ReXavatars: Evaluating Generated Patient Images for Clinical Realism and Reliability

Oishi Banerjee, MS
oishi_banerjee@g.harvard.edu
Department of Biomedical Informatics
Harvard Medical School, Boston, MA

Alexandra N. Willauer, MD
Department of Medicine, Division of Gastroenterology
Massachusetts General Hospital, Boston, MA

Pranav Rajpurkar, PhD
Department of Biomedical Informatics
Harvard Medical School, Boston, MA

Abstract

Generative image models enable the creation of realistic "patient avatars" that can undergo virtual examinations, with potential applications in medical education and AI agent benchmarking. We evaluate Gemini-3's ability to generate such avatars using clinician-verified prompts and clinician assessment of photorealism, composition, and clinical accuracy. We find that images are highly realistic and well-composed, often achieving an impressive level of structural and textural detail. We further find that Gemini-3 can maintain anatomical and environmental consistency across viewpoints or poses in 60% and 70% of cases, respectively, a key requirement for interactive scenarios such as virtual examinations. Despite these strengths, clinical accuracy is inconsistent. When testing without additional constraints, 45 of 80 pathologies tested are generated without errors; performance degrades when we increase task complexity by requiring specific presentations or combinations of pathologies. Additionally, we find notable weaknesses in numerical reasoning tasks, with a 45% success rate on tasks requiring accurate counting. These results highlight the promise of patient avatars as a new interface for education and evaluation, while showing directions to improve clinical accuracy.

1. Introduction

A cornerstone of the practice of medicine is the physical exam. This skill requires clinicians to evaluate the patient based on visual, auditory, and tactile inspection and recognize a variety of normal and abnormal physical exam findings (Bickley et al., 2021; Elder et al., 2017). Despite technological advances in medicine, the physical exam remains a crucial skill for clinicians, yet past work on AI for medical imaging has focused on subspecialized areas, such as fundoscopy images in ophthalmology (Zhou et al., 2022; Masalkhi et al., 2026; Pandey et al., 2025) or wounds and other skin lesions in dermatology (Sarp et al., 2021; Ghorbani et al., 2019; Izu-Belloso et al., 2025). To the best of our knowledge, our work is the

first to investigate image generation to simulate the visual component of a comprehensive physical exam in the clinical environment.

In this work, we test whether a state-of-the-art image generation model can convincingly represent a wide range of abnormal findings that may be seen during the physical exam, building a test suite covering 80+ unique medical conditions and 360 different variations. Through clinician-verified prompts, we test whether avatars can realistically capture general conditions (e.g. "syndactyly") and more specific presentations (e.g. "syndactyly affecting one hand with webbing connecting only the second and third fingers at the distal interphalangeal joints"). Furthermore, we investigate whether models can generate patient "avatars" that remain consistent across frames, reliably representing a specific patient's physiology and environment while responding to different instructions. These patient avatars enable interactive experiences, where a user can conduct a dynamic examination by asking for new viewpoints and poses. Using Gemini-3 as a state-of-the-art image generation engine, we find that current avatars can convincingly represent complex conditions (Figure 1), simulating 45 of 80 core pathologies without error when we request no additional elements. They also maintain consistency across multiple viewpoints and poses in 60% and 70% of experiments, respectively. However, there are still frequent inaccuracies, particularly when we incorporate additional constraints around numerical reasoning and other fine details.

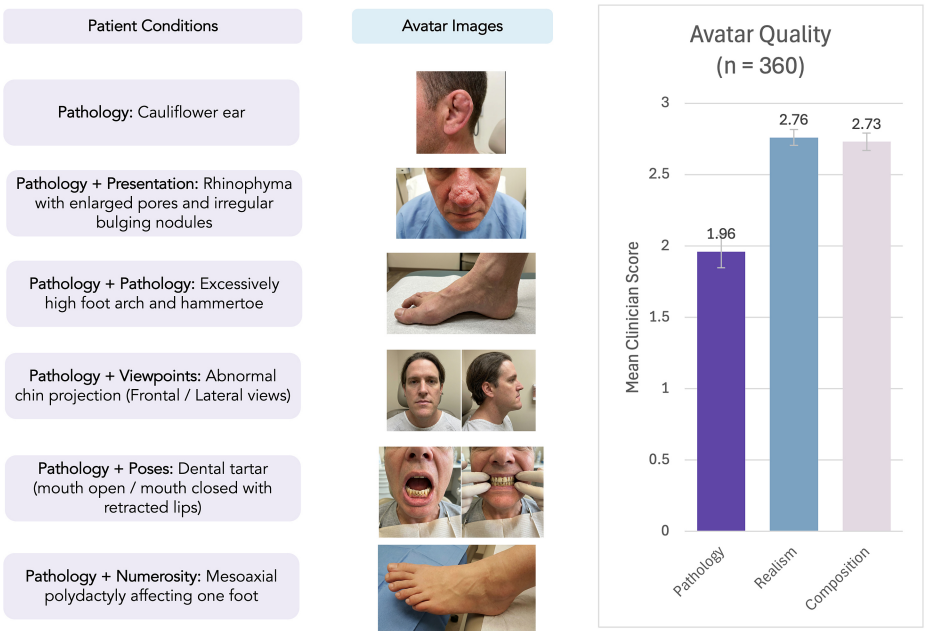


Figure 1: A state-of-the-art image generation model can simulate a wide range of pathologies, showing an impressive command of textural, structural, and color-based abnormalities. While it generally excels in realism and composition, clinical accuracy is still inconsistent, especially on more complex tasks.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Patient avatars can bridge the gap between static AI datasets and dynamic clinical scenarios.** Avatars introduce a new paradigm of "virtual patient exams" for evaluating clinical agents and also have potential applications in medical education.
- **Current AI-generated medical photographs show strong realism and consistency across frames.** Images generally achieve high scores on photorealism and composition. Additionally, we find that 60–70% of avatars remain anatomically and environmentally consistent across two frames, a vital capability for enabling interactions.
- **Current images exhibit promising but inconsistent performance when depicting pathologies.** Many pathologies were simulated accurately by themselves, but not when they co-occurred with other pathologies. Similarly, the model struggled to execute some specific instructions involving fine-grained details or numerical reasoning. Taken together, there is room for improvement in composability and controllability.

