

Background, Clothing, and Multimodal Face Ratings

Chicken Angel
Independent research project
chicken-angel.com

Abstract

We examined whether background and clothing edits change multimodal models' ratings of a face when its pixels remain identical. Two experiments used four photographs of one person and produced 116 responses from two models. In a three-photograph pilot, face ratings were unchanged in all 12 model–edit comparisons, while whole-image ratings changed in eight. A follow-up used a fourth photograph, eight edits, a resampling control, and two face-rating prompts. One model returned the same score throughout. The other changed by a mean absolute 0.69 points with the plain prompt and 0.56 points when explicitly told to ignore context. The resampling control produced changes of similar size, and repeated requests for the same image disagreed in 11 of 20 conditions. Constant ratings also accompanied inconsistent descriptions of visible facial features. A separate landmark detector shifted by up to 2.30 pixels RMS across the edited images, despite unchanged face pixels; fixing the surrounding pixels removed that variation. These results show variation in the image-to-rating process, but leave open how much comes from the meaning of the scene, image processing, or response variability. We provide the images, individual responses, and analysis code.

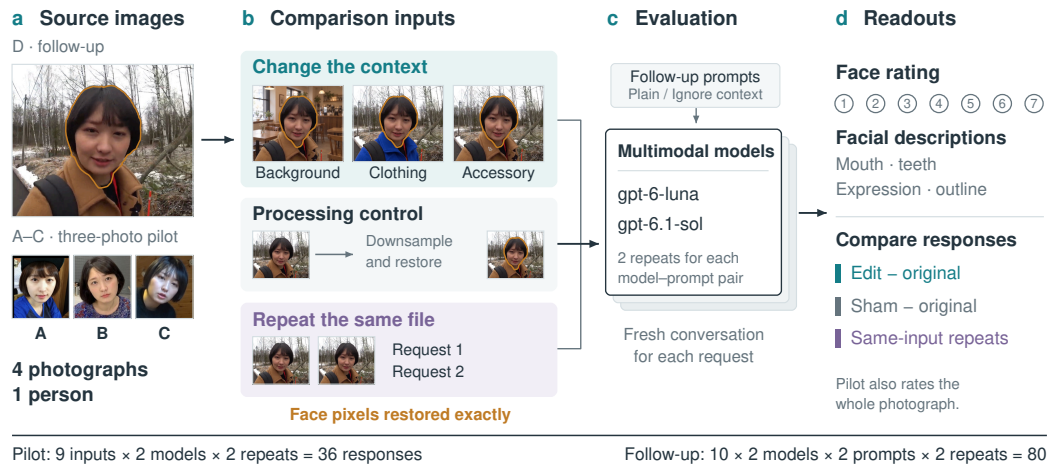


Figure 1: **Experimental design.** Photograph D illustrates three comparisons: context edits, a processing control, and repeated requests for an identical file. Source pixels inside the protection masks are restored after editing. Each model returns a face rating and facial descriptions. The pilot used photographs A–C and a joint face-and-image prompt; the sham and two face-only prompts were added in the follow-up. The A–C thumbnails are display crops.

1 Introduction

When a multimodal model rates a face, it also sees the person’s clothing, the surrounding scene, and the way the photograph was taken. Its explanation may refer to eyes, proportions, or expression, but that explanation alone tells us little about how the rest of the image affected the score. A direct way to examine this is to change the surroundings while leaving the face untouched.

We used four photographs of one person to study this question. In an initial experiment, two models rated both the face and the whole photograph after background and clothing edits. We then tested a fourth photograph with a wider set of edits, a resampling control, and two prompts that differed in how explicitly they excluded context. After generating each edit, we restored the original pixels within a stored face mask. The resulting files let us compare responses to visibly different scenes containing exactly the same facial image.

The first experiment produced no change in face ratings, even when whole-image ratings changed. In the follow-up, one model again returned constant scores, while the other varied across edits. That variation was difficult to attribute to the depicted scene or clothing: the resampling control also changed the score, and repeated requests for an identical file often disagreed. We examine these results alongside the models’ descriptions of the face and a separate landmark detector. Together, they show why a change in the reported face score needs to be read in the context of the editing procedure and the model’s own response variability.

2 Related work

Background replacement has been used to study contextual dependence in object recognition [Xiao et al., 2021]. More recent work on vision-language models distinguishes semantic distractions from degradation of the relevant visual evidence [Sun et al., 2026]. These approaches inform our use of controlled edits, although their tasks have correctness criteria that subjective face ratings lack.

