quality. Since BRISQUE, SSIM, and LPIPS were originally developed and trained for natural scene images. Consequently, these metrics do not adequately capture the diagnostic relevance, anatomical accuracy, or subtle pathological features required for medical imaging, highlighting the need for an expert clinical review.

4.2.1. No Reference Image Evaluation

BRISQUE was applied to all generated images to obtain no-reference quality scores for each model across the seven prompt categories. Table 7 presents the average BRISQUE scores for each model and the prompt category.

Table 7. Mean BRISQUE Scores by Prompt Category and Model

Category	Model	Mean BRISQUE
General Prompt	Copilot	11.37
	GPT o4 mini high	2.80
Specific Prompt	Copilot	17.51
	GPT o4 mini high	−0.25
Similarity	Copilot	22.59
	GPT o4 mini high	0,20
Blend Generation	GPT o4 mini high	3.08
Adjustment	Copilot	7.79
	GPT o4 mini high	13.14
Extra-Category (Rare Cases)	Copilot	12.16
	GPT o4 mini high	2.41
Detailed Prompt	GPT o4 mini high	23.83

GPT o4 mini high consistently achieved substantially lower (better) BRISQUE scores than Copilot in general, specific prompts, and similarity (2.80 vs. 11.37; 0.25 vs. 17.51; 0.25 vs. 17.51), as well as in rare-case scenarios, indicating superior locational naturalness. Conversely, Copilot slightly outperformed GPT o4 mini high on the adjustment tasks (7.79 vs. 13.14), and the GPT o4 mini high model struggled most under highly detailed prompts (Prompt Detail, GPT = 23.83), suggesting that extreme prompt specificity may degrade the perceived image quality. The slightly negative BRISQUE value for GPT o4 mini high in the Specific Prompt category reflects the fact that BRISQUE, being trained on natural scenes, can yield outlier or negative scores when applied to out-of-domain inputs such as greyscale medical X-rays. Because BRISQUE is not optimised for dental imaging, these results serve as supplemental indicators alongside the dentist's evaluations rather than definitive measures of image quality.

4.2.2. Paired Image Evaluation

In the image-to-image prompt category, where each output is generated from a specific reference image, reference-based metrics were applied to quantify fidelity. Specifically, the SSIM and LPIPS were calculated between each synthetic image and its corresponding ground truth. Table 8 summarises these results for the six image pairs generated by each model.

Table 8. Mean ± SD SSIM and LPIPS Scores for Copilot And GPT O4 Mini High

Model	Count	SSIM (Mean ± Std)	LPIPS (Mean ± Std)
Copilot	6	0.458 ± 0.027	0.454 ± 0.071
GPT o4 mini high	6	0.442 ± 0.029	0.401 ± 0.056

While SSIM scores were slightly higher for Copilot (0.458 vs. 0.442), indicating marginally better structural similarity to the reference images, GPT o4 mini high outperformed Copilot in LPIPS (0.401 vs. 0.454), suggesting superior perceptual and signal-level fidelity. Both Copilot and GPT o4 mini high attained only moderate fidelity to their reference images (SSIM = 0.44–0.46; LPIPS = 0.40–0.45), highlighting that image-conditioned generation yields visually plausible but structurally different outputs.

4.2.3. Expert Side-by-Side Comparison

To further investigate the limitations of image-conditioned generation, a dental expert performed side-by-side evaluations of a randomly selected subset of synthetic–ground-truth image pairs. Each pair was examined for clinically relevant discrepancies in anatomical detail, image clarity, and artefact presence. Two representative examples, both generated by the GPT o4 mini high, are presented in Figures 5 and 6. In each case, a synthetic image is presented alongside its ground-truth counterpart, along with an absolute-difference map that shows the regions of difference.

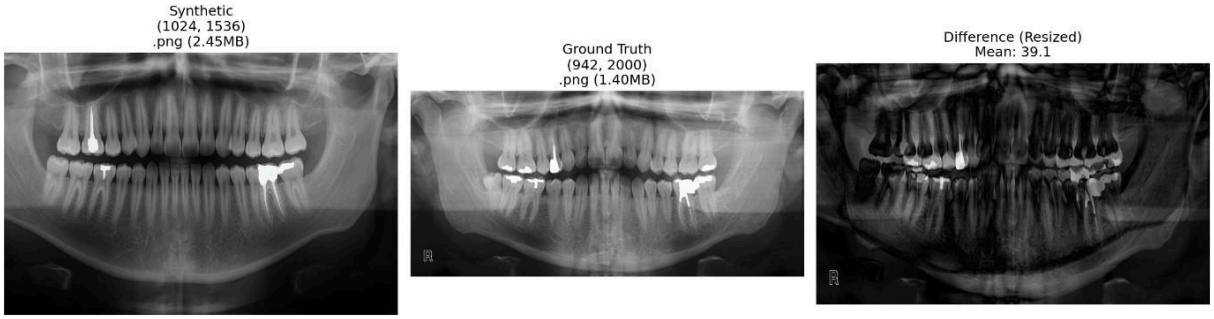


Figure 5. GPT o4 Mini High Output (Left), Ground-Truth Panoramic X-Ray (Centre), and Absolute Difference Map

The restorations on teeth 46 to 37 were imitated in terms of location and shape with close resemblance to the original. However, the restorations on teeth 27 to 47 are absent in the generated image. The root canal fillings on teeth 36 to 14 were incorrectly placed in terms of both location and shape. Overall, the images exhibit marked discrepancies in appearance that are noticeable even to non-expert observers.