2. Related Work

Recent work on synthetic medical images has concentrated on standardized diagnostic imaging modalities, such as radiographs and fundoscopic imaging (Moroianu et al., 2025; Pandey et al., 2025; Zhou et al., 2022; Masalkhi et al., 2026) rather than ordinary photographs. On the photography side, synthetic-image research has mostly remained within narrow subspecialties, especially dermatology and ophthalmology (Malihi et al., 2023; Sarp et al., 2021; Ghorbani et al., 2019; Izu-Belloso et al., 2025). The diverse photographs needed for a whole-body physical exam have been included in some medical image benchmarks (Banerjee et al., 2026; Zhang et al., 2024); however, these works have focused on image comprehension rather than generation. Notably, Banerjee et al. (2026) found that Gemini-3 substantially outperformed peer models in comprehension of medical photographs, motivating our choice of model for this study. Outside of medicine, some image-generation benchmarks have scored images on fine-grained dimensions, such as counting and object/concept co-occurrence (Ghosh et al., 2023; Chen et al., 2025), which we likewise consider here.

Another related line of work studies virtual patients, standardized-patient simulation, and OSCE-style evaluation with large language models. These studies focus primarily on LLM-driven dialogue agents that act as the standardized patient and/or clinician, often emphasizing text-only interactions (Zeng et al., 2025; Rao et al., 2026; Bentegeac et al., 2025). Studies evaluating multimodal conversational and visual assessment capabilities remain the exception rather than the norm (Schmidgall et al., 2026; Saab et al., 2026; Johri et al., 2025; Kang et al., 2025). Moreover, when images have been incorporated into virtual-patient systems, they are typically not generated by the LLM itself nor linked to the emerging patient avatar (Voigt et al., 2025). Thus, our study provides an initial framework for and assessment of generative images, rather than retrieval-based images, for LLM-simulated patient avatars.

3. Methods

3.1. Prompt Generation

We constructed a diverse, clinician-verified prompt set designed to probe both high-level realism and fine-grained clinical accuracy in the generation of patient avatars. Our prompts are largely focused on a core set of **80 conditions** that (1) can be clearly observed by visual examination (as opposed to requiring internal imaging), (2) feature a diverse range of anatomical regions (Table 1), and (3) cover multiple types of conditions (Table 2), including congenital conditions, musculoskeletal disorders, dental pathology, and other findings of acute and chronic illness. We de-emphasize dermatological conditions, as these have received more attention in prior literature. Appendix A contains the complete list of conditions.

Table 1: Condition Coverage by Anatomical Region

Region	**Example Conditions**
Face/Head	Strabismus; Cleft lip and palate; Cauliflower ear
Spine/Neck	Scoliosis; Torticollis; Cervical lymphadenopathy
Torso	Pectus carinatum; Scapular winging
Upper Limb	Dupuytren's contracture; Arm amputation; Arachnodactyly
Lower Limb	Bunion; Baker's cyst; Knee valgus

Table 2: Condition Coverage by Type

Type	**Example Conditions**
Congenital abnormalities	Microcephaly; Syndactyly
Musculoskeletal disorders	Shoulder dislocation; Arm amputation
Dental pathology	Dental caries; Hypodontia
Findings of acute illness	Abnormal pupil size; Cervical lymphadenopathy
Findings of chronic illness	Rhinophyma; Dental caries

The prompts were organized into six categories to capture different evaluation dimensions (Table 3). Five of the categories stemmed from our core set of 80 conditions. In our first category **("Pathology")**, we simply asked for an image depicting the condition. In the second category **("Pathology+Presentation")**, we requested a specific presentation of the condition, including fine-grained details about location, severity, coloration, and other aspects. In the third category **("Pathology+Pathology")**, we requested images showing two conditions appearing together; each of our 80 conditions appears in exactly two of these prompts. We initially paired conditions randomly and then manually rearranged them to ensure that both conditions were compatible and could reasonably be demonstrated in a single image. In the fourth category **("Pathology+Viewpoints")**, we requested an image with two frames, showing a single patient with the condition from two specific views (e.g. frontal view and lateral view). We randomly sampled 40 conditions for this category. In the fifth category **("Pathology+Poses")**, we requested an image with two frames, now showing a single patient with the condition performing two different, clinically meaningful poses while viewed from specific angles. For example, to evaluate digital clubbing, the requested poses were "fingers extended, providing camera with a dorsal view of hands" and "lateral

view of index fingers, with other fingers curled in, allowing assessment of Lovibond angles". We again randomly sampled 40 conditions for this fifth category, replacing conditions in which clinicians were unlikely to request two different poses.

In the sixth category **("Pathology + Numerosity")**, we focused on a different set of 12 conditions that test whether models can count accurately (Appendix B). We included some small abnormalities that can appear in clusters (e.g. warts, moles, insect bites). We also included conditions that require models to count digits or teeth to achieve proper placement (e.g. syndactyly connecting second and third fingers, tooth decay affecting only three incisors). Finally, we included conditions that required models to produce an ordinary anatomical feature but in an atypical quantity (e.g. oligodactyly affecting toes, supernumerary nipple). We included variations on each of these conditions, resulting in 40 prompts total.

Table 3: Prompt Categories

Category	**# Samples**	**Example**
Pathology	80	Polydactyly affecting fingers
Pathology + Presentation	80	Mesoaxial polydactyly affecting one hand
Pathology + Pathology	80	Asymmetric arm length and polydactyly
Pathology + Viewpoints	40	Frontal and lateral view of dorsal hump on nose
Pathology + Poses	40	Polydactyly with hands raised vs. resting on table
Pathology + Numerosity	40	Tooth decay affecting only three incisors

There were **80** prompts for each of the "Pathology," "Pathology+Presentation" and "Pathology+Pathology" categories, and **40** for "Pathology+Viewpoints," "Pathology+Poses," and "Pathology+Numerosity" categories. While some candidate prompts were initially generated using GPT-5, all prompts were **reviewed and verified by a board-certified internal medicine physician** to ensure that they were realistic, diverse, and visually interpretable. The prompts were also revised as necessary to avoid implausible or unobservable combinations (e.g., conditions requiring incompatible viewpoints) and to increase the diagnostic challenge while preserving clarity. Finally, for prompts involving anatomical exposure (e.g., torso conditions), we explicitly specified "non-sexual medical image of male patient" to comply with safety constraints in image generation systems.

3.2. Image Generation

We generated images using the Gemini-3-Pro-Image-Preview model via the Google GenAI Python SDK; full generation settings and prompt templates are provided in Appendix C. We chose Gemini-3 because it substantially outperformed peer models in a recent visual question-answering benchmark covering similar medical photographs (Banerjee et al., 2026). Given that approximately 50 hours of clinician effort were required for prompt validation and image evaluation, we prioritized a thorough evaluation of one strong model rather than a shallow comparison across multiple models. For comparison, a smaller analysis of 80

GPT-5.5 pathology images yielded lower realism scores (2.12 vs. 2.78), with frequently exaggerated presentations; results are reported in Appendix D.