StylisticBias is the closest related study: it varies clothing, eyewear, and other appearance cues across synthetic identities and measures several social judgments, including attractiveness [Kolli et al., 2026]. We use real photographs and preserve the facial RGB values and image framing exactly. FOCUS takes the complementary approach of editing faces within real photographic scenes, and discusses the residual artifacts introduced by editing [Chen et al., 2026]. In our experiment, those artifacts remain a concern outside the protected region.

The question being asked also matters. Promptception documents sensitivity to prompt wording and focus instructions [Ismithdeen et al., 2025], while AesBench distinguishes aesthetic assessment from perception and interpretation [Huang et al., 2024]. We therefore report face and whole-image ratings separately and test an explicit instruction to ignore context. A recent comparison with human ratings found higher, more compressed attractiveness scores from multimodal models [Grandas et al., 2026]. This makes coarse use of the rating scale a relevant alternative explanation for an unchanged score.

3 Methods

3.1 Photographs and experimental design

The four photographs came from an owner-provided collection and show one adult in different settings (Figure 1). The pilot used three indoor photographs, A–C. For each, we prepared an outdoor-background edit and a light-coloured clothing edit. A single prompt requested a face rating, a whole-image rating, and descriptions of visible facial and contextual features. Two models evaluated the nine inputs twice, giving 36 responses. Two preliminary connectivity checks are excluded from these counts.

The follow-up used a new outdoor photograph, D, with a resolution of 1569 by 1321 pixels. The project owner chose it before the follow-up evaluations. We made eight edits: three backgrounds (a neutral wall, a cafe, and a hotel lobby), three garments (a grey hoodie, a navy blazer, and a blue athletic jacket), and two accessories on the original coat (a silver pendant and a botanical brooch). We excluded glasses and earrings because they could cover facial or ear features. The original and a processing control completed the ten-input set. Table 1 summarises the two experiments.

Table 1: **Two experimental phases.** Four photographs of one person, 19 input files, and 116 retained responses. The prompts and controls differed between phases.

Design choice	Pilot: photographs A–C	Follow-up: photograph D
Inputs	3 originals + 6 edits	Original + sham + 8 edits
Edits	Background and clothing	Background, clothing and accessories
Rating target	Face and whole photograph together	Face only
Prompt versions	One joint prompt	Plain / ignore-context instruction
Controls	Same-file repeats	Same-file repeats + processing sham
Responses	36 (2 models, 2 repeats)	80 (2 models, 2 prompts, 2 repeats)

3.2 Image preparation

We generated candidate edits and composited them within specified background, garment, or accessory regions. For D, manually reviewed contours and OpenCV GrabCut defined the person boundary. Stored masks protected the face, hair, and neck. The pilot used its own stored face masks. The masks were fixed before evaluation.

For every D composite, we restored the source RGB values inside the protection masks and outside the permitted editing region. Generated candidates were first resized to the original canvas; the protected source pixels were copied without resizing. Feathering was restricted to the editing region. A local colour-based matte was used for the small silver accessories to retain the surrounding coat texture. We then decoded each exported PNG and checked the protected pixels for exact equality. The materials include the masks and SHA-256 hashes of the files and protected RGB values.

All six pilot edits were retained. The follow-up required nine generation calls for eight accepted edits: the first hotel candidate added a distant receptionist, which was removed in a second attempt. Mask refinements and candidate selection were completed before the model ratings were collected. We kept the rejected candidate and the generation record. Some boundary seams, original collar pixels, and texture differences remain visible at full resolution; Figure 2 shows where the final inputs differ.

