

Article

Drawing the Surgical Blueprint: Evaluating ChatGPT Versus Gemini Across Diverse Plastic Aesthetic Procedures

Gianluca Marcaccini ¹, Ishith Seth ^{2,*}, Omar Shadid ², Roberto Cuomo ¹ and Warren M. Rozen ²¹ Plastic and Reconstructive Surgery, Department of Medicine, Surgery and Neuroscience, University of Siena, 53100 Siena, Italy; gianlu32@gmail.com (G.M.)² Department of Plastic and Reconstructive Surgery, Peninsula Health, 2 Hastings Road, Melbourne, VIC 3199, Australia

* Correspondence: iseth@phcn.vic.gov.au

Abstract

Preoperative facial markings are critical to surgical precision and aesthetic outcomes in plastic surgery, yet remain operator-dependent and variably documented. Generative artificial intelligence (AI), particularly large multimodal models, offers potential for the automated illustration of surgical plans. This study compares the performances of ChatGPT-4o and Gemini Advanced in generating standardised preoperative markings for aesthetic facial procedures. Methods: Six text prompts describing common facial aesthetic surgeries were developed using established marking protocols. Each prompt was submitted once to ChatGPT-4o and Gemini Advanced, yielding twelve illustrations. Three board-certified plastic surgeons independently evaluated the images using a five-domain Likert scale assessing incision clarity, anatomical accuracy, template conformity, clinical usefulness, and overall graphic quality. A composite score out of 25 was calculated. Data were analysed using paired *t*-tests, and interrater reliability was assessed with intraclass correlation coefficients. Results: ChatGPT-4o significantly outperformed Gemini Advanced in composite scores (mean 18.0 ± 1.4 vs. 13.9 ± 1.6 , $p = 0.001$, Cohen's $d = 1.69$). Superior performance was noted across all domains, particularly in clarity (mean difference 0.83, $p = 0.002$) and graphic quality (mean difference 0.90, $p = 0.001$). Interrater reliability was good (ICC = 0.82). Discussion: ChatGPT-4o demonstrated higher fidelity in translating surgical prompts into anatomically appropriate, clinically useful illustrations. However, neither system achieved the precision required for clinical implementation without revision. These models may serve as adjuncts in education and preliminary planning. Future work should explore model fine-tuning, surgeon-guided generation, and performance in reconstructive procedures.

Keywords: artificial intelligence; facial aesthetic surgery; preoperative markings; ChatGPT; multimodal large language models



Academic Editor: Andrew A. Gumbs

Received: 26 May 2025

Revised: 17 June 2025

Accepted: 25 June 2025

Published: 27 June 2025

Citation: Marcaccini, G.; Seth, I.; Shadid, O.; Cuomo, R.; Rozen, W.M. Drawing the Surgical Blueprint: Evaluating ChatGPT Versus Gemini Across Diverse Plastic Aesthetic Procedures. *Surgeries* **2025**, *6*, 51. <https://doi.org/10.3390/surgeries6030051>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Precise cutaneous markings remain critical to safety and aesthetic quality in facial plastic surgery. By translating a three-dimensional assessment of the patient's anatomy into two-dimensional guides drawn directly onto the skin, these markings dictate the exact placement and orientation of incisions, the extent of soft-tissue excision, and the vectors along which deeper layers will be advanced. Millimetric inaccuracies have been associated with prolonged operative times, higher rates of scar malposition, ectropion, asymmetry,

and patient dissatisfaction [1]. Additionally, markings serve a medicolegal function by documenting the operative plan discussed during informed consent [2].

Facial markings are still executed freehand with felt-tip pens despite their central role. The process is operator-dependent, varies with experience, and is often documented only through intraoperative photographs of inconsistent quality, which limits remote consultation, resident feedback, and longitudinal analysis of surgical learning curves [3]. Digital planning tools have become a routine part of skeletal cranio-maxillofacial surgery. Yet, comparable software for soft-tissue facial procedures is scarce, expensive, and rarely integrated into day-to-day clinical workflows [4,5].

Generative artificial intelligence offers a potential alternative. Modern text-to-image diffusion engines convert detailed natural language prompts into high-resolution illustrations in seconds. Early descriptive studies have employed these systems for synthetic clinical photographs, patient education, and teaching diagrams. However, systematic assessments reveal proportional errors and occasional hallucinations of additional anatomical features, highlighting the need for domain-specific validation [6,7].

Multimodal large language model platforms such as ChatGPT-4o and Gemini Advanced integrate diffusion engines that can be guided by highly structured prompts. These models possess extensive medical lexical knowledge, making them uniquely suited for generating anatomically accurate, publication-quality illustrations of facial preoperative markings. Preliminary reports suggest that large language models outperform junior residents in written knowledge assessments and exhibit differences in spatial reasoning tasks [8,9]. However, no peer-reviewed study has objectively compared their ability to translate surgical prompts into clinically acceptable marking diagrams for aesthetic facial procedures.

This study aims to evaluate and compare the capacity of ChatGPT-4o and Gemini Advanced to reproduce accurate standard facial preoperative markings, using expert surgeon ratings as the reference standard. The results aim to assist clinicians needing quick visual aids, educators designing competency-based curricula, and developers seeking to improve the medical reliability of foundational generative models.

2. Materials and Methods

This prospective *in silico* comparison was conducted in accordance with the Declaration of Helsinki. Only synthetic illustrations were produced; no real patient data or identifiers were involved. Therefore, institutional review board approval and informed consent were not required.