For single-frame categories, each prompt requested a single image. While we instructed the model to produce images with a single frame, it produced multi-frame images 13.5% of the time (38 instances out of 280; e.g. showing a patient with and without the condition). In these cases, we recorded the score of the highest-scoring frame for Pathology, Pathology+Presentation, and Pathology+Numerosity categories. Some Pathology+Pathology images contained two frames, typically centering one pathology per frame. We only penalized the Composition score for failing to capture both pathologies in one frame. For multi-frame categories (Pathology+Viewpoints, Pathology+Poses), each prompt requested an image with two frames corresponding to the specified variations.

3.3. Evaluation Protocol

All generated images were evaluated by the clinician on three primary dimensions:

- **Pathology**: Whether the target condition is depicted accurately and without introducing separate abnormalities. For the Pathology+Pathology conditions, we scored both pathologies separately and then averaged them.
- **Realism**: Whether the generated image is perceived as realistic. Realism errors include anatomically implausible presentations as well as errors in environmental physics.
- **Composition**: Whether the viewpoint, pose, clothing, and scene elements are well-chosen, making it possible to clearly see the abnormality without obstructions. Composition scoring also considers whether images contain one or two frames as specified.

We also used the following category-specific criteria:

- **Presentation**: Whether the model executed the specific details requested for this presentation of the pathology. We note that it is possible to have a high Pathology score and a low Presentation score (e.g. if an image shows the right abnormality but in the wrong location) or a low Pathology score with a high Presentation score (e.g. if an image precisely displays the requested abnormality but also introduces a new, unwanted abnormality).
- **Consistency**: Whether the model maintained consistency in the patient's anatomy and surroundings across frames. This dimension applies to Pathology+Viewpoints and Pathology+Poses prompts.
- **Viewpoint**: Whether the model captured the requested camera angles for Pathology+Viewpoints prompts. We averaged the scores for both frames.
- **Pose**: Whether the model captured the requested poses for Pathology+Poses prompts, while following any anatomical constraints arising from the pathology. We averaged the scores for both frames.
- **Counting**: Whether the model counted successfully for Pathology+Numerosity prompts.

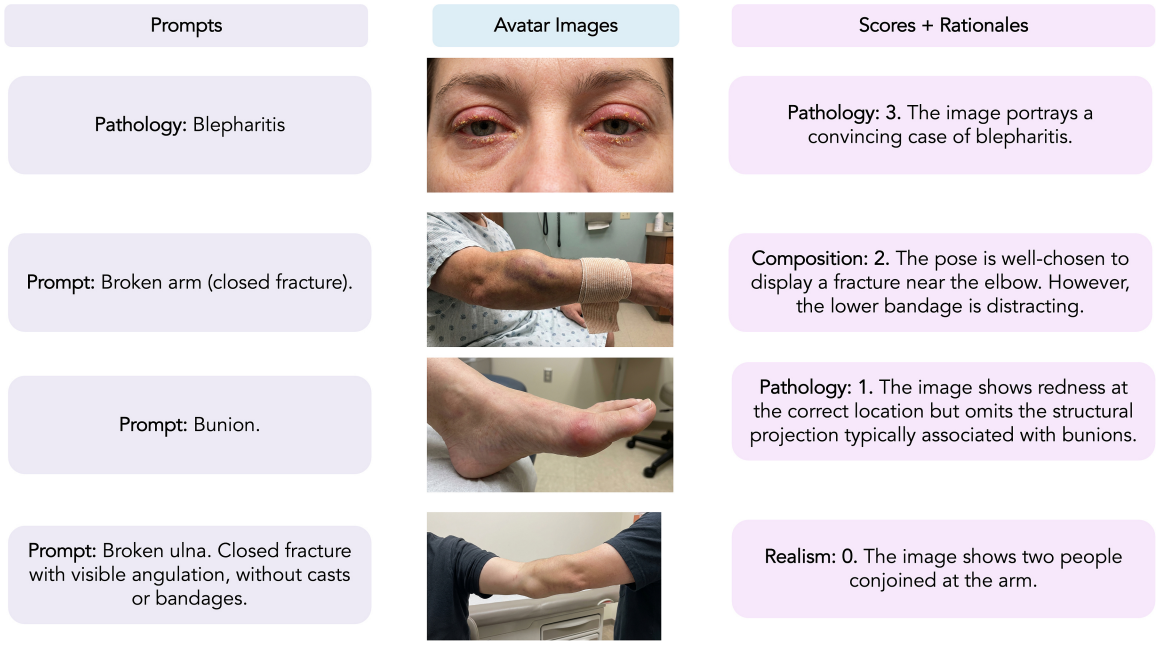


Figure 2: Examples of our clinician evaluation process, with scores ranging from 0 to 3.

Using these scoring dimensions, a clinician scored each image from 0–3. A 3 reflects perfect performance, while a 2 reflects near-perfect performance with minor inaccuracy. A 1 indicates that an image has some promise but contains one or more significant errors. Images that are entirely incorrect or have no abnormality present (i.e., image produced with normal characteristics) receive a 0. For the "Counting" dimension, scores were binarized, with images receiving either a 0 or a 3. We provide examples of each error severity in Figure 2. We report 95% confidence intervals (t-based for mean clinician scores, Wilson for proportions of images without errors). Additionally, half of the images from each category were randomly sampled and scored by a second clinician; we averaged clinicians' scores where available. For inter-rater reliability, Gwet's AC2 and the mean absolute difference across each dimension were reported. Lastly, it is important to represent diverse patient avatars, so we evaluated current gender distributions in the generation of patient avatars (Appendix E).

4. Results

4.1. Realism and Composition

Figure 3 summarizes scoring for the Pathology, Pathology+Presentation, and Pathology+Pathology categories. Overall, Gemini-3 produced images with excellent photorealism and composition. Mean realism scores were high across all categories, ranging from 2.55 (95% CI: [2.29, 2.81]) to 2.91 (95% CI: [2.83, 2.99]). These scores indicate that the generated images are broadly perceived as visually convincing. Similarly, composition scores remained strong across categories, ranging from 2.39 (95% CI: [2.21, 2.57]) to 2.95 (95% CI: [2.89, 3.01]). In general, the model reliably selected the appropriate camera angles, poses, and scene

elements. In particular, the images typically depicted patients with clear, unobstructed views of the relevant anatomy, using clothing and positioning that highlight—rather than obscure—the target pathology. The lowest Composition scores occurred in the Pathology+Pathology category, where the model struggled to produce two separate anatomical abnormalities simultaneously in a single image.