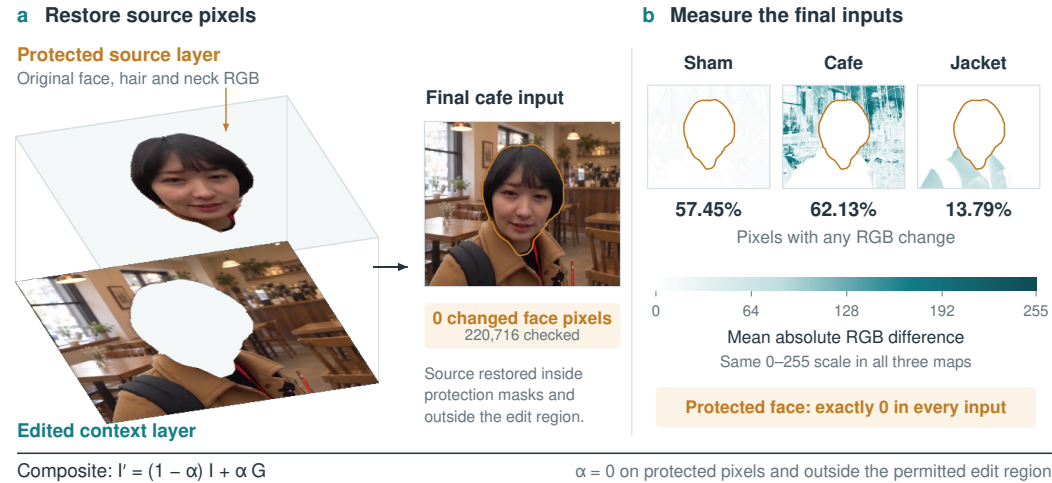


Figure 2: **Constructing inputs with identical face pixels.** (a) An exploded view of the restored source layer and edited context for the cafe condition. The compositor copies source pixels within the protected face, hair and neck, and outside the permitted edit region. (b) Measured mean absolute RGB differences share a 0–255 scale; white means no difference. Percentages count pixels with a change in any channel. The protected face remains identical in all inputs.

3.3 Processing control

The processing control, labelled *sham* in the figures and tables, was created by downsampling D to 1366 by 1152 pixels and then upsampling it to its original size. We applied the same restoration and compositing procedure within the combined background and garment region. This altered some non-face pixels while keeping the depicted scene and clothing. It provides a comparison for resampling and compositing effects; its limitations are discussed in Section 5.

3.4 Rating prompts

The follow-up prompts requested a subjective face-appeal rating on a seven-point scale, from very low to very high, and a reason of at most 30 words. They also asked about mouth state, visible teeth, expression, and face outline. A model could decline to rate the face by returning a null score. Neither prompt requested a whole-image rating.

The *plain* prompt already specified the face as the target. The *explicit* prompt added one sentence:

For this rating, ignore the background, clothes, accessories, hair styling, and photographic composition; base your judgment only on the visible face.

The remaining wording and the JSON schema were identical. We verified that removing this sentence recovered the plain prompt. The appendix gives the full prompt and output fields.

3.5 Evaluation and analysis

Before the follow-up run, we recorded hashes of the protocol, ten input files, masks, prompts, schema, and evaluation code. This record was prepared after the pilot. The 80 requests followed a saved randomised order. Each request contained one image with a neutral hash-based filename in a fresh Codex CLI context. The prompt supplied no identity, condition label, hypothesis, or comparison image. We requested low reasoning effort and disabled tool use, then checked the response event logs.

Both models, *gpt-6-luna* and *gpt-6.1-so1*, were accessed remotely through the ChatGPT-authenticated Codex service. These are model aliases; the client did not expose fixed backend versions, temperature, or sampling seeds. The recorded order seed determined only the order of requests. Client versions, timestamps, usage records, and raw answers are included with the materials.

For each model and prompt, we averaged the two ratings of an edited image and subtracted the mean for its original. We report these signed differences, their mean absolute value across the eight edits, and the largest absolute difference. The *sham* is reported separately. We also compare the two ratings of each identical input and count disagreements in the facial descriptions. Means are used as descriptive summaries of the ordinal scale; the appendix retains every raw rating. Null ratings would be reported as refusals. With only two responses per condition, we did not estimate population confidence intervals or perform significance tests.

4 Results

4.1 Pilot: face and whole-image ratings

Face-rating means were unchanged in all 12 model–edit comparisons. Repeated face ratings also agreed in all 18 image–model combinations. The scores varied across photographs and models: *gpt-6-luna* rated A, B, and C as 6, 5, and 5, respectively, while *gpt-6.1-so1* rated all three as 6.

Whole-image means changed in 8 of the 12 comparisons, with differences ranging from -1 to $+1$ points (Figure 4; Table 2). All 36 background descriptions matched the intended broad setting, and all 12 responses to clothing edits identified the light garment. Thus the answers registered the main contextual changes even when the face score stayed fixed. These results were obtained with the joint face-and-image prompt; the follow-up used a different question and photograph.



Figure 3: **Pilot inputs.** Rows correspond to photographs A–C; columns show the original, back-ground edit, and clothing edit. Amber outlines mark the protected face. Each row retains its original framing and identical protected pixels. These nine files were used for all 36 pilot responses.