AI platforms: The study compared ChatGPT-4o (OpenAI, built 24 April 2025) with Gemini Advanced (Google DeepMind, released 16 April 2025). Both systems were accessed through their standard web interfaces during the same forty-eight-hour period to minimise version drift. No plug-ins, extensions, or manual postprocessing were applied.

Prompt preparation: The authors wrote all of the prompts based on established facial surgery marking protocols. The prompts were written by two board-certified plastic surgeons and reviewed by the author team to ensure alignment with standard marking practices. Each text specifies orientation, incision course, excision outline, and lift vectors, replicating the information typically drawn on the patient's skin.

1 Upper blepharoplasty

“Draw a preoperative surgical marking for upper blepharoplasty. Show a front view of a human face with clear surgical markings: upper eyelid crease, planned incision line along the natural fold, lateral extension (crow's-feet area), and areas of skin excision. The face should have a neutral expression, light skin tone, and open eyes. Use a realistic anatomical medical illustration style on a white background.”

2 Endoscopic forehead (brow) lift

“Draw a preoperative surgical marking for an endoscopic forehead lift (brow lift). Show a front view of a human face with clear markings: small incision lines hidden within the hairline in parietal and temporal areas, arrows indicating the vector of lift (upward and lateral), and areas of forehead and brow tissue mobilisation. Include outlines of the brow arches. The face should have a neutral expression, light skin tone, and hair pulled back. Use a realistic anatomical medical illustration style on a white background.”

3 Temporal (lateral brow) lift

“Draw the preoperative surgical markings for a temporal lift (lateral brow lift). Show a front view of a human face with clear markings: incision lines hidden within the temporal scalp just behind the lateral hairline. Indicate oblique upward and lateral lift vectors with arrows, focusing on the eyebrow’s tail, and highlight the mobilisation zone in the temporal region. The face should have a neutral expression, light skin tone, and hair pulled back. Use a realistic anatomical medical illustration style on a white background.”

4 Direct brow lift

“Draw the preoperative surgical markings for a direct brow lift. Show a front view of a human face with clear markings: elliptical excision lines immediately above the brow arches, following the upper edge of each eyebrow. Highlight planned skin excision and indicate the upward lift direction with arrows. The face should have a neutral expression, light skin tone, and clearly defined eyebrows. Use a realistic anatomical medical illustration style on a white background.”

5 Sub-nasal (bull-horn) lip lift

“Draw the preoperative surgical marking for a subnasal lip lift (bullhorn technique). Show a front view of a human face with clear markings: a curved excision outline beneath the base of the nose, along the columella and alar bases, resembling a bull-horn shape. Indicate the upward vector of lift on the upper lip with arrows. The face should have a neutral expression, light skin tone, and visible lips. Use a realistic anatomical medical illustration style on a white background.”

6 Deep-plane face-and-neck lift

“Draw the preoperative surgical markings for a deep plane face and neck lift. Show a front view of a human face with detailed markings: preauricular and retroauricular incision lines extending into the temporal scalp. Indicate the deep dissection zones, especially the masseter muscle’s midface and anterior border. Use arrows to show the vectors of lift in a superolateral direction for the cheek, jawline, and neck. Include anatomical landmarks such as the zygomatic arch, mandibular angle, and sternocleidomastoid border. The face should have a neutral expression, light skin tone, and hair pulled back. Use a realistic anatomical medical illustration style on a white background.”

Each prompt was submitted once to ChatGPT-4o and once to Gemini Advanced, resulting in twelve PNG images saved exactly as the platforms delivered them.

Expert Evaluation: Three board-certified plastic surgeons with experience ranging from eleven to eighteen years independently reviewed each illustration. Utilising a five-point Likert scale, they assessed the clarity of the incision lines, anatomical accuracy, conformity to recognised marking templates, perceived clinical usefulness, and overall graphic quality. Summation of the five criteria (range 5–25) constituted the composite score. The reviewers were blinded to the model’s identity and operated independently.

Outcome Measures: The primary endpoint was the difference in composite scores between ChatGPT-4o and Gemini Advanced. Secondary endpoints encompassed domain-specific score differences and interrater reliability.

Statistical Analysis: The scores were entered into Microsoft Excel and analysed using Python 3.11 (pandas, SciPy). The normality of paired differences was assessed with the Shapiro–Wilk test. Normally distributed data were compared using paired two-tailed Student’s *t*-tests, while non-normally distributed data were compared using Wilcoxon signed-rank tests. Statistical significance was established at $p < 0.05$. Effect sizes were expressed as Cohen’s *d* for parametric data and rank-biserial correlation for non-parametric data. Interrater reliability was calculated using the two-way random intraclass correlation coefficient ICC (2, 1) and interpreted according to conventional thresholds. Data Availability: All AI-generated images, anonymised scoring sheets, and analysis scripts are stored in the institutional repository and can be requested from the corresponding author upon reasonable inquiry.

3. Results

The three reviewers assessed the AI-generated diagrams once each from ChatGPT-4o and Gemini Advanced for the six procedures: upper blepharoplasty (Figure 1), endoscopic forehead (Figure 2), temporal lift (Figure 3), direct brow lift (Figure 4), sub-nasal lip lift (Figure 5), and deep-plane face-and-neck lift (Figure 6).