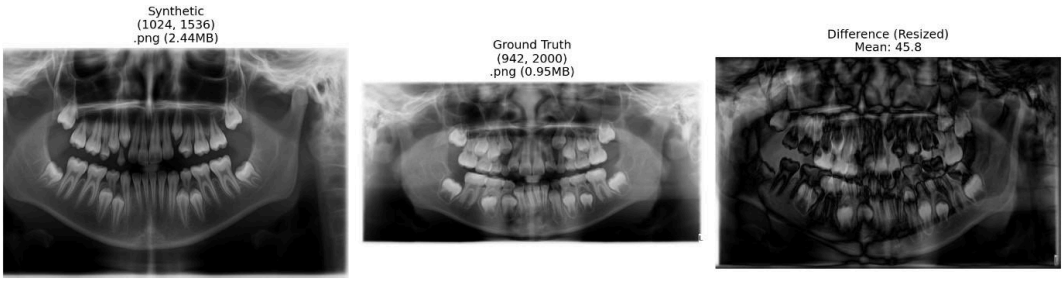


Figure 6. GPT o4 Mini High Output (Left), Ground-Truth Panoramic X-Ray (Centre), and Absolute Difference Map

An image representing the mixed dentition period was attempted to be replicated; however, despite the correct positioning of some impacted teeth, discrepancies in both the number and morphology of the teeth were observed. Overall, the synthetic and ground-truth images showed clear anatomical and quantitative differences, particularly in the representation of the fine structures and the contrast distribution. These inconsistencies were noticeable even to non-expert observers and highlight the current limitations of image-conditioned generation in producing clinically reliable dental panoramic images.

4.3. Objective Metric Evaluation Correlation with Dentist Evaluation

The expert's clinical assessments for each generative model were quantitatively validated against the corresponding objective pairwise-comparison metrics of SSIM and LPIPS. Pearson correlation coefficients were computed to quantify the linear associations between the subjective evaluations (e.g., diagnostic utility, artefact severity) and the SSIM/LPIPS scores. Table 9 reports the Spearman correlation coefficients (ρ) and associated *p*-values linking two objective image-quality metrics (SSIM and LPIPS) to the dentist's four rating criteria for Copilot-generated panoramic images; only associations that achieved, or closely approached, statistical significance are reported. Statistical significance was assessed at α = 0.05.

Table 9. Spearman Correlation Coefficients (P) Between SSIM/LPIPS Scores and Expert Evaluation Criteria for Copilot-Generated Panoramic Images.

Metric	Rating criterion	ρ	p
SSIM	Artefact Free	0.83	0.04
LPIPS		−0.83	0.04
SSIM	Image Clarity	0.68	0.14
LPIPS		−0. 68	0.14
SSIM	Clinical/Educational Utility	0.51	0.3
LPIPS		−0.51	0.3
SSIM	Visual Realism	0.51	0.3
LPIPS		−0. 51	0.3

For Copilot, both metrics correlate very strongly with the "artefact-free" criterion ($|\rho| = 0.83$, $p = 0.04$). The opposite signs for SSIM and LPIPS reflect their scoring conventions (higher SSIM versus lower LPIPS indicating better quality). Correlations with image clarity were moderate to strong ($|\rho| \approx 0.68$) but failed to reach statistical significance, likely reflecting the limited sample size. Associations with clinical utility and visual realism were only moderate ($|\rho| \approx 0.51$) and likewise non-significant, indicating that these similarity-based metrics are less sensitive to broader diagnostic attributes.

Table 10 summarises the Spearman correlation coefficients between SSIM/LPIPS scores and the dentist's ratings for GPT o4 mini high outputs, limiting the presentation to those associations that attained, or narrowly missed, statistical significance.

Table 10. Spearman Correlation Coefficients (P) Between SSIM/LPIPS Scores and Expert Evaluation Criteria for GPT o4 Mini High- Generated Panoramic Images.

Metric	Rating criterion	ρ	p
LPIPS	Clinical/Educational Utility	0.85	0.03
LPIPS	Anatomical Accuracy	0.72	0.11

The image-patch similarity metric LPIPS showed a very strong and statistically significant correlation with the dentist's assessment of clinical utility (ρ = 0.85, p = 0.03), indicating that higher similarity to ground-truth images aligns closely with greater diagnostic usefulness in GPT-generated outputs. The association with anatomical accuracy was moderately strong (ρ = 0.72) but not significant at the current sample size (p = 0.11), suggesting a promising yet inconclusive relationship. Because only these two correlations met the predefined reporting threshold, further data and complementary metrics would be necessary to establish a more comprehensive quantitative validation of the GPT-based image-conditioned generation.

Overall, the quantitative results provide partial support for the expert's ratings in detecting artefacts; however, a larger dataset and additional metrics that extend beyond pairwise similarity are required to comprehensively evaluate the quality of the outputs.

4.4. Image Adjustment Comparison

Finally, the "adjustment" category was investigated, in which the models were tasked either with inserting predefined clinical features (e.g., carious lesions, root-canal fillings) or with enhancing blurred images while preserving their anatomical integrity. For this analysis, each original image was displayed alongside its corresponding synthetically adjusted version to facilitate direct visual comparison. The dental expert then conducted a detailed evaluation of every pair, considering both the fidelity of the introduced pathologies and the effectiveness of the enhancements in terms of clarity, contrast, and preservation of fine anatomical structures. These assessments collectively provide a nuanced appraisal of the models' performance in the adjustment category.

4.4.1. Image Adjustment Evaluation

In the first stage, each model was instructed to modify the supplied image by inserting a predefined clinical feature while preserving all other anatomical and radiographic characteristics. This experiment evaluated the models' ability to generate targeted, clinically relevant variations from an existing image, a capability that could prove invaluable when domain-specific data are scarce. The exact prompt used for this task is provided below in Prompt 8.

Prompt 8:

"Please look at the attached **panoramic** dental X-ray (original.png). Produce a new X-ray that is identical except

- Add a filled root canal in the upper right second molar (radiopaque filling in the canal).