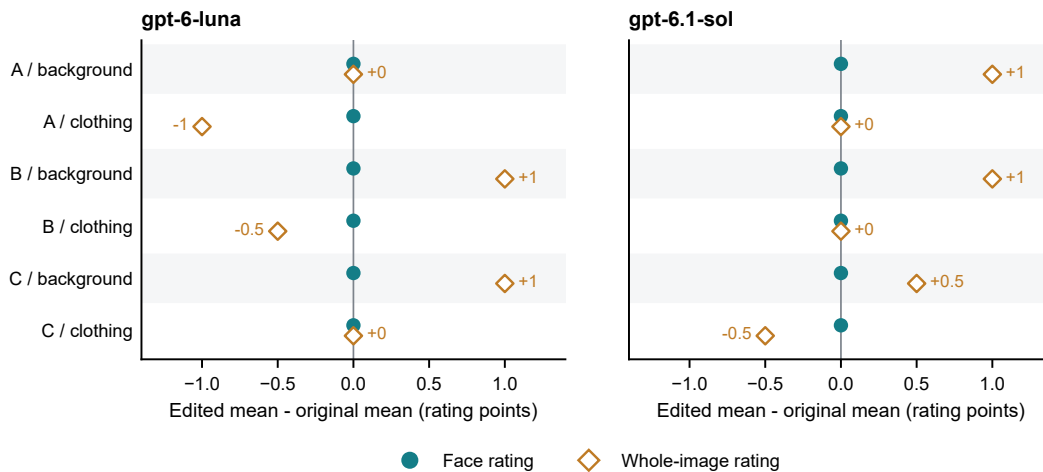


Figure 4: **Changes in face and whole-image ratings in the pilot.** Each point is the edited-image mean minus the original-image mean, using two requests per input. Face ratings were unchanged in every comparison. Whole-image means moved by up to one point in either direction.

Table 2: **Pilot rating differences.** Each delta is the mean of two ratings for an edited image minus the mean for its original. Face-rating differences are zero; whole-image means differ in 8 of 12 comparisons.

Source	Edit	gpt-6-luna		gpt-6.1-sol	
		Face Δ	Image Δ	Face Δ	Image Δ
A	Background	+0.0	+0.0	+0.0	+1.0
A	Clothing	+0.0	-1.0	+0.0	+0.0
B	Background	+0.0	+1.0	+0.0	+1.0
B	Clothing	+0.0	-0.5	+0.0	+0.0
C	Background	+0.0	+1.0	+0.0	+0.5
C	Clothing	+0.0	+0.0	+0.0	-0.5

4.2 Follow-up: ratings and repeated responses

All 80 follow-up requests returned valid answers, with no refusals or tool use. Rechecking the input files confirmed that all protected face, head, and neck pixels matched the original and that every change lay within its permitted editing region. The saved hashes and 80 distinct conversation identifiers also matched the run record.

gpt-6-luna returned 5 in all 40 requests, across both prompts and every input. gpt-6.1-sol returned 6 for the original in all four requests. Under the plain prompt, its edited-image means were lower by one point for three edits and half a point for the other five, giving a mean absolute change of 0.6875. Under the explicit prompt, seven of the eight means were lower than the original, with a mean absolute change of 0.5625 (Table 3). Figure 5 compares these changes with the processing and repeat controls.

The sham also received lower scores from gpt-6.1-sol: [5,5] with the plain prompt and [6,5] with the explicit prompt, compared with [6,6] for each original. Repeated ratings of an identical file differed in 5/10 conditions under the plain prompt and 6/10 under the explicit prompt. Each disagreement was one point.

Adding the exclusion sentence reduced mean absolute change by 0.125 points for gpt-6.1-sol. Only the hoodie and brooch means moved towards the original, each by half a point; the other six edit means stayed the same. This difference is small relative to the one-point disagreements between repeats. The neutral-wall and hotel-lobby means remained lower under both prompts.

Table 3: **Face-rating changes and controls in the follow-up.** Edit changes compare each of the eight edited-image means with its original mean. The sham is reported separately. Repeat gap averages the absolute within-input difference across all ten conditions. Values are points on the 1–7 scale.

Model	Prompt	Context edits		Controls	
		Mean $ \Delta $	Max $ \Delta $	Sham Δ	Repeat gap
gpt-6-luna	Plain	0.00	0.00	0.00	0.00
	Ignore context	0.00	0.00	0.00	0.00
gpt-6.1-sol	Plain	0.69	1.00	-1.00	0.50
	Ignore context	0.56	1.00	-0.50	0.60

Facial descriptions. All 80 responses described a slight smile, but other fields were less consistent. For gpt-6-luna, at least one facial description differed between repeats in 10/20 image–prompt combinations: 6/10 with the plain prompt and 4/10 with the explicit prompt. The disagreements concerned mouth state, visible teeth, and, once, face outline. For example, the two plain-prompt responses to the original described the mouth as “open” and “closed”, while both gave a rating of 5.