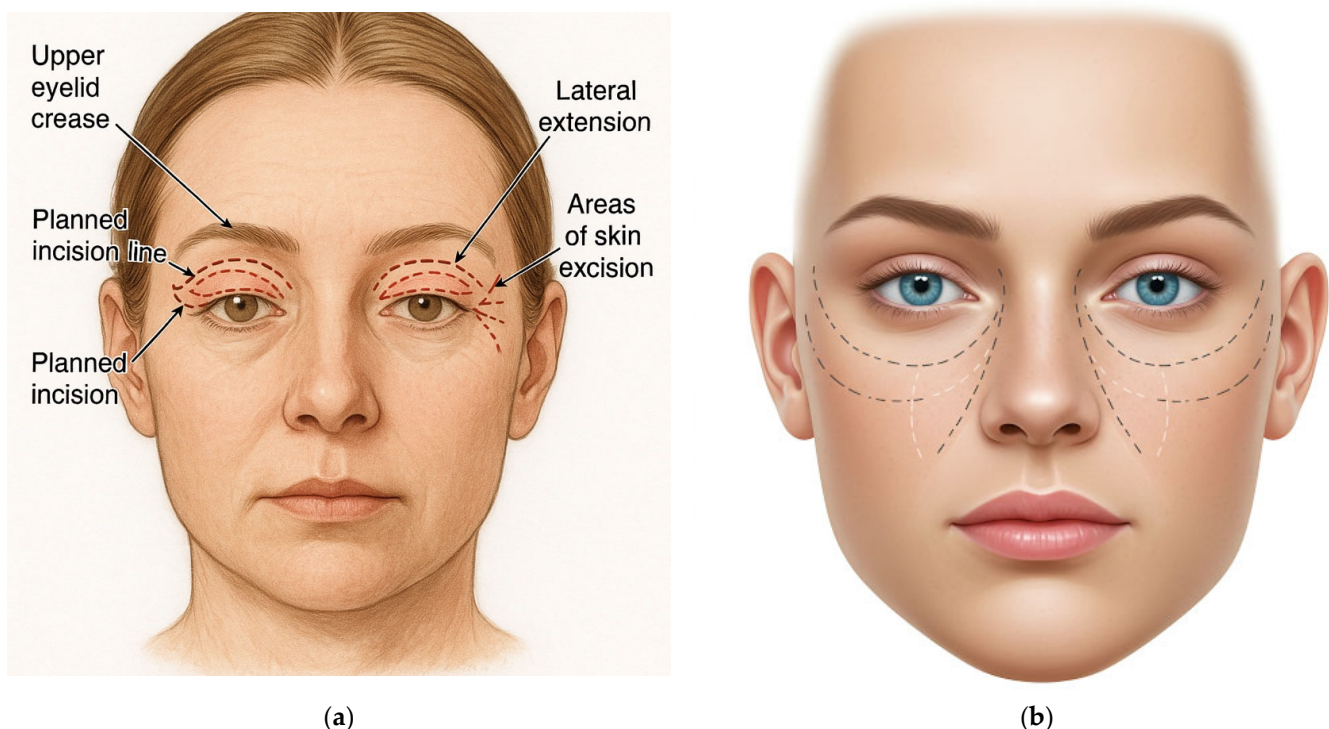
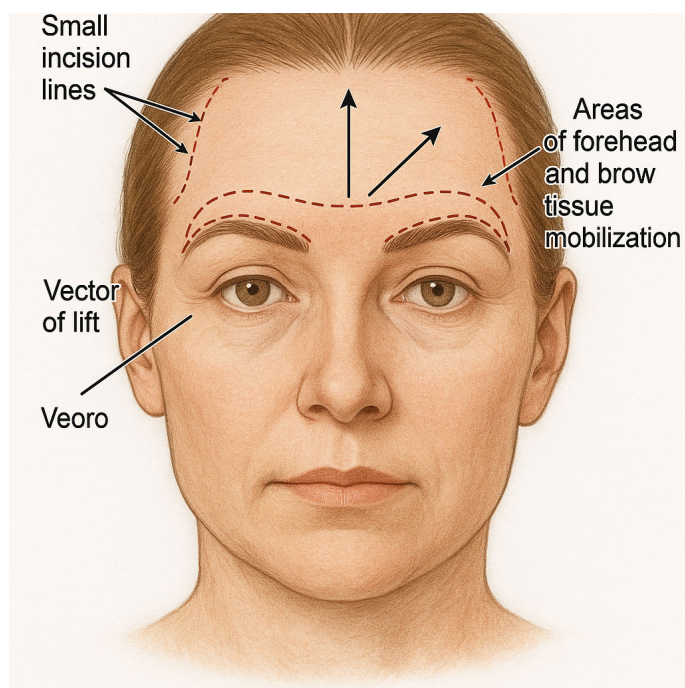
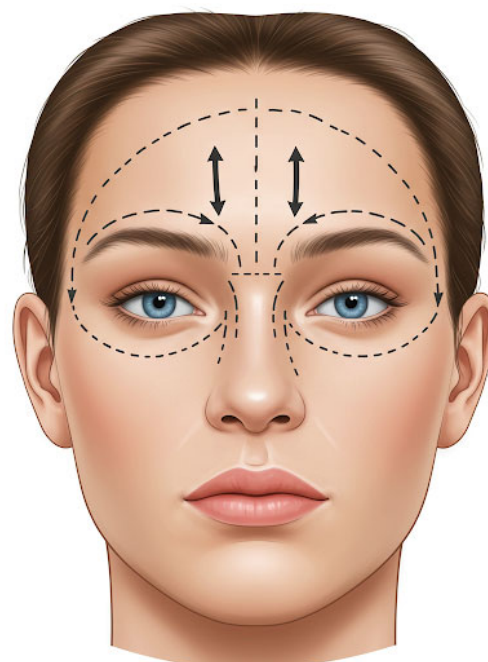


Figure 1. AI markings for upper blepharoplasty. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.