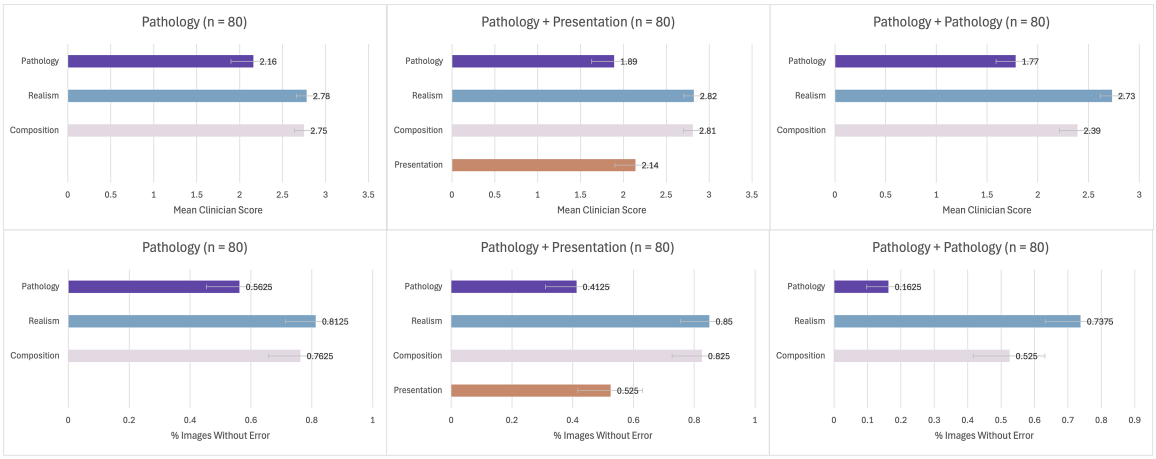


Figure 3: Results from Pathology, Pathology+Presentation, and Pathology+Pathology categories. We find that images are generally realistic and well-composed. The majority of images also display the correct pathology in isolation, though performance drops when requesting a specific presentation or a combination of two pathologies.

4.2. Single-Frame Images

We found that 56.2% (45/80) (95% CI: [0.45, 0.67]) of images perfectly depicted the requested finding in our "Pathology" category, with an average score of 2.16 (95% CI: [1.90, 2.42]). However, more complex prompts proved challenging. When moving from the plain "Pathology" prompts to the "Pathology+Presentation" prompts, mean Pathology scores dropped to 1.89 (95% CI: [1.63, 2.15]); 41.2% (33/80) (95% CI: [0.31, 0.52]) of images were free of pathology errors, while 52.5% (42/80) were free of presentation errors. We noticed additional issues when combining conditions in the "Pathology+Pathology" category, with a mean Pathology score of 1.77 (95% CI: [1.58, 1.97]) and only 16.2% (13/80) (95% CI: [0.10, 0.26]) of images representing both pathologies perfectly. Generally, even if a pathology could be displayed correctly in isolation, this did not guarantee success once additional elements were added (Fig. 4). Lastly, single-frame prompts produced multi-frame outputs 13.5% of the time (38 instances out of 280 total), and points were docked from the composition score accordingly.

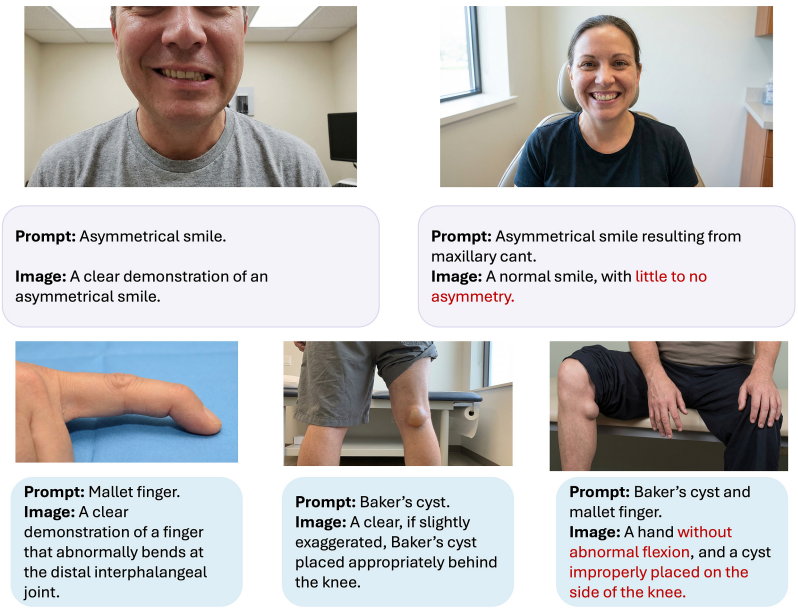
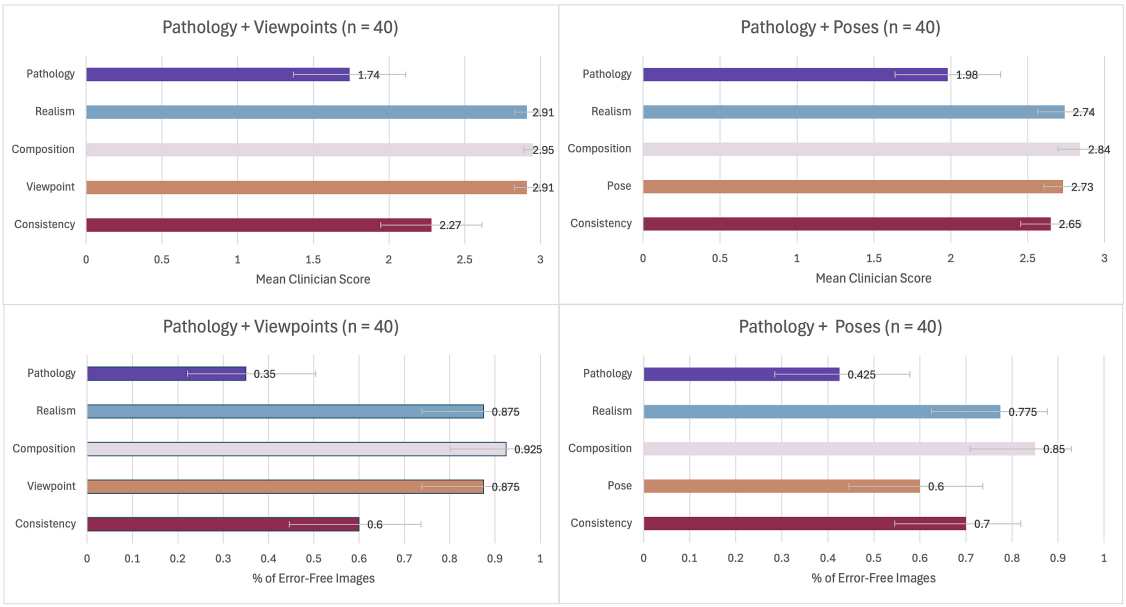


Figure 4: In some cases, images can represent a pathology in isolation, but not when additional elements are requested.