- Add an occlusal **carious** lesion in the lower left first molar (radiolucent area on the biting surface).

Keep everything else the same."

Figure 7 presents the ground-truth image (left) with its corresponding Copilot adjusted counterpart (right).

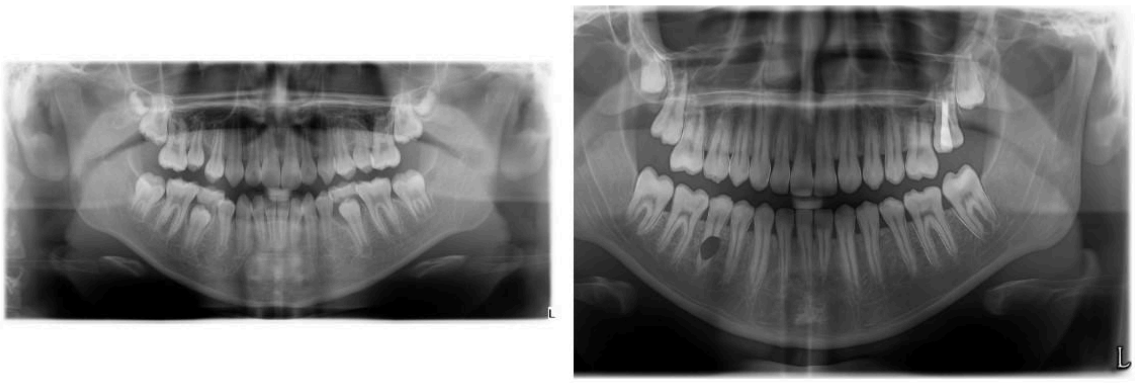


Figure 7. Side-By-Side Comparison Of The Reference Panoramic Radiograph (Left) And The Copilot-Generated, Clinically Adjusted Image (Right).

Dentist evaluation of Figure 7: In the real image, primary teeth are present in the mandible, whereas these teeth are absent in the generated image. The tooth structures and morphologies differ between the two. Although the positions of the impacted teeth in the maxilla are similar, their shapes and structures remain different, resulting in distinct images. Only the forms of the upper central incisors show some resemblance. There are also errors in the commands given for the replicated image.

The canal filling intended to be generated in the upper left second molar is accurately positioned, and the shape of the canal filling can be considered close to the real image. However, the occlusal carious lesion intended to be placed on the lower left second molar is located incorrectly, with an inaccurate position and a form that does not correspond to reality.

Figure 8 presents the ground-truth image (left) with its corresponding GPT o4 mini high-adjusted counterpart (right).

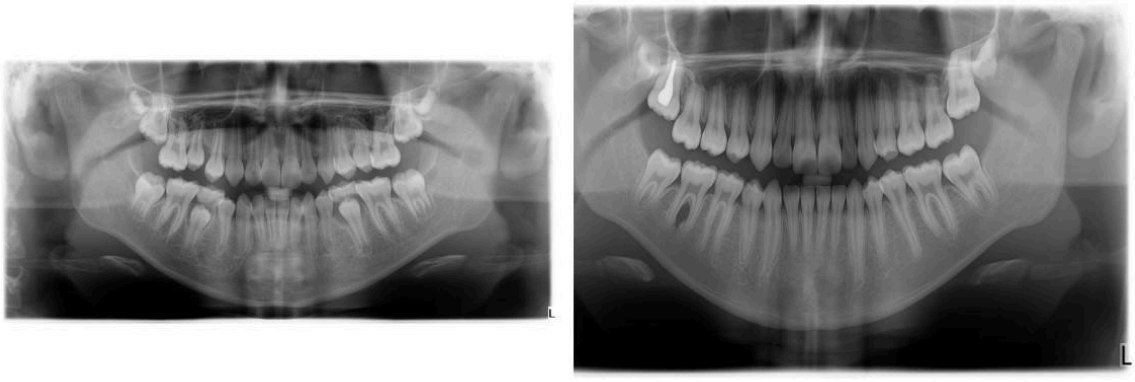


Figure 8. Side-By-Side Comparison of The Reference Panoramic Radiograph (Left) And The GPT o4 Mini High Generated, Clinically Adjusted Image (Right).

Dentist evaluation of Figure 8: In the real image, primary teeth are present in the mandible, whereas these teeth are absent in the generated image. The tooth structures and morphologies differ between the two.

Although the positions of the impacted teeth in the maxilla appear similar, their structures and morphologies remain different, resulting in distinct images. The commands given for the replicated image were also not successfully executed.

The canal filling intended to be generated in the upper left second molar is incorrectly positioned and appears instead on the upper right second molar. However, the shape of the canal filling can be considered close to the real image. The occlusal carious lesion intended to be placed on the lower left second molar is located incorrectly, with an inaccurate position and a form that does not correspond to reality.

When applying these commands, special attention must be paid to the concepts of right and left. According to the panoramic X-ray image displayed on the screen, the right-left orientation does not necessarily correspond to the patient's actual right and left sides. It is essential to provide a clear command indicating whether the right side of the image corresponds to the patient's right side or not, as the AI may misinterpret and hallucinate. To reduce these anatomical hallucinations, we recommend using spatially explicit prompts, such as visual orientation markers in the input images, and training with directionally annotated datasets to improve the model's anatomical consistency.

4.4.2. Image Enhancement Evaluation

In the second stage of this category, each model was supplied with a blurred or otherwise degraded panoramic radiograph and instructed to generate an enhanced version that preserved all underlying anatomical and radiographic features. This evaluation probes the models' ability to recover diagnostically usable images from sub-optimal inputs, potentially reducing the need for conventional preprocessing steps when deep-learning workflows encounter acquisition artefacts that standard image-processing methods cannot fully correct. The exact wording of the enhancement prompt is provided in Prompt 9.