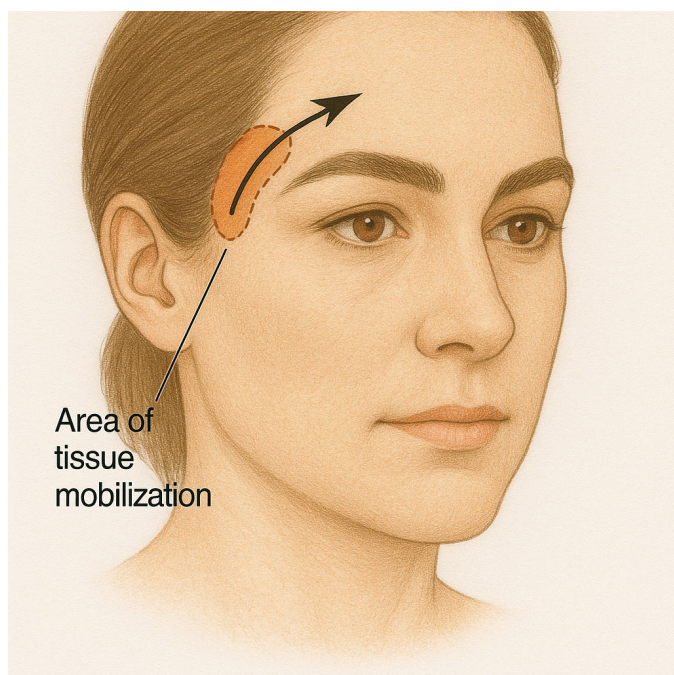


(a)

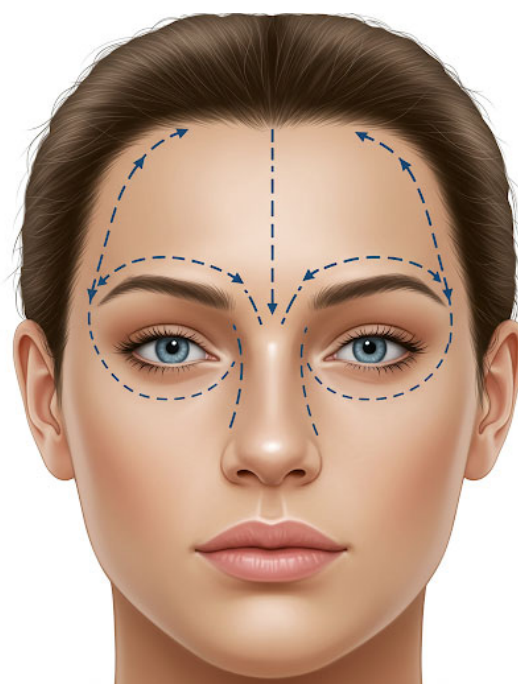


(b)

Figure 2. AI markings for endoscopic forehead lift. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.

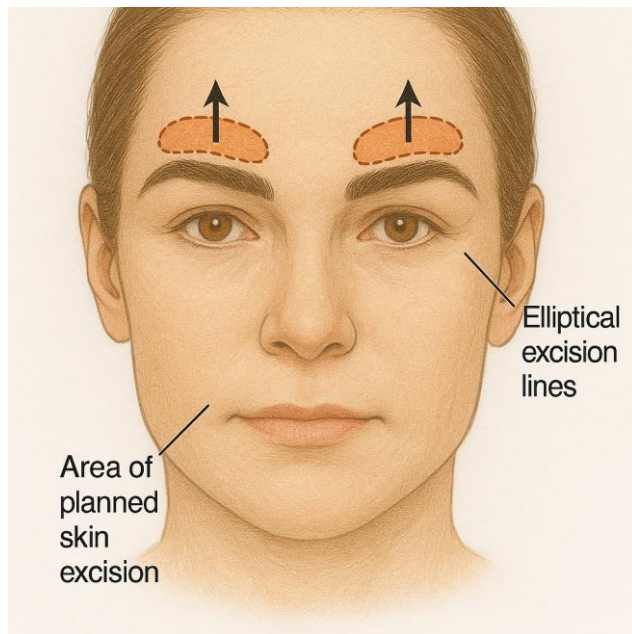


(a)

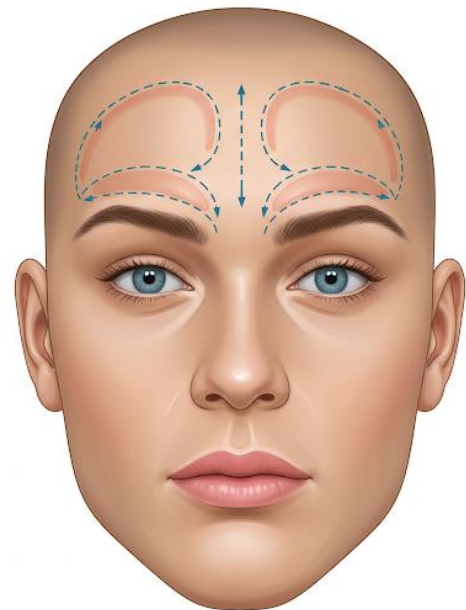


(b)

Figure 3. AI markings for temporal (lateral brow) lift. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.