4.3. Multi-Frame Images

We found mixed success with multi-frame images, which show patients from two viewpoints or with two distinct poses. In the "Pathology+Viewpoints" category, only 35.0% (n=14/40, 95% CI: [0.22, 0.50]) of images were free from pathology errors, with a mean Pathology score of 1.74 (95% CI: [1.37, 2.11]). However, 87.5% (n=35/40, 95% CI: [0.74, 0.95]) of images featured the correct viewpoints with an average Viewpoint score of 2.91 (95% CI: [2.83, 2.99]), and 60.0% (n=24/40, 95% CI: [0.45, 0.74]) maintained consistency across both frames, receiving a mean Consistency score of 2.27 (95% CI: [1.94, 2.61]). We observed stronger performance in the "Pathology+Poses" category, where 42.5% (n=17/40, 95% CI: [0.29, 0.58]) of images were free of pathology errors with an average Pathology score of 1.98 (95% CI: [1.63, 2.32]). The average Pose score was 2.73 (95% CI: [2.61, 2.86]), and 60.0% (n=24/40, 95% CI: [0.45, 0.74]) of images were free of pose errors. Additionally, we again observed that the model was capable of preserving anatomical and environmental consistency across frames, with an average clinician score of 2.65 (95% CI: [2.45, 2.85]), and 70.0% (n=28/40, 95% CI: [0.55, 0.82]) of images being free of consistency errors.

We show examples of successful multi-frame images in Fig. 5. Unsuccessful multi-frame images frequently involved directional errors (e.g., attributing the abnormality to the left hand in one viewpoint and then to the right hand in the other viewpoint) or misplaced furniture (e.g., the background remained static while the patient was rotated instead of changing the entire spatial orientation between viewpoints), rather than deeper anatomical errors. For completeness, we annotated the lowest-scoring frame of unwanted multi-frame images and recomputed our results for the "Pathology" and "Pathology+Presentation"



Maintaining Environmental and Anatomical Consistency

Dental Tartar (Frontal / Oblique Views)

Arm Amputation (Frontal / Lateral Views)

Clinodactyly
Patient lifting hands up with fingers extended, frontal view / Both hands resting on a table, palm down and fingers extended to the extent possible, superior view

Abnormal Tilt of Ear
(Neutral sitting position, head facing forward, frontal view / Neutral sitting position, head turned to the side, frontal view)

Figure 5: In 60% of "Pathology+Viewpoints" cases and 70% of "Pathology+Poses" cases, we observed no consistency errors, even when portraying detailed features such as teeth. This work suggests that avatars can remain coherent across interactions.

images (Appendix F); using the lowest-scoring frame instead of the highest-scoring frame had little impact on our findings.

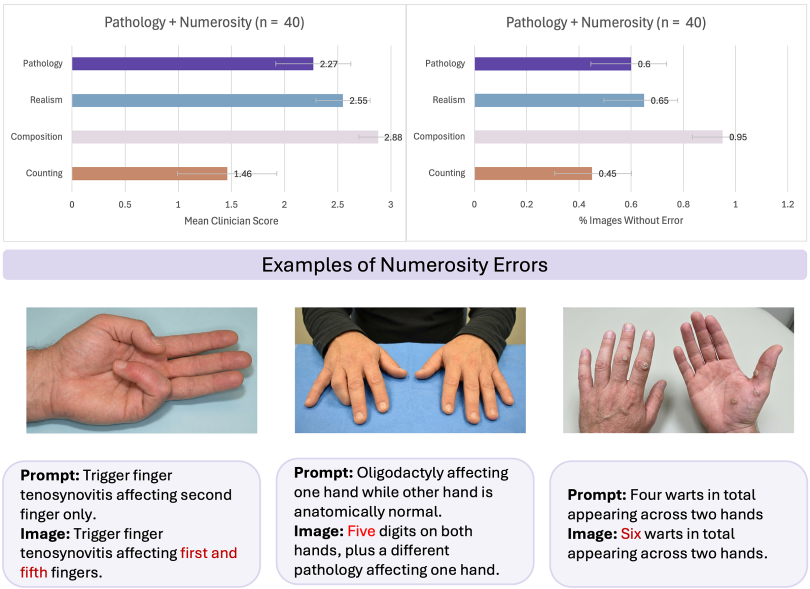


Figure 6: Results from the Pathology+Numerosity category. We find that Gemini-3 struggles with numerosity, and this subsequently affects several tasks, including identifying specific digits and displaying a specific number of lesions.

4.4. Numerosity

The "Pathology+Numerosity" category revealed a consistent error pattern around counting (Fig. 6). The images received average scores of 2.27 (95% CI: [1.92, 2.63]), 2.55 (95% CI: [2.29, 2.81]), and 2.88 (95% CI: [2.69, 3.06]) for Pathology, Realism, and Composition, respectively. However, they received an average score of only 1.46 (95% CI: [1.00, 1.93]) for Counting, as only 45.0% (n=18/40, 95% CI: [0.31, 0.60]) of images successfully executed the quantitative aspect of their prompt. This struggle with numerosity affected a wide range of tasks. For example, the model struggled to localize pathologies to specific digits or generate images with an abnormal number of digits. When generating abnormalities such as warts and insect bites that can appear in groups, it could not consistently generate the target quantity, instead showing a tendency to produce more abnormalities than needed.

4.5. Inter-rater reliability

A random sample of half of the images from each category was scored by a second clinician. We computed Gwet's AC2 and the mean absolute difference across each dimension and found strong overall agreement (Table 4).

Table 4: Inter-rater agreement between clinician evaluators, measured by mean absolute difference and Gwet's AC2. Mean absolute difference is reported on the 0–3 scoring scale. Both metrics are reported with 95% confidence intervals.