Prompt 9:

> Please examine the attached panoramic dental X-ray image. Generate a new, enhanced version optimised for deep learning analysis by performing the following adjustments:
>
> - Increase overall brightness and uniformize contrast so that tooth structures (enamel, dentine, roots) are clearly visible without overexposure.
>
> - Reduce noise and blur to sharpen the edges of all teeth, roots, and alveolar bone.
>
> - Preserve all anatomical features exactly (tooth positions, bone contours, soft tissue shadows), only improving clarity.
>
> - The output should appear like a clean, high-quality radiograph suitable for automated segmentation or classification algorithms.
>
> Generate a new image based on the given prompt.

To illustrate the enhancement task, Figure 9 contrasts a severely blurred panoramic radiograph with the corresponding Copilot generated reconstruction.

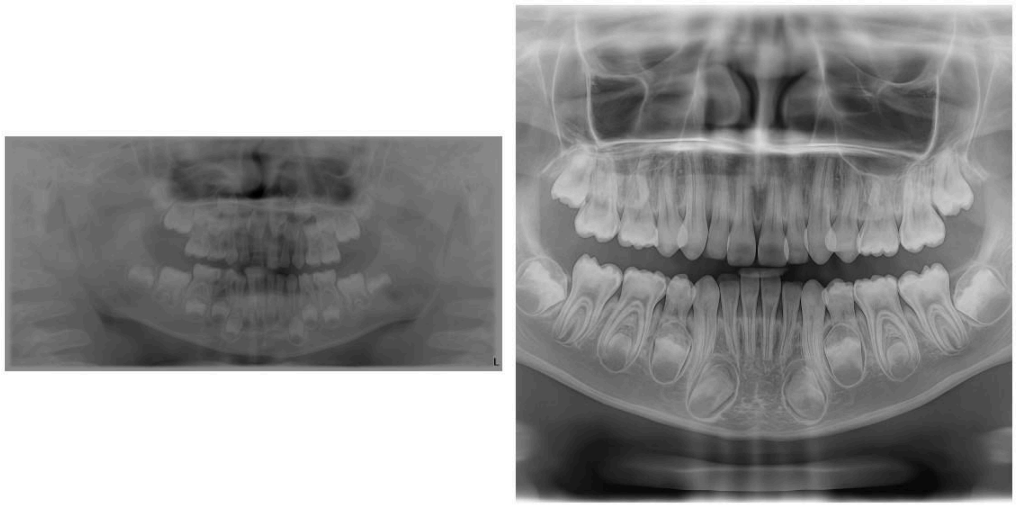


Figure 9. Comparison of the degraded input radiograph (left) and the Copilot-generated enhancement (right) produced in response to Prompt 9.

Dentist evaluation of Figure 9: A blurry radiographic image was intended to be enhanced and replicated while preserving all teeth and anatomical structures. In the generated image, the teeth and anatomical features appear clearer. However, differences exist in tooth morphology. In the mandible, both the unerupted and erupted teeth match the original image in number and position although their shapes and structures differ. In the maxilla, a large portion of the unerupted teeth are missing in the generated image. The number and morphology of the maxillary teeth in the generated image differ from those in the original.

To illustrate the enhancement task, Figure 10 contrasts a severely blurred panoramic radiograph with the corresponding GPT o4 mini high-generated reconstruction.

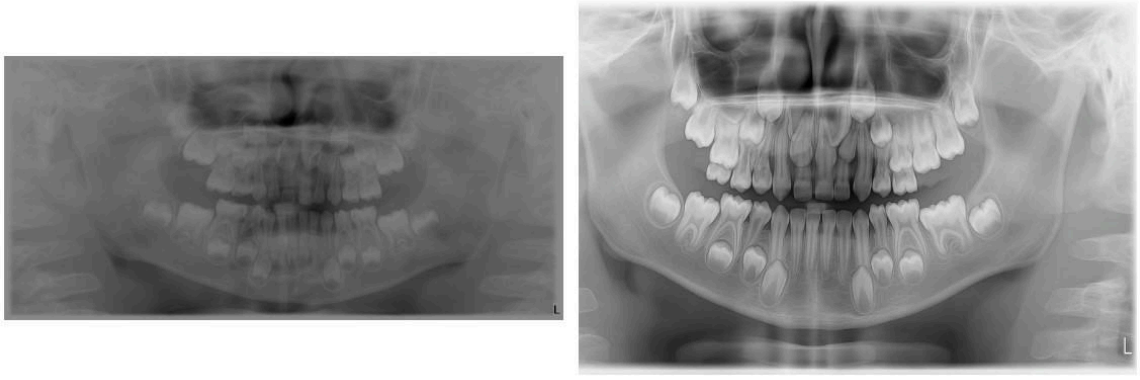


Figure 10. Comparison of the degraded input radiograph (left) and the Copilot-generated enhancement (right) produced in response to Prompt 9.

Dentist evaluation of Figure 10: A blurry radiographic image was replicated while preserving all anatomical structures. The resulting image appeared clearer. The number of erupted and unerupted teeth in both the mandible and maxilla matched the original image, and their positions were similar. However, the tooth morphology and the structures differed from the original. Despite some errors, the radiographic image replicating the mixed dentition generated by GPT was found to be closer to reality compared to Copilot's replication.

5. Discussion

This study investigated how two multimodal generative systems, GPT o4 mini high and Copilot, can synthesise panoramic dental radiographs that satisfy clinical expectations while diminishing data scarcity. An expert dentist examined sixty-one images across five quality dimensions: visual realism, anatomical accuracy, clinical/educational utility, image clarity, and freedom from artefacts.