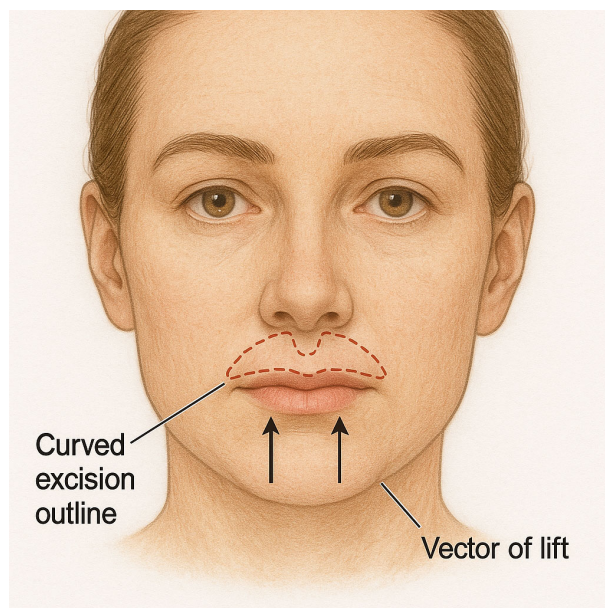


(a)

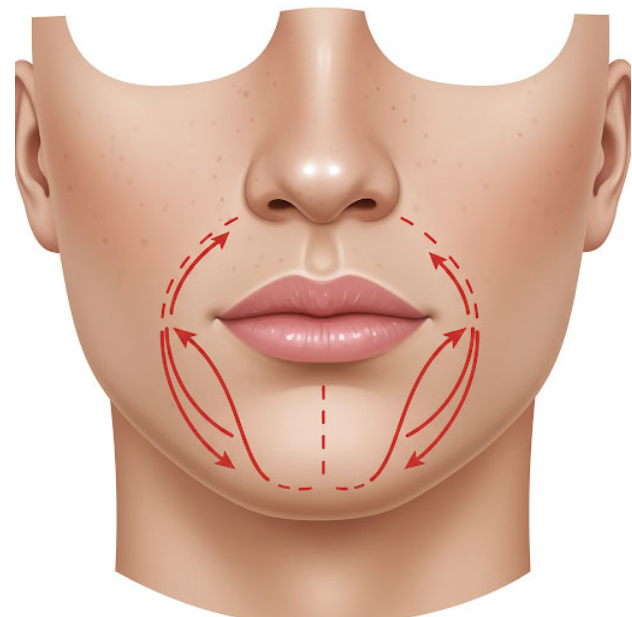


(b)

Figure 4. AI markings for direct brow lift. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.



(a)



(b)

Figure 5. AI markings for sub-nasal (bull horn) lift. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.

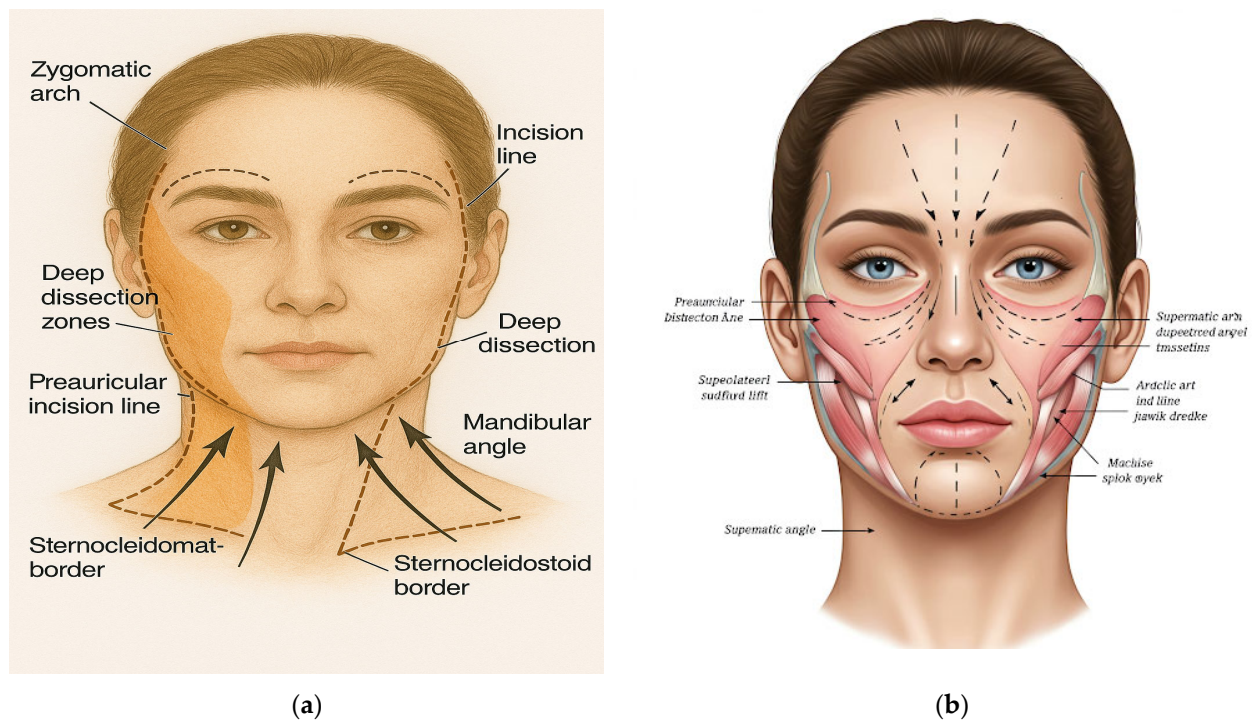


Figure 6. AI markings for deep-plane face-and-neck lift. (a) AI markings by ChatGPT-4o; (b) AI markings by Gemini Advanced.