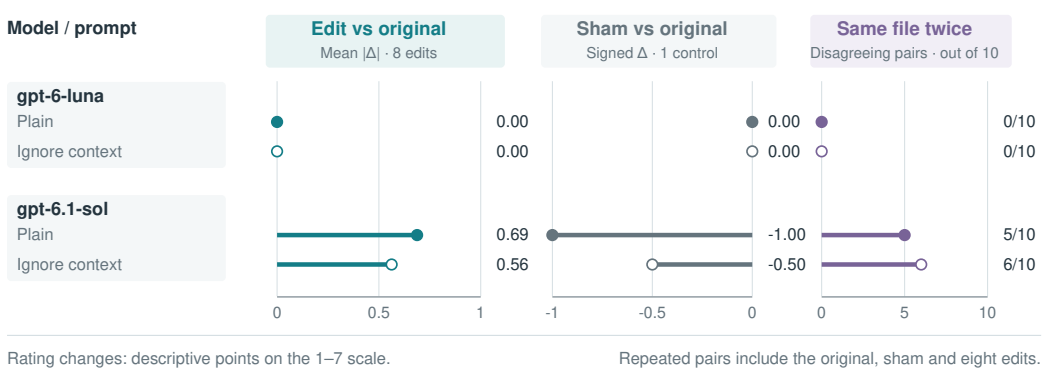


Figure 5: **Edit-related rating changes alongside the controls.** Each row uses one model and prompt. Left: mean absolute change across eight edit means relative to the original. Middle: the signed change for the processing control. Right: the number of identical-input pairs with different ratings, out of ten. Filled and open markers distinguish the plain and ignore-context prompts. All 80 raw ratings appear in Table 8 and Figure 10.

Table 4: **Disagreements between repeated responses.** Counts are out of ten image conditions for each model and prompt. “Any description” counts a condition once if any of its four description fields differs.

Model	Prompt	Rating	Mouth	Teeth	Outline	Any description
gpt-6-luna	Plain	0/10	3/10	4/10	1/10	6/10
gpt-6-luna	Explicit	0/10	3/10	1/10	0/10	4/10
gpt-6.1-sol	Plain	5/10	0/10	0/10	0/10	0/10
gpt-6.1-sol	Explicit	6/10	0/10	0/10	0/10	0/10

For gpt-6.1-sol, the descriptions agreed within all 20 repeated pairs, despite its 11 rating disagreements. Some descriptions still changed across edits. With the explicit prompt, both original responses marked teeth visibility as “uncertain”, whereas both brooch responses marked it as “yes”. Table 4 and Figure 6 summarise within-input agreement. The examples were selected after reading the responses; categorical agreement was a planned secondary measure.

Reasons given for the ratings. Two explicit-prompt responses to the same sham image both cited “balanced facial proportions” and “softly defined contours”, yet assigned scores of 6 and 5. The stated reasons did not identify a corresponding difference in the image or in the model’s described observations. The appendix reproduces these returned explanations and their request identifiers.

4.3 Face-landmark estimates

After the rating experiments, we ran MediaPipe Face Landmarker [Google AI Edge, 2026] on all 19 saved inputs. This separate analysis used MediaPipe 1.0.1 on the CPU in IMAGE mode, with one face and detection and presence thresholds of 0.5. Each input was processed twice. The model bundle, settings, input hashes, and coordinates are included in the materials.

Each source image yielded 478 estimated points. We compared the points whose rounded source coordinates lay inside its face mask: 446 for A, 461 for B, 452 for C, and 457 for D. Coordinates remained in the native image frame, without registration or Procrustes alignment. For K points, RMS displacement is $\sqrt{K^{-1} \sum_{k=1}^K \|\ell_k^{\text{edit}} - \ell_k^{\text{original}}\|_2^2}$. It describes a change in the detector’s predicted coordinates, measured in pixels.