5.1. Summary of the Key Findings

This study investigated the capacity of two multimodal generative systems, GPT o4 mini high and Copilot, to synthesise clinically viable panoramic dental radiographs that address data scarcity challenges. An expert dentist evaluated 61 images across five dimensions: visual realism, anatomical accuracy, clinical/educational utility, image clarity, and artefact absence. The findings revealed consistently above-average quality across all dimensions, with strong internal consistency confirming these criteria as unified quality indicators.

Regarding RQ1, both GPT-4o mini high and Copilot produced panoramic images judged by the expert dentist to be above average in clinical quality. The GPT o4 mini high demonstrated superior performance in visual realism, anatomical fidelity, and overall diagnostic utility, with median expert ratings frequently matching or exceeding those for authentic radiographs. Copilot delivered comparably clear and artefact-free outputs, making it well-suited for scenarios requiring minimal image cleanup. Overall, both models can generate clinically usable synthetic dental radiographs, with GPT o4 mini high favouring diagnostic authenticity and Copilot favoring cleanliness.

Prompt design significantly influenced the outcomes, addressing RQ3. General language prompts yielded the most realistic images, while rare-condition scenarios produced the cleanest outputs. Conversely, overly detailed prompts sometimes degraded the quality, indicating limitations in handling excessive complexity. Adjustment and enhancement prompts succeeded in inserting pathology indicators and restoring blurred images, although with imperfect anatomical accuracy. These findings underscore that carefully calibrated prompt design, balancing specificity against the model's generative limits, is critical for maximising both the diversity and clinical validity of synthetic dental images, especially in complex or rare scenarios.

In relation to RQ2, image-conditioned and blend-generation tasks yielded plausible reconstructions of the source radiographs, but neither model reproduced the originals exactly. Quantitative metrics (SSIM ≈ 0.44–0.46; LPIPS ≈ 0.40–0.45) and expert similarity ratings confirmed that image-conditioned synthesis introduces meaningful structural variation. Blend prompts produced stable outputs with narrow variability, suggesting that blending can reliably augment datasets with realistic anatomical mixtures. Although these methods offer medically valid and quantitatively consistent data to mitigate imaging scarcity, their output must be examined by domain experts before downstream use.

Correlational analysis showed that visual realism was strongly linked to clinical utility (ρ=0.915) and anatomical accuracy (ρ=0.733), while image clarity contributed moderately. Artefact scores correlated robustly with all dimensions, indicating that noise reduction supports but does not solely determine quality. Quantitative metrics (BRISQUE, SSIM, LPIPS) reinforced the expert assessments without replacing them, although both models generated structurally divergent images from the references, highlighting the inherent limitations of image-conditioned synthesis.

Adjustment tasks (e.g., adding pathologies) and enhancements (e.g., deblurring) produced visibly altered outputs despite instructions to preserve the originals. GPT o4 mini high more accurately replicated the

mixed dentition, while both systems improved the clarity and positional accuracy despite the morphological variances.

Practical considerations included the necessity of explicit specification of laterality and image dimensions in prompts, as right–left orientation in panoramic views can invert patient anatomy if unspecified. Detailed prompts beyond a certain threshold tend to reduce the visual quality and are best avoided. Image-conditioned generation did not reproduce the originals exactly but increased diversity, which can be valuable for augmenting under-represented classes once the outputs are vetted.

In summary, this study confirms that both the GPT o4 mini high and Copilot can successfully generate clinically usable synthetic panoramic dental radiographs, each with complementary strengths. The efficacy of image-conditioned and blended generation offers promising avenues for dataset augmentation, although expert validation remains essential. Finally, prompt formulation critically impacts image quality and clinical validity, emphasising the need for a balanced and carefully designed input to optimise AI-generated dental images for practical applications.

5.2. Medical Expert Report

The expert dentist's qualitative evaluation revealed significant differences between the models in terms of anatomical fidelity, pathological accuracy, and overall realism, which provided insight into their strengths and limitations.

Anatomical and Structural Realism The GPT o4 mini high consistently generated images that were closer to clinical reality, especially in terms of tooth morphology and anatomical structures. For example, the pulp chambers and root canals were generally more plausible in the GPT o4 mini high images, although some details of the upper molar root canals remained indistinct. Compared to Copilot's images, Copilot's images typically exhibited less natural transitions between enamel and dentine, along with pulp forms and crown shapes that do not adhere to typical anatomical variations. This uniformity suggests that Copilot tends to repeat patterns rather than provide true anatomical diversity, which may limit its usefulness for clinical training or diagnostic enhancement.

Pathological Feature Representation It was demonstrated that both models have the capacity to depict dental pathologies, including caries and fillings. However, GPT o4 mini high exhibited superior performance in the clinical plausibility of rendering these features. For example, GPT-o4's depiction of root canal fillings and carious lesions was often more anatomically accurate and appropriately localised. However, there were some artificial features, including translucent areas and symmetrical fillings, which may indicate generation artefacts on occasion. In contrast, Copilot's pathological representations were less consistent, with some fillings appearing overly uniform or incorrectly positioned and unnatural border lines similar to the lamina dura, which are rarely obvious on actual radiographs.

Image Quality and Artefact Simulation The GPT o4 images were generally more realistic and clear, and they demonstrated a more convincing simulation of common imaging artefacts such as motion blur and contrast differences. The images produced by Copilot, on the other hand, occasionally contained artificial effects, such as blurry textures and Photoshop-like overlays, which diminished their clinical credibility. Both models caused a deviation from their routine clinical appearance. While both could depict bone trabeculation realistically, errors were observed in the replication of the periodontal structures. However, the GPT o4-generated images demonstrated greater success in simulating bone loss and cyst-like formations.