The results from the paired differences met the Shapiro–Wilk normality test ($p > 0.05$), enabling parametric statistics. The composite score, presented on a 25-point scale, averaged 18.0 ± 1.4 for the images produced with ChatGPT-4o and 13.9 ± 1.6 for those created with Gemini Advanced; the mean paired difference was 4.1 points, resulting in $p = 0.001$ and a Cohen d of 1.69 (Table 1).

Table 1. Comparative evaluation of ChatGPT-4o and Gemini Advanced across five expert-rated domains and their composite scores.

Domain (5-pt Likert)	ChatGPT-4o Mean $\pm$ SD	Gemini Adv. Mean $\pm$ SD	Mean Paired Diff.	p-Value	Cohen D
Clarity of incision lines	4.1 ± 0.4	3.3 ± 0.5	0.83	0.002	1.30
Anatomical accuracy	3.5 ± 0.6	2.7 ± 0.7	0.83	0.004	1.12
Conformity to templates	3.3 ± 0.5	2.5 ± 0.6	0.78	0.006	1.07
Clinical usefulness	3.4 ± 0.6	2.6 ± 0.6	0.83	0.005	1.21
Overall graphic quality	3.7 ± 0.4	2.8 ± 0.5	0.90	0.001	1.52
Composite score (5–25)	18.0 ± 1.4	13.9 ± 1.6	4.08	0.001	1.69

Analysis of individual domains revealed consistent findings. The clarity of incision lines had means of 4.1 ± 0.4 for ChatGPT-4o and 3.3 ± 0.5 for Gemini Advanced, with a paired difference of 0.83 points ($p = 0.002$, $d = 1.30$). Anatomical accuracy showed mean values of 3.5 ± 0.6 and 2.7 ± 0.7 , respectively; the 0.83-point difference achieved $p = 0.004$ and $d = 1.12$. Conformity to recognised marking templates averaged 3.3 ± 0.5 compared to 2.5 ± 0.6 , yielding a 0.78-point gap ($p = 0.006$, $d = 1.07$). Perceived clinical usefulness recorded means of 3.4 ± 0.6 for ChatGPT-4o and 2.6 ± 0.6 for Gemini, with the 0.83-point difference correlating to $p = 0.005$ and $d = 1.21$. Overall graphic quality averaged 3.7 ± 0.4 against 2.8 ± 0.5 ; the 0.90-point difference indicated $p = 0.001$ and $d = 1.52$. To improve statistical transparency, 95% confidence intervals for all domain-specific mean differences

were calculated and are reported in Table 2. Due to the small sample size, confidence intervals were relatively wide and should be interpreted with caution.

Table 2. The 95% confidence intervals for the mean paired differences across the evaluation domains.

Domain	Mean Paired Diff.	95% CI (Lower)	95% CI (Upper)
Clarity of incision lines	0.83	−0.81	2.47
Anatomical accuracy	0.83	−1.07	2.73
Conformity to templates	0.78	−1.09	2.65
Clinical usefulness	0.83	−0.93	2.59
Overall graphic quality	0.90	−0.62	2.42
Composite score (5–25)	4.08	−2.13	10.29

Interrater reliability for the composite score was good, with an intraclass correlation coefficient ICC (2,1) of 0.82 and a 95% confidence interval from 0.71 to 0.91. Domain-specific ICC values ranged between 0.74 for conformity and 0.86 for clarity, demonstrating consistent agreement among the three reviewers.

4. Discussion

This study compared two large multimodal language model platforms for the narrowly defined task of generating preoperative facial markings based on standardised textual prompts. The primary finding indicates that ChatGPT-4o attained higher mean scores than Gemini Advanced across all five expert-rated domains and secured a composite score four points greater on a twenty-five-point scale. Given that the paired differences were substantial in both statistical and practical terms, the results position ChatGPT-4o as the prevailing leader among general-purpose generative systems for this specific surgical illustration.

The initial significant outcome pertains to the clarity of incision lines. ChatGPT-4o produced continuous, well-contrasted strokes that adhered to the requested trajectories, whereas Gemini Advanced frequently exhibited interruptions or deviations. Continuous incision paths are essential in surgical markings as they delineate dissection planes and provide guidance during intraoperative orientation. Previous studies assessing AI-generated infographics for patient education reported comparable advantages of ChatGPT-based images in preserving line coherence [10,11]. In this context, the enhanced integrity of the lines correlated with elevated ratings of clinical usefulness, indicating that the text encoder within ChatGPT-4o effectively conveys spatial instructions to its diffusion engine with greater precision than its competitor.

The second major observation pertains to anatomical precision and adherence to established templates. The reviewers indicated that ChatGPT-4o generally maintained key facial proportions, including upper-lid crease height, brow apex location, and auricular landmarks. At the same time, Gemini Advanced more frequently misaligned these structures and occasionally directed lift arrows toward unrealistic orientations. Studies on AI anatomical illustration have underscored the persistent issues of proportional drift and landmark omission as consistent shortcomings of earlier generative models [11–13]. The current data demonstrate measurable progress; however, they also confirm that even the top-performing system still does not meet specialists' expectations, particularly when representing subtle three-dimensional considerations, such as the transition between tarsal and pretarsal skin.