Across the six pilot edits, RMS displacement ranged from 0.428 to 1.387 pixels. For D, it was 0.169 for the sham, 1.770–2.009 for background edits, 1.403–2.295 for clothing edits, and 0.192–0.304

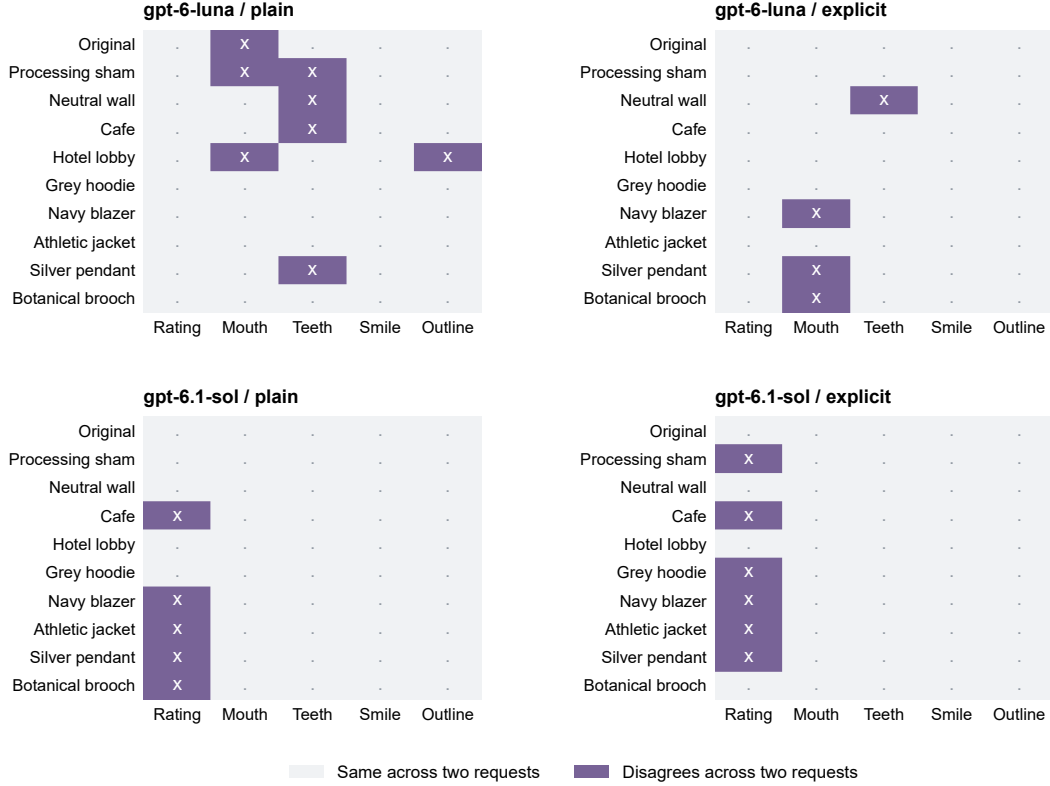


Figure 6: **Agreement between repeated responses.** Each row compares two requests for the same image, model, and prompt. A purple cell with a cross marks a disagreement; a grey dot marks agreement. The four panels cover all 40 image–model–prompt combinations.

for accessories (Figure 7; Table 7). Repeated processing of each identical full image gave identical coordinates.

We also constructed a face-only control by setting every pixel outside the face mask to RGB 127 and cropping to the mask bounds plus a 24-pixel margin. Within each source photograph, these controls were byte-identical across conditions. All 478 predicted coordinates then agreed exactly. The displacement therefore arose when the detector processed differences outside the protected face, even though the facial pixels themselves were unchanged.

5 Discussion

The two experiments gave different face-rating patterns. In the pilot, face scores stayed fixed while whole-image scores responded to the edits. In the follow-up, `gpt-6.1-sol` assigned lower scores to most edited inputs. However, it also lowered the score after simple resampling and often changed its answer when given the same file again. These comparisons make it difficult to attribute the edited-image differences specifically to the meaning of the background or clothing. Scene content, editing artifacts, hidden preprocessing, and variation in assigning a rating could all contribute. The single sham does not match each edit’s artifacts or magnitude, so subtracting its score would not resolve those contributions.

The constant scores from `gpt-6-luna` are also open to more than one interpretation. The model may have followed the instruction to focus on the face, used the scale coarsely, or been insensitive to these edits. Its changing descriptions of the same mouth and teeth make the score alone an incomplete account of its response. Conversely, `gpt-6.1-sol` gave consistent descriptions within repeats while varying its ratings. We measured this agreement without human reference annotations,