Clinical and Educational Usability Considering a practical point of view, the outputs of GPT-o4 are more suitable for clinical teaching and AI training purposes due to their higher anatomical and pathological validity. Copilot's limitations in anatomical variability and artefact realism suggest that it may be more suitable for preliminary augmentation tasks rather than complicated clinical applications. The occasional unrealistic features in both models ınderscore the necessity for careful curation and expert review before the integration of synthetic images within diagnostic workflows.

Additional Comparative Insights Further analysis showed distinct differences in how GPT o4 mini high and Copilot handle specific dental features and image enhancements. Despite some morphological discrepancies, GPT o4 mini high more accurately preserved the number and position of erupted and unerupted teeth. However, Copilotfrequently omitted unerupted teeth in the maxilla and demonstrated greater variation in tooth morphology.

Both models struggled with command execution, frequently misinterpreting and misplacing the right and left orientations. However, despite Copilot also exhibiting command-related inaccuracies (these errors), it demonstrated relatively better accuracy in positioning certain features. It is important to explain that, according to the clinical evaluation protocols, the patient's right and left dental profiles correspond inversely to the right and left positions on the screen. Specifically, the right side of the screen represents the patient's left dental profile. Therefore, when giving prompts and interpreting prompt results, it is necessary to explicitly state this information to ensure accurate orientation and avoid confusion. Regarding image enhancement, both models improved clarity from the initially blurry inputs. GPT-o4's enhancements maintained closer anatomical fidelity, while Copilot's outputs, although clearer, lacked consistency in tooth morphology and completeness.

Overall, GPT-o4's synthetic images of mixed dentition and pathological features were judged to be more realistic and clinically plausible compared to Copilot's, which tended to generate images with missing structures and less anatomical variability. Overall, GPT-o4 is the better generator: it produces images that are significantly more realistic, more anatomically accurate, and more clinically useful. This holds even when controlling the prompt content (in the paired test). In contrast, Copilot creates cleaner but less realistic images: it avoids artefacts and may produce sharper visuals. However, these technical qualities do not lead to greater realism or clinical usefulness.

6. Threats To Validity

Generative AI offers great potential to synthesise dental X-ray images, but certain threats may undermine the validity and clinical applicability of these technologies. In this section, we highlight the main threats to the validity of our evaluation.

Algorithmic Bias and Fairness One notable threat is algorithmic bias and fairness. Training datasets often reflect imbalances, such as the overrepresentation of common conditions or data from high-resource settings, while underrepresenting minority groups or rare diseases. This may lead to AI models performing well in the majority group but poorly in underrepresented groups, potentially resulting in existing health disparities. Addressing this requires diverse and well-annotated datasets, fairness-aware training methodologies , and a comprehensive evaluation across demographic and clinical subgroups.

Hallucinations and Clinical Reliability It is referred to as hallucinations when AI produces inaccurate or fabricated medical content due to noisy or incomplete data, insufficient diversity, or a lack of domain-specific

training. AI-generated images can be affected by such errors, resulting in a loss of trust in the technology and posing a threat to patient safety if misleading features lead to misdiagnosis or improper treatment. The mitigation of hallucinations requires rigorous data curation, extensive training on diverse clinical conditions, including rare pathologies, and automated detection methods to identify and flag unreliable outputs.

Exclusive Reliance on a Single Evaluator Another limitation concerns the number of evaluators, as this study relied on the assessment of a single dental specialist. While this expert has nearly ten years of experience as a specialist and is theoretically well-qualified, involving multiple evaluators could provide a broader perspective and reduce potential subjective bias. Incorporating assessments from a diverse panel of clinicians would strengthen the reliability and generalizability of the findings.

Addressing these threats to validity requires the use of more diverse, multi-institutional datasets and greater data variability to ensure robustness and reliability before clinical deployment.

7. Conclusion

In this study, two large-scale multimodal language models, GPT o4 mini high and Copilot, were assessed for their ability to generate synthetic panoramic dental radiographs suitable for both clinical and educational purposes. In total, sixty-one images were produced, 34 using GPT o4 mini high and twenty-seven using Copilot; each image was reviewed independently by a PhD-qualified dentist. The expert dentist evaluated each image, and the resulting assessments were converted into scores on a five-point Likert scale evaluating visual realism, anatomical accuracy, clinical and educational utility, image clarity, and absence of artefacts. The median rating across all 61 images was 4 out of 5, with visual realism also achieving a median score of 4, indicating that most synthetic images were deemed satisfactory and often indistinguishable from authentic radiographs.

After the dental evaluation, quantitative quality indicators were employed as supportive evidence to further investigate the findings. BRISQUE provided no reference spatial naturalness scores, while SSIM and LPIPS quantified similarity in image-conditioned tasks. GPT o4 mini high recorded substantially lower, therefore better, BRISQUE values than Copilot in general and specific prompts, as well as in rare-case scenarios. Copilot achieved slightly cleaner outputs in the adjustment task, whereas GPT o4 mini high performed weakest under highly detailed prompts, suggesting that extreme prompt specificity can degrade the perceived quality. Qualitative analysis revealed distinct trends between the models. GPT o4 mini high demonstrated superior realism, anatomical accuracy, and overall clinical utility, while Copilot produced images with fewer visible artefacts. Both expert evaluation and similarity metrics confirmed that direct image-conditioned prompting did not exactly replicate the source radiographs but encouraged a wider range of plausible variations. Adjustment prompts, such as those adding pathologies or enhancing blurred images, were generally effective, though not perfectly accurate, indicating the potential for augmenting limited or damaged datasets. Notably, the category featuring rare scenarios yielded the cleanest, artefact-free images, whereas the highly detailed prompt category, completed only by GPT o4 mini high, fell short of clinical expectations.