A third finding concerns overall graphic quality. Both platforms produced images deemed satisfactory for teaching or counselling; however, none were considered accurate enough for direct transfer onto a patient's skin without revision. This conclusion aligns

with recent comparative studies in hand fracture management and rehabilitation plan generation, where ChatGPT-4o provided the most clinically relevant content but still required human oversight before implementation [14]. Together, these results reinforce the growing consensus that contemporary LLMs can enhance visual communication but do not match the precision of specialty-trained medical illustrators.

Several limitations temper the interpretation of our findings. The prompt library was confined to six facial aesthetic procedures, and sensitive anatomical regions such as breasts and genital areas were intentionally excluded to adhere to current generative AI policies. This scope limits the generalizability of our findings to other surgical subspecialties, such as reconstructive procedures, body contouring, or orthoplastic markings. Performance may vary significantly across anatomical regions and use cases. Future research should explore a broader range of procedures to assess whether the observed trends hold beyond facial aesthetics. Consequently, the conclusions cannot be generalised to full-body markings or procedures involving areas the models are not authorised to depict. Each prompt was submitted only once to each platform; in clinical practice, surgeons may request multiple generations and manually select the most suitable frame, a workflow that might narrow the observed performance gap. This one-shot design was intentional, aiming to assess each model's baseline performance without human intervention. While clinical users may iterate, our goal was to test whether a single, structured prompt could yield a usable result without manual optimisation or selection bias. Evaluation relied on a subjective five-point Likert scale, although interrater reliability was satisfactory. While our use of a five-point Likert scale provided structured expert evaluation, it remains inherently subjective. Future studies should incorporate objective metrics, such as geometric overlap analysis, landmark accuracy, or deviation from validated surgical templates, to quantify performance more precisely and reduce evaluator bias. We also acknowledge the absence of expert-drawn reference markings. Direct comparison with surgeon-generated diagrams could better contextualise AI performance and should be included in future validation studies. Lastly, both platforms are evolving rapidly; the 48 h testing window minimised version drift but does not capture performance consistency over time. Future evaluations should consider repeated assessments across longer periods to account for model updates and discontinuities. Finally, both ChatGPT-4o and Gemini Advanced are general-purpose models not specifically trained on medical or surgical datasets. Their training corpora are not publicly disclosed and likely lack detailed, validated anatomical illustrations. This limitation may explain several of the inaccuracies observed in vector orientation and landmark placement. Achieving clinical-grade precision will likely require purpose-built generative models fine-tuned on curated surgical datasets under expert supervision.

Future research should address these limitations. A more extensive prompt set encompassing body-contouring or reconstructive markings, multiple generations per prompt, and objective alignment analysis would provide a more comprehensive performance landscape. Fine-tuning diffusion engines on curated surgical illustration datasets may enhance landmark placement and arrow orientation. Interactive surgeon-in-the-loop interfaces, where the user anchors critical points before final rendering, could integrate the speed of generative AI with expert precision. Prospective studies that evaluate operative time, resident comprehension, or the quality of patient consent when AI-generated drawings supplement traditional methods would elucidate real-world utility. Beyond validation, future work should also address clinical integration pathways. Regulatory frameworks must clarify the use of AI-generated surgical planning tools, particularly regarding liability in case of misrepresentation or error. Implementation into surgical workflows may require surgeon-in-the-loop systems that allow expert oversight and final approval. Educational platforms could integrate these models for interactive training modules, while preoperative

consultations may benefit from visual aids that enhance patient understanding. Such applications would require not only technical refinement but also ethical, legal, and institutional alignment [15,16]. Cost-effectiveness is another important consideration. AI-generated illustrations, while not yet clinically sufficient, may offer a scalable and low-cost alternative to professional medical illustration services, particularly in educational and preliminary planning contexts. Future studies should assess whether their use reduces time or resource demands in preoperative workflows [15,16]. Given the rapid pace of model evolution, periodic reassessment will be indispensable in capturing capability gains that may arise within short time frames.

5. Conclusions

ChatGPT-4o generated facial preoperative markings that were clearer, more anatomically accurate, and more compliant with standard templates than those produced by Gemini Advanced. However, both systems demonstrated inaccuracies in vector direction and subtle landmark placement; therefore, neither meets the standard required for unsupervised clinical application. These technologies should thus be regarded as promising adjuncts for education and preliminary visualisation in facial aesthetic surgery, while their role in other domains remains to be established. Continuous advancements in multimodal architecture, targeted fine-tuning, and regulatory guidance are expected to reduce the disparity between AI output and surgical requirements. Future investigations should monitor this progression while upholding rigorous validation protocols.