Dimension	Mean Absolute Difference	Gwet's AC2 Agreement
Pathology	0.67 [0.54, 0.79]	0.64 [0.53, 0.75]
Realism	0.27 [0.18, 0.35]	0.94 [0.91, 0.97]
Composition	0.33 [0.24, 0.43]	0.91 [0.88, 0.95]
Presentation	0.78 [0.45, 1.10]	0.65 [0.41, 0.90]
Consistency	0.40 [0.17, 0.63]	0.88 [0.78, 0.99]
Pose	0.42 [0.17, 0.68]	0.77 [0.51, 1.00]
Viewpoint	0.10 [0.00, 0.25]	0.93 [0.81, 1.00]
Counting	0.45 [0.00, 0.93]	0.70 [0.36, 1.00]

5. Discussion

Our results demonstrate that modern generative models have made substantial progress toward producing clinically meaningful patient avatars. Across a wide range of conditions, Gemini-3 generates images that are consistently photorealistic and well-composed, receiving high clinician scores on both dimensions. The model reliably selects appropriate camera angles, poses, and scene elements, often presenting the pathology in a clear and unobstructed manner. Additionally, we see emerging capabilities around simulating complex pathologies, with 45 of our 80 conditions being generated without errors in the baseline "Pathology" category. We found that models can achieve impressively lifelike detail across a wide array of physiological and structural abnormalities that encompass many anatomic locations. Additionally, we observed promising performance when requiring models to show a single patient using multiple viewpoints or poses, a core capability for bringing a patient avatar to life. In the future, these frames could potentially be used to ground video generation methods, producing patient avatar videos.

At the same time, our results highlight directions for future improvement. When asked to generate particular presentations of a condition, the model can struggle to incorporate the requested details, sometimes compromising the core pathology to do so. We see similar issues when requesting images that show two pathologies; pathologies that are successfully generated separately sometimes cause errors when combined. Additional work is needed to improve composability and controllability, allowing the specification of fine-grained details. Furthermore, we observed poor numerical reasoning, which also limits the model's command of fine details.

These results open the door to a "virtual physical exam," with potential applications in medical education and clinical AI. For example, trainees could hone their physical exam skills using patient avatars, especially for abnormal findings seen in rare medical conditions. Alternatively, AI-generated patient avatars could be used for simulating telemedicine encounters, which focus on visual findings without any touch-based component. Excitingly, patient avatars can also advance the training and evaluation of clinical AI agents, providing

an interactive testbed where agents can practice conducting exams without risk to patient safety. The prompts, generated images, and scores are available here.

Limitations. First, we do not evaluate turn-based or interactive generation, where a patient avatar is updated across multiple sequential prompts. Although Gemini-3 demonstrates reasonable consistency within multi-frame images, maintaining consistency across a longer interaction (e.g., a simulated clinical encounter) may be substantially more challenging. Second, we observed limited patient demographics, with patients varying in apparent gender and age but generally having lighter skin tones. Additional work should explore whether Gemini-3 can maintain accuracy while depicting more diverse patients. Third, we generate and evaluate only a single image per prompt. This setup measures Gemini-3's base level of performance, but more complex strategies could improve reliability, such as sampling multiple candidates and selecting the best output or using iterative refinement (e.g., actor-critic or self-correction). We expect that such approaches will meaningfully increase performance and represent an important direction for future work.

Acknowledgments

Many thanks to Christine Zhou, MD, and Sung Eun Kim, MD, for assistance in image evaluation. O.B. was supported by the Biswas Family Foundation's Transformative Computational Biology Grant in collaboration with the Milken Institute.

References

Oishi Banerjee, Sung Eun Kim, Alexandra N. Willauer, Julius M. Kernbach, Abeer Rihan Alomaish, Reema Abdulwahab S. Alghamdi, Hassan Rayhan Alomaish, Mohammed Baharoon, Xiaoman Zhang, Julian Nicolas Acosta, Christine Zhou, and Pranav Rajpurkar. ReXInTheWild: A unified benchmark for medical photograph understanding, 2026. URL https://arxiv.org/abs/2603.19517.

R. Bentegeac et al. ECOSBot: a multicenter validation pilot study of a generative AI tool for OSCE-based nephrology training. *Clinical Kidney Journal*, 18(10):sfaf308, 2025. doi: 10.1093/ckj/sfaf308.

Lynn S. Bickley, Peter G. Szilagyi, and Richard M. Hoffman. *Bates' Guide to Physical Examination and History Taking, 13e*. Lippincott Williams & Wilkins, a Wolters Kluwer business, January 2021. ISBN 978-1-496398-17-8.

Kaijie Chen, Zihao Lin, Zhiyang Xu, Ying Shen, Yuguang Yao, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. R2I-Bench: Benchmarking reasoning-driven text-to-image generation. 2025. URL https://arxiv.org/abs/2505.23493.

Andrew T. Elder, I. C. McManus, A. Patrick, K. Nair, L. Vaughan, and Jane Dacre. The value of the physical examination in clinical practice: an international survey. *Clinical Medicine*, 17(6):490–498, Dec 2017. doi: 10.7861/clinmedicine.17-6-490.

Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. DermGAN: Synthetic generation of clinical skin images with pathology. 2019. URL https://arxiv.org/abs/1911.08716.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Rosa Maria Izu-Belloso, Rafael Ibarrola-Altuna, and Alex Rodriguez-Alonso. Generative adversarial networks in dermatology: A narrative review of current applications, challenges, and future perspectives. *Bioengineering*, 12(10), 2025. ISSN 2306-5354. doi: 10.3390/bioengineering12101113. URL https://www.mdpi.com/2306-5354/12/10/1113.

S. Johri et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86, 2025. doi: 10.1038/s41591-024-03328-5.

S. Kang et al. Physical examination identification in medical education videos: Zero-shot multimodal AI with temporal sequence optimization study. *JMIR AI*, 4:e76586, 2025. doi: 10.2196/76586.

Leila Malihi, Ursula Hübner, et al. Can synthetic images improve CNN performance in wound image classification? In *Caring is Sharing – Exploiting the Value in Data for Health and Innovation*, Studies in Health Technology and Informatics, 2023. doi: 10.3233/SHTI230311.

M. Masalkhi, K. Sporn, R. Kumar, et al. Ophthalmic image synthesis and analysis with generative adversarial network artificial intelligence. *Journal of Digital Imaging*, 39:732–754, 2026. doi: 10.1007/s10278-025-01519-1.