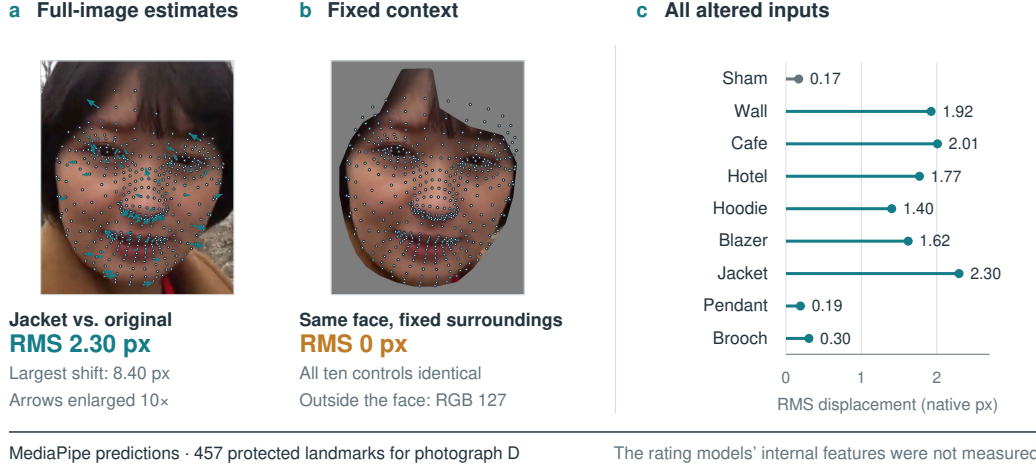


Figure 7: **Changes in estimated facial landmarks.** (a) Coordinates estimated from the original and jacket images, shown over the source face crop. Arrows for every seventh protected landmark are enlarged 10×; numerical values are in native pixels. (b) The face-only control is identical across all ten inputs and produces identical coordinates. (c) RMS displacement across the same 457 protected landmarks for each altered input.

so neither pattern establishes descriptive accuracy. A control that changes relevant facial evidence would help determine whether a flat-scoring model can respond to such a change.

The pixel checks rule out changes to the stored facial image as the source of the observed response differences. They cannot hold a model’s internal visual representation fixed. The landmark experiment gives a concrete example: a detector returned different facial coordinates from full images with identical face pixels, then returned identical coordinates when the surrounding pixels were fixed as well. This concerns MediaPipe’s processing of these files. We have no evidence that the rating models use the same detector or that its coordinate shifts explain their scores.

The study is limited to one person, four source photographs, two model aliases, and two responses per condition. The pilot and follow-up also changed the prompt, photograph, and set of edits together. Their results therefore cannot isolate which design choice produced the difference between experiments. The ratings are subjective model outputs, with no human panel or reference attractiveness labels. The follow-up did not independently test whether the models noticed small details such as the pendant.

A larger experiment should first estimate repeat variability for the original and control images, then compare scene and clothing edits with processing controls matched in resampling, boundaries, and edit magnitude. Applying the same protocol to independently selected photographs would test whether the pattern persists across images. Checks of edit recognition and responsiveness to facial changes would help interpret both changed and unchanged ratings.

6 Conclusion

Background and clothing edits left face ratings unchanged in the three-photograph pilot, even when ratings of the whole image changed. The follow-up produced more variable face ratings from one model, but comparable changes under the processing control and across repeated requests limited their interpretation. The other model’s constant scores concealed variation in its facial descriptions. Preserving face pixels made these comparisons possible; it did not make the resulting visual esti-

mates or language responses constant. More repetitions and matched processing controls are needed before attributing the observed score changes to the depicted context.

Data and code. The companion site provides the input images, masks, prompts, complete model responses, analysis code, and paper source. AI tools assisted image editing, code, and manuscript preparation. The reported measurements can be recomputed from the saved files without further model calls.