Author Contributions: Conceptualisation, I.S. and W.M.R.; methodology, O.S. and I.S.; software, O.S.; validation, I.S., W.M.R. and R.C.; formal analysis, I.S.; investigation, O.S.; resources, W.M.R.; data curation, O.S.; writing—original draft preparation, I.S. and O.S.; writing—review and editing, I.S., R.C. and G.M.; visualisation, O.S.; supervision, W.M.R. and G.M.; project administration, I.S.; funding acquisition, W.M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ICC	Intraclass Correlation Coefficient
IRB	Institutional Review Board
PNG	Portable Network Graphics
SD	Standard Deviation
3D	Three-Dimensional
2D	Two-Dimensional

References

1. Borille, G.B.; Pereira Filho, G.A.; Zancanaro, M.; Ribeiro, V.W.; Giannini, R.; Carvalho, L. Optimizing the Incision Marking in Abdominoplasty: 2-year Series. *Plast. Reconstr. Surg.-Glob. Open* **2025**, *13*, e6640. [[CrossRef](#)] [[PubMed](#)]
2. Ellsworth, W.A.; Basu, C.B.; Iverson, R.E. Perioperative considerations for patient safety during cosmetic surgery—Preventing complications. *Can. J. Plast. Surg.* **2009**, *17*, 9–16. [[CrossRef](#)] [[PubMed](#)]

3. Schreiber, J.E.; Stern, C.S.; Garfein, E.S.; Weichman, K.E.; Tepper, O.M. A Novel Approach to Surgical Markings Based on a Topographic Map and a Projected Three-Dimensional Image. *Plast. Reconstr. Surg.* **2016**, *137*, 855e–859e. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Myers, P.L.; Nelson, J.A.; Rosen, E.B.; Allen, R.J.J.; Disa, J.J.; Matros, E.M. Virtual Surgical Planning for Oncologic Mandibular and Maxillary Reconstruction. *Plast. Reconstr. Surg.-Glob. Open* **2021**, *9*, e3672. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Borisenko, A.S.; Sharobaro, V.I.; Avdeev, A.E.; Burkhonova, N.S.; Fisun, A.O. Using 3D Imaging to Preoperatively Plan Facelift Procedures for the Lower Third of the Face and Neck. *Cosmetics* **2024**, *11*, 28. [\[CrossRef\]](#)
6. Kumar, A.; Burr, P.; Young, T.M. Using AI Text-to-Image Generation to Create Novel Illustrations for Medical Education: Current Limitations as Illustrated by Hypothyroidism and Horner Syndrome. *JMIR Med. Educ.* **2024**, *10*, e52155. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Muhr, P.; Pan, Y.; Tumescheit, C.; Kübler, A.-K.; Parmaksiz, H.K.; Chen, C.; Orozco, P.S.B.; Lienkamp, S.S.; Hastings, J. Evaluating Text-to-Image Generated Photorealistic Images of Human Anatomy. *Cureus* **2024**, *16*, e74193. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Atkinson, C.J.; Seth, I.; Xie, Y.; Ross, R.J.; Hunter-Smith, D.J.; Rozen, W.M.; Cuomo, R. Artificial Intelligence Language Model Performance for Rapid Intraoperative Queries in Plastic Surgery: ChatGPT and the Deep Inferior Epigastric Perforator Flap. *J. Clin. Med.* **2024**, *13*, 900. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Hubany, S.S.; Scala, F.D.; Hashemi, K.; Kapoor, S.B.; Fedorova, J.R.B.; Vaccaro, M.J.B.; Ridout, R.P.B.; Hedman, C.C.B.; Kellogg, B.C.; Barone, A.A.L. ChatGPT-4 Surpasses Residents: A Study of Artificial Intelligence Competency in Plastic Surgery In-service Examinations and Its Advancements from ChatGPT-3.5. *Plast. Reconstr. Surg.-Glob. Open* **2024**, *12*, e6136. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Temsah, M.-H.; Alhuzaimi, A.N.; Almansour, M.; Aljamaan, F.; Alhasan, K.; Batarfi, M.A.; Altamimi, I.; Alharbi, A.; Alsuhaibani, A.A.; Alwakeel, L.; et al. Art or Artifact: Evaluating the Accuracy, Appeal, and Educational Value of AI-Generated Imagery in DALL-E 3 for Illustrating Congenital Heart Diseases. *J. Med. Syst.* **2024**, *48*, 54. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Noel, G.P.J.C. Evaluating AI-powered text-to-image generators for anatomical illustration: A comparative study. *Anat. Sci. Educ.* **2024**, *17*, 979–983. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Haider, S.A.; Prabha, S.; Gomez-Cabello, C.A.; Borna, S.; Pressman, S.M.; Genovese, A.; Trabilsy, M.; Galvao, A.; Aziz, K.T.; Murray, P.M.; et al. A Validity Analysis of Text-to-Image Generative Artificial Intelligence Models for Craniofacial Anatomy Illustration. *J. Clin. Med.* **2025**, *14*, 2136. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Borji, A. Qualitative failures of image generation models and their application in detecting deepfakes. *Image Vis. Comput.* **2023**, *137*, 104771. [\[CrossRef\]](#)
14. Atkinson, C.J.; Seth, I.; Seifman, M.A.; Rozen, W.M.; Cuomo, R. Enhancing Hand Fracture Care: A Prospective Study of Artificial Intelligence Application With ChatGPT. *J. Hand Surg. Glob. Online* **2024**, *6*, 524–528. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Marcaccini, G.; Seth, I.; Novo, J.; McClure, V.; Sacks, B.; Lim, K.; Ng, S.K.-H.; Cuomo, R.; Rozen, W.M. Leveraging Artificial Intelligence for Personalized Rehabilitation Programs for Head and Neck Surgery Patients. *Technologies* **2025**, *13*, 142. [\[CrossRef\]](#)
16. Marcaccini, G.; Seth, I.; Xie, Y.; Susini, P.; Pozzi, M.; Cuomo, R.; Rozen, W.M. Breaking Bones, Breaking Barriers: ChatGPT, DeepSeek, and Gemini in Hand Fracture Management. *J. Clin. Med.* **2025**, *14*, 1983. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.