In summary, GPT o4 mini high can generate diagnostically valuable panoramic images but may introduce more artefacts, whereas Copilot typically produces cleaner images with slightly less realism. When guided by carefully crafted prompts specifying laterality, field of view, and intended modifications, both models show promise in alleviating data scarcity, addressing privacy concerns, and enriching rare clinical scenarios. However, expert validation remains essential before these synthetic images are used in downstream deep

learning workflows. As a result of this study, it can be concluded that multimodal generative AI can produce clinically acceptable synthetic dental images that can support AI development while mitigating data limitations.

Peer Review Externally peer-reviewed.
Conflict of Interest The author have no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Author Details

Elif Yilmaz
[1] İstanbul University, Restorative Dentistry, Institute of Graduate Studies in Health Sciences, İstanbul, Türkiye
0000-0001-5713-4286

Abdullah Huzeyfe Köse
[2] İstanbul Technical University, Department of Electrical and Electronic Engineering, İstanbul, Türkiye
[3] Altınbaş University, Department of Electrical and Electronic Engineering, İstanbul, Türkiye
0000-0002-8008-6929

Rahime Yilmaz
[4] İstanbul Aydin University, Department of Computer Engineering, İstanbul, Türkiye
0000-0003-4079-2260

References

[1] R. Baza. 2024. Evaluating artificial intelligence in dental radiography.
[2] F. Umer and N. Adnan. 2024. Generative artificial intelligence: synthetic datasets in dentistry. BDJ Open 10, 1, 13. https://doi.org/10.1038/s41405-024-00144-9
[3] F. Schwendicke and J. Krois. 2022. Data dentistry: how data are changing clinical care and research. J. Dent. Res. 101, 1 (Jan. 2022), 21–29. https://doi.org/10.1177/00220345211041658
[4] L. T. Reyes, J. K. Knorst, F. R. Ortiz, and T. M. Ardenghi. 2021. Scope and challenges of machine learning-based diagnosis and prognosis in clinical dentistry: A literature review. J. Clin. Transl. Res. 7, 4, 523.
[5] T. Wang, Y. Lei, Y. Fu, J. F. Wynne, W. J. Curran, T. Liu, and X. Yang. 2021. A review on medical imaging synthesis using deep learning and its clinical applications. J. Appl. Clin. Med. Phys. 22, 1, 11–36.
[6] L. R. Koetzier, J. Wu, D. Mastrodicasa, A. Lutz, M. Chung, W. A. Koszek, et al. 2024. Generating synthetic data for medical imaging. Radiology 312, 3, e232471.
[7] A. Hosseini and A. Serag. 2025. Is synthetic data generation effective in maintaining clinical biomarkers? Investigating diffusion models across diverse imaging modalities. Front. Artif. Intell. 7, 1454441.
[8] A. B. Abdusalomov, R. Nasimov, N. Nasimova, B. Muminov, and T. K. Whangbo. 2023. Evaluating synthetic medical images using artificial intelligence with the GAN algorithm. Sensors 23, 7, 3440.
[9] M. Farhadi Nia, M. Ahmadi, and E. Irankhah. 2025. Transforming dental diagnostics with artificial intelligence: Advanced integration of ChatGPT and large language models for patient care. Front. Dent. Med. 5, 1456208.
[10] Z. Yang, Z. Yao, M. Tasmin, P. Vashisht, W. S. Jang, F. Ouyang, et al. 2025. Unveiling GPT-4V's hidden challenges behind high accuracy on USMLE questions: Observational study. J. Med. Internet Res. 27, e65146.
[11] S. Tian, Q. Jin, L. Yeganova, P. T. Lai, Q. Zhu, X. Chen, et al. 2024. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief. Bioinform. 25, 1, bbad493.

[12] H. Fukuda, M. Morishita, K. Muraoka, S. Yamaguchi, T. Nakamura, I. Yoshioka, et al. 2025. Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination. J. Dent. Sci. 20, 1, 368–372.

[13] N. Sharmin and A. K. Chow. From text to tissue: Artificial intelligence-generation of biological images. J. Dent. Educ. (in press).

[14] M. Silhadi, W. B. Nassrallah, D. Mikhail, D. Milad, and M. Harissi-Dagher. 2025. Assessing the performance of Microsoft Copilot, GPT-4 and Google Gemini in ophthalmology. Can. J. Ophthalmol. (in press).

[15] D. Stephan, A. Bertsch, M. Burwinkel, S. Vinayahalingam, B. Al-Nawas, P. W. Kämmerer, and D. G. Thiem. 2024. AI in dental radiology —improving the efficiency of reporting with ChatGPT: Comparative study. J. Med. Internet Res. 26, e60684.

[16] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2012. No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21, 12, 4695–4708.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13, 4, 600–612.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 586–595.

[19] M. Büttner, R. Rokhshad, J. Brinz, J. Issa, A. Chaurasia, S. E. Uribe, et al. 2024. Core outcomes measures in dental computer vision studies (DentalCOMS). J. Dent. 150, 105318.

[20] H. M. S. S. Herath, H. M. K. K. M. B. Herath, N. Madusanka, and B. I. Lee. 2025. A systematic review of medical image quality assessment. J. Imaging 11, 4, 100.

[21] S. Pedersen, S. Jain, M. Chavez, V. Ladehoff, B. N. de Freitas, and R. Pauwels. 2025. Pano-GAN: A deep generative model for panoramic dental radiographs. J. Imaging 11, 2, 41.

[22] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, et al. 2023. Denoising diffusion probabilistic models for 3D medical image generation. Sci. Rep. 13, 1, 7303.

[23] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, et al. 2023. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Sci. Rep. 13, 1, 12098.