Stefania L. Moroianu, Christian Bluethgen, Pierre Chambon, Mehdi Cherti, Jean-Benoit Delbrouck, Magdalini Paschali, Brandon Price, Judy Gichoya, Jenia Jitsev, Curtis P. Langlotz, and Akshay S. Chaudhari. Improving performance, robustness, and fairness of radiographic AI models with finely-controllable synthetic data, 2025. URL https://arxiv.org/abs/2508.16783.

Prashant U Pandey, Jonathan A Micieli, Stephan Ong Tone, Kenneth T Eng, Peter J Kertes, and Jovi C Y Wong. Realistic fundus photograph generation for improving automated disease classification. *British Journal of Ophthalmology*, 109(7):791–798, 2025. ISSN 0007-1161. doi: 10.1136/bjo-2024-326122. URL https://bjo.bmj.com/content/109/7/791.

A. S. Rao et al. Large language model performance and clinical reasoning tasks. *JAMA Network Open*, 9(4):e264003, 2026. doi: 10.1001/jamanetworkopen.2026.4003.

K. Saab et al. Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 32(5):1726–1736, 2026. doi: 10.1038/s41591-026-04371-0.

Salih Sarp, Murat Kuzlu, Emmanuel Wilson, and Ozgur Guler. WG2AN: Synthetic wound image generation using generative adversarial network. *The Journal of Engineering*, 2021(5):286–294, 2021. doi: 10.1049/tje2.12033. URL https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/tje2.12033.

S. Schmidgall et al. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02674-7.

H. Voigt et al. LLM-powered virtual patient agents for interactive clinical skills training with automated feedback. *arXiv preprint arXiv:2508.13943*, 2025. doi: 10.48550/arXiv.2508.13943.

J. Zeng et al. Embracing the future of medical education with large language model-based virtual patients: Scoping review. *Journal of Medical Internet Research*, 27:e79091, 2025. doi: 10.2196/79091.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering, 2024. URL https://arxiv.org/abs/2305.10415.

Yi Zhou, Boyang Wang, Xiaodong He, Shanshan Cui, and Ling Shao. DR-GAN: Conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. *IEEE Journal of Biomedical and Health Informatics*, 26(1):56–66, 2022. doi: 10.1109/JBHI.2020.3045475.

Appendix A. 80 Core Pathologies

Table 5: List of 80 main conditions.

Condition	**Condition**
Microcephaly	Frontal bossing
Temporomandibular joint dislocation	Strabismus
Blepharitis	Unilateral drooping eyelid
Abnormal eye protrusion	Ulnar drift
Coloboma affecting the eyelid	Madarosis affecting eyelids
Abnormal pupil size	Eyebrow ptosis
Anophthalmia	Overbite
Underbite	Asymmetric mouth opening
Macroglossia	Asymmetric smile
Abnormal chin projection	Swollen parotid glands
Cleft palate	Preauricular pit
Asymmetrical ear size	Microtia
Abnormal tilt of ear	Notch in ear
Cauliflower ear	Off-center columella
Cleft lip	Rhinophyma
Saddle nose deformity	Dorsal hump in nose
Arrhinia	Nasal tip collapse
Scoliosis	Kyphosis
Shoulder dislocation	Scapular winging
Hyperlordosis	Torticollis
Cervical lymphadenopathy	Neck hump
Goiter	Webbed neck
Pectus excavatum (non-sexual, medical image of male patient)	Pectus carinatum (non-sexual, medical image of male patient)
Broken arm (closed fracture)	Polydactyly affecting fingers
Asymmetric arm length	Arm amputation

Table 6: List of 80 main conditions (continued).

Condition	Condition
Arachnodactyly	Brachydactyly
Digital clubbing affecting fingers	Clinodactyly affecting fingers
Syndactyly affecting fingers	Mallet finger
Dupuytren's contracture	Trigger finger tenosynovitis
Boutonniere deformity	Swan neck contracture
Excessively high foot arch	Excessively flat foot arch
Bunion	Hammertoe
Foot drop	Gout affecting foot
Hip abductor weakness	Leg amputation
Broken leg (closed fracture)	Subluxation of patella
Knee joint effusion	Knee valgus
Knee varus	Genu recurvatum
Baker's cyst	Dental caries
Gum hyperplasia	Hypodontia
Dental tartar	Tooth abfraction

Appendix B. 12 Numerosity Pathologies

Table 7: List of 12 conditions for testing numerical reasoning.

Condition	Condition
Trigger finger tenosynovitis	Syndactyly
Tooth decay	Supernumerary nipple
Oligodactyly affecting hand	Polydactyly affecting hand
Oligodactyly affecting foot	Polydactyly affecting foot
Warts	Pimples
Insect bites	Moles

Appendix C. Image Generation Settings and Prompt Templates

We generated images using Gemini-3-Pro-Image-Preview through the Google GenAI Python SDK. We used the Standard `v1beta` API with default hyperparameters and no additional preprocessing. The package version was `google-genai==1.59.0`. Images were generated at a resolution of 1408×768 pixels, with one sample generated per prompt.

Prompt Template for Pathology, Pathology+Presentation, and Pathology+Numerosity

For the Pathology, Pathology+Presentation, and Pathology+Numerosity categories, we used the following prompt template:

Generate a single clinical photograph. Requirements:

- Absolutely NO text, labels, arrows, diagrams, charts, or any graphical overlays on the image.
- Show an adult patient with a clear, unmistakable presentation of this condition, while also following any additional instructions: [Condition prompt here]
- If multiple poses or viewpoints are requested, make one image for each. For example if frontal and lateral views are requested, you need to make two images.
- Never put multiple panels or frames in a single image.
- Only output images. Do not include text or other annotations.
- Choose pose, viewpoint, framing, clothing, and setting as needed so the condition is obvious to a clinician.
- Use natural lighting, realistic skin and fabric, and a photorealistic style.
- Avoid depicting other medical abnormalities unless they would necessarily result from the condition.

Prompt Template for Pathology+Pathology, Pathology+Poses, and Pathology+Viewpoints

For the Pathology+Pathology, Pathology+Poses, and Pathology+Viewpoints categories, we used the following prompt template:

Generate a single clinical photograph. Requirements:

- Absolutely NO text, labels, arrows, diagrams, charts, or any graphical overlays on the image.
- Show an adult patient with a clear, unmistakable presentation of this condition, while also following any additional instructions: [Condition prompt here]
- If multiple poses or viewpoints are requested, include them in a single image by using multiple panels or frames.
- Only output images. Do not include text or other annotations.
- Choose pose, viewpoint, framing, clothing, and setting as needed so the condition is obvious to a clinician.
- Use natural lighting, realistic skin and fabric, and a photorealistic style.
- Avoid depicting other medical abnormalities unless they would necessarily result from the condition.