[24] B. De Wilde, A. Saha, M. de Rooij, H. Huisman, and G. Litjens. 2023. Medical diffusion on a budget: Textual inversion for medical image generation. arXiv preprint arXiv:2303.13430. https://arxiv.org/abs/2303.13430

[25] K. Kokomoto, R. Okawa, K. Nakano, and K. Nozaki. 2021. Intraoral image generation by progressive growing of generative adversarial network and evaluation of generated image quality by dentists. Sci. Rep. 11, 1, 18517.

[26] S. Y. Jeong, E. J. Bae, H. S. Jang, S. Na, and S. Y. Ihm. 2024. Study on virtual tooth image generation utilizing CF-fill and Pix2pix for data augmentation. *Sci. Rep.* 14, 1 (2024), 26772.

[27] S. Pedersen, S. Jain, M. Chavez, V. Ladehoff, B. N. de Freitas, and R. Pauwels. 2025. Pano-GAN: A deep generative model for panoramic dental radiographs. *J. Imaging* 11, 2 (2025), 41.

[28] M. Chaichuk, S. Gautam, S. Hicks, and E. Tutubalina. 2025. Prompt to polyp: Clinically-aware medical image synthesis with diffusion models. *arXiv preprint arXiv:2505.05573*. https://arxiv.org/abs/2505.05573

[29] J. Rousseau, C. Alaka, E. Covili, H. Mayard, L. Misrachi, and W. Au. 2023. Pre-training with diffusion models for dental radiography segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, Cham, Switzerland, 174–182.

[30] R. Di Via, F. Odone, and V. P. Pastore. 2025. Self-supervised pre-training with diffusion model for few-shot landmark detection in X-ray images. In *2025 IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 3886–3896.

[31] I. E. Hamamci, S. Er, E. Simsar, A. Sekuboyina, M. Gundogar, B. Stadlinger, et al. 2023. Diffusion-based hierarchical multi-label object detection to analyze panoramic dental X-rays. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, Cham, Switzerland, 389–399.

[32] F. Khader, G. Müller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, et al. 2023. Denoising diffusion probabilistic models for 3D medical image generation. *Sci. Rep.* 13, 1 (2023), 7303.

[33] A. L. Y. Hung, K. Zhao, H. Zheng, R. Yan, S. S. Raman, D. Terzopoulos, and K. Sung. 2023. Med-cDiff: Conditional medical image generation with diffusion models. *Bioengineering* 10, 11 (2023), 1258.

[34] Y. Shi, A. Abulizi, H. Wang, K. Feng, N. Abudukelimu, Y. Su, and H. Abudukelimu. 2024. Diffusion models for medical image computing: A survey. *Tsinghua Sci. Technol.* 30, 1 (2024), 357–383.

[35] F. Umer and N. Adnan. 2024. Generative artificial intelligence: Synthetic datasets in dentistry. *BDJ Open* 10 (2024), 13. https://doi.org/10.1038/s41405-024-00144-9

[36] S. Pedersen, S. Jain, M. Chavez, V. Ladehoff, B. N. de Freitas, and R. Pauwels. 2025. Pano-GAN: A deep generative model for panoramic dental radiographs. *J. Imaging* 11, 2 (2025), 41.

[37] C. Gao, B. D. Killeen, Y. Hu, R. B. Grupp, R. H. Taylor, M. Armand, and M. Unberath. 2023. Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nat. Mach. Intell.* 5, 3 (2023), 294–308.

[38] S. A. Koohpayegani. 2024. Improving efficiency of deep learning models. Ph.D. Dissertation. University of California, Davis, CA, USA.

[39] D. B. Dangi, B. B. Dangi, and O. Steinbock. 2025. Evaluation of GPT-4o and GPT-4o-Mini's vision capabilities for compositional analysis from dried solution drops. *ACS Omega* 10, 18 (2025), 18955–18959.

[40] M. Sourek. 2025. *Artificial Intelligence in Architecture and the Built Environment: The Revolution Yet to Come*. CRC Press, Boca Raton, FL, USA.

[41] P. Rouzrokh, B. Khosravi, S. Faghani, M. Moassefi, M. M. Shariatnia, P. Rouzrokh, and B. Erickson. 2025. A current review of generative AI in medicine: Core concepts, applications, and current limitations. *Curr. Rev. Musculoskelet. Med.* (2025), 1–21.

[42] Y. Zhang, F. Ye, L. Chen, F. Xu, X. Chen, H. Wu, et al. 2023. Children's dental panoramic radiographs dataset for caries segmentation and dental disease detection. *Sci. Data* 10, 1 (2023), 380.

[43] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2011. Blind/referenceless image spatial quality evaluator. In *Proc. 45th Asilomar Conf. on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, 723–727. https://doi.org/10.1109/ACSSC.2011.6190099

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (Apr. 2004), 600–612.

[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 586–595.

[46] Z. Hanusz, J. Tarasinska, and W. Zielinski. 2016. Shapiro–Wilk test with known mean. *REVSTAT Stat. J.* 14, 1 (2016), 89–100. https://doi.org/10.57805/revstat.v14i1.180

[47] T. W. MacFarland and J. M. Yates. 2016. Kruskal–Wallis H-test for oneway analysis of variance (ANOVA) by ranks. In *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Springer, Cham, Switzerland, 177–211.

[48] D. G. Pereira, A. Afonso, and F. M. Medeiros. 2015. Overview of Friedman's test and post-hoc analysis. *Commun. Stat. Simul. Comput.* 44, 10 (2015), 2636–2653.

[49] R. F. Woolson. 2005. Wilcoxon signed-rank test. In *Encyclopedia of Biostatistics*, P. Armitage and T. Colton (Eds.). John Wiley & Sons, Hoboken, NJ, USA.

[50] J. Hauke and T. Kossowski. 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest. Geogr.* 30, 2 (2011), 87–93. https://doi.org/10.2478/v10117-011-0021-1