Appendix D. GPT-5.5 Analysis

As a point of comparison for Gemini-3, we generated a smaller set of images with GPT-5.5, using the same 80 core "Pathology" prompts and the same 0–3 rubric described in the Methods (one image per condition, 80 images total). This comparison set was annotated by a single board-certified clinician. As elsewhere, we report means with 95% t-based confidence intervals and, per dimension, the proportion of error-free images (a score of 3) with 95% Wilson confidence intervals.

On this set, GPT-5.5 achieved a mean Pathology score of 2.15 (95% CI [1.91, 2.39]), with 53.8% (95% CI [0.43, 0.64]) of images free of pathology errors; a mean Realism score of 2.12 (95% CI [1.94, 2.31]), with 37.5% (95% CI [0.28, 0.49]) error-free; and a mean Composition score of 2.90 (95% CI [2.82, 2.98]), with 92.5% (95% CI [0.85, 0.97]) error-free. Relative to Gemini-3 graded on the same 80 conditions (Table 8), pathology accuracy was comparable (2.15 vs. 2.16) and composition was similar (2.90 vs. 2.75), but GPT-5.5 was markedly less photorealistic (2.12 vs. 2.78). Qualitatively, GPT-5.5 more often produced caricatured or otherwise unrealistic presentations, as seen in Figure 7.

Table 8: GPT-5.5 mean scores on the 80 "Pathology" category images, each scored 0–3 by a single clinician, with 95% t-based confidence intervals. Gemini-3 scores on the same 80-condition "Pathology" category are shown for reference.

Dimension	GPT-5.5 (1 clinician)	Gemini-3
Pathology	2.15 [1.91, 2.39]	2.16 [1.90, 2.42]
Realism	2.12 [1.94, 2.31]	2.78 [2.66, 2.90]
Composition	2.90 [2.82, 2.98]	2.75 [2.64, 2.86]

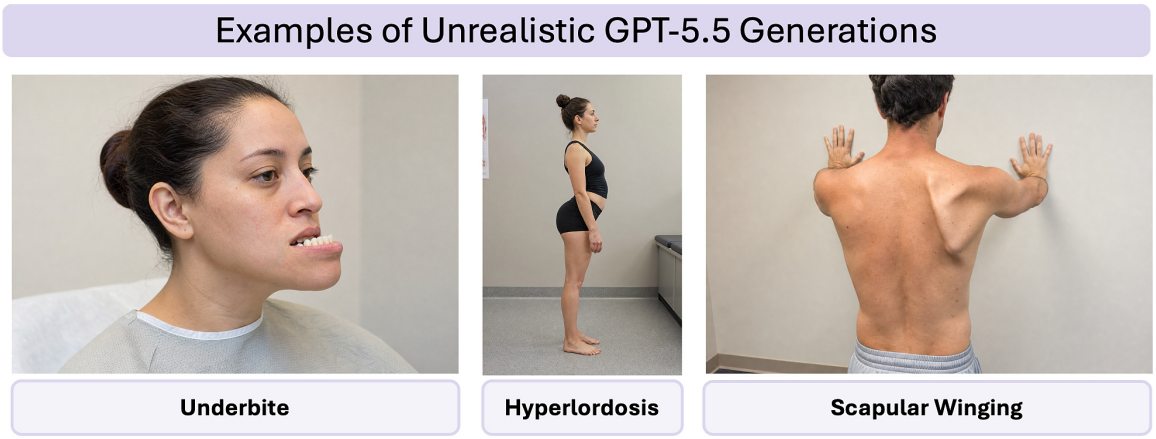


Figure 7: Examples of unrealistic GPT-5.5 generations.

Appendix E. Gender Distribution and Score Analysis

To further characterize demographic patterns in the generated patient avatars, a clinician annotator manually classified the apparent gender of each image. We identified 75 images depicting female patients, 145 images depicting male patients, and 140 images where apparent gender was unclear, such as images focused only on hands. This distribution suggests that future work should explicitly prompt image-generation models to provide more balanced representation across gender.

Despite this imbalance, quality and clinical accuracy were broadly similar between images depicting female and male patients. For example, average Pathology scores were 1.99 for images depicting female patients and 1.98 for images depicting male patients. Table 9 reports metric-level scores by apparent gender.

Table 9: Average scores by apparent gender. Images were manually classified as depicting female patients, male patients, or unclear apparent gender. The table reports total averages and counts, as well as averages and counts for images classified as female or male.

Metric	Total Avg.	Total N	Female Avg.	Female N	Male Avg.	Male N
Pathology	1.96	360	1.99	75	1.98	145
Realism	2.76	360	2.77	75	2.78	145
Composition	2.73	360	2.71	75	2.68	145
Counting	1.46	40	0.00	3	1.50	6
Pose	2.73	40	2.77	10	2.79	17
Consistency	2.46	80	2.73	20	2.64	33
Presentation	2.14	80	2.50	15	2.38	29
Viewpoint	2.91	40	3.00	10	2.89	16

Appendix F. Multi-Frame Scoring Sensitivity

As noted in the Methods, some single-frame prompts unexpectedly produced multi-frame images, for which we recorded the score of the highest-scoring frame. To assess the sensitivity of our findings to this choice, we annotated the lowest-scoring frame of these unwanted multi-frame images and recomputed our results for the "Pathology" and "Pathology+Presentation" categories, now including the second clinician's scores from the inter-rater analysis. As shown in Table 10, using the lowest-scoring frame instead of the highest-scoring frame had little impact on our findings.

Table 10: Mean scores using the highest-scoring versus the lowest-scoring frame of unwanted multi-frame images, for the "Pathology" and "Pathology+Presentation" categories. Scores include both clinicians' annotations where available.

	Pathology		Pathology+Presentation	
Metric	Highest-Scoring	Lowest-Scoring	Highest-Scoring	Lowest-Scoring
Pathology	2.156	2.081	1.894	1.781
Realism	2.781	2.781	2.825	2.812
Composition	2.750	2.725	2.806	2.781
Presentation	—	—	2.144	2.100

The remaining unwanted multi-frame images appeared in response to "Pathology+Pathology" prompts, which were prone to depicting the two pathologies in two different frames. In these cases, both pathologies were graded even when shown in different frames, and the Composition score was instead docked for failing to depict both pathologies in a single frame.