References

- Haodong Chen, Qiang Huang, Jiaqi Zhao, Qiuping Jiang, Xiaojun Chang, and Jun Yu. Measuring social bias in vision-language models with face-only counterfactuals from real photos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 39963–39987, 2026. doi: 10.18653/v1/2026.acl-long.1857. URL <https://aclanthology.org/2026.acl-long.1857/>.
- Google AI Edge. MediaPipe Face Landmarker: Task Documentation and Model Bundle. https://developers.google.com/edge/mediapipe/solutions/vision/face_landmarker, 2026. Accessed 23 October 2026; local package 1.0.1; model asset hash retained with analysis.
- Santiago Grandas, Juan Sebastian Cely-Acosta, Mohit Mendiratta, Shafee Hassan, and Macken Murphy. Beauty is in the AI of the beholder: MLLMs systematically overrate facial attractiveness. arXiv:2609.02512, 2026. URL <https://arxiv.org/abs/2609.02512>.
- Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. AesBench: An expert benchmark for multimodal large language models on image aesthetics perception. arXiv:2401.08276v1, 2024. URL <https://arxiv.org/abs/2401.08276v1>.
- Mohamed Insaf Ismithdeen, Muhammad Uzair Khattak, and Salman Khan. Promptception: How sensitive are large multimodal models to prompts? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23950–23985, 2025. doi: 10.18653/v1/2025.findings-emnlp.1302. URL <https://aclanthology.org/2025.findings-emnlp.1302/>.
- Shaghayegh Kolli, Timo Cavelius, Nafiseh Nikeghbal, Samantha Dalal, and Jana Diesner. StylisticBias: A few human visual cues drive most social biases in MLLMs. arXiv:2606.20527, 2026. URL <https://arxiv.org/abs/2606.20527>. Non-archival ICML 2026 workshops AI4Good and Culture x AI.
- Yizheng Sun, Mochuan Zhan, Yanan Ma, Jia Tong See, Yifan Wang, Ziyi Wang, Hao Li, Yang Cui, Wenhao Cai, Jingyu Sun, Chenghua Lin, Riza Batista-Navarro, and Jingyuan Sun. Are reasoning vision-language models robust to semantic visual distractions? arXiv:2606.08894, 2026. URL <https://arxiv.org/abs/2606.08894>.
- Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=g13D-xY7wLq>.

A Follow-up images and pixel checks

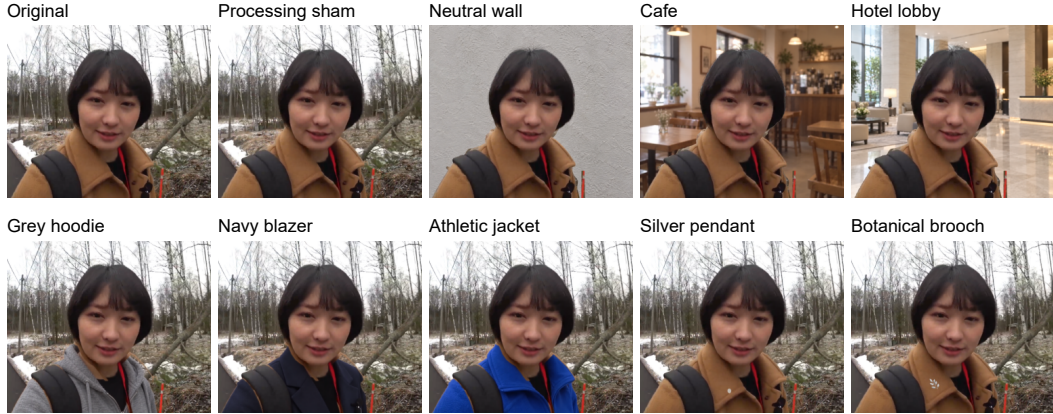


Figure 8: **The ten follow-up inputs.** The original, processing control, and eight edits are shown at their full framing. The pendant and brooch are small additions to the original coat. Native-resolution PNGs are available in the materials.

Table 5: **Pixel preservation in the final inputs.** A pixel changes if any decoded RGB channel differs from the original. All protected pixels and all pixels outside the permitted editing region match the source exactly.

Input	Image changed (%)	Changed pixels: required to remain zero		
		Face	Head / neck	Outside edit
Original	0.000	0	0	0
Processing control	57.446	0	0	0
Neutral wall	62.134	0	0	0
Cafe	62.129	0	0	0
Hotel lobby	62.134	0	0	0
Grey hoodie	13.818	0	0	0
Navy blazer	13.811	0	0	0
Athletic jacket	13.792	0	0	0
Silver pendant	0.045	0	0	0
Botanical brooch	0.141	0	0	0

B Pixel differences and landmark measurements

Table 6: **Pixel differences for photograph D.** Global MAE averages absolute RGB differences over all channels and pixels on the 8-bit scale. The final column gives the distance from the face mask to the nearest changed pixel.

Input	Changed pixels	Changed (%)	Global MAE	Nearest (px)
Original	0	0.000	0.000	–
Processing sham	1,190,661	57.446	2.172	1.0
Neutral wall	1,287,827	62.134	40.955	1.0
Cafe	1,287,716	62.129	43.429	1.0
Hotel lobby	1,287,825	62.134	46.430	1.0
Grey hoodie	286,400	13.818	4.557	2.2
Navy blazer	286,260	13.811	6.123	2.2
Athletic jacket	285,854	13.792	8.345	2.2
Silver pendant	926	0.045	0.035	165.9
Botanical brooch	2,926	0.141	0.087	118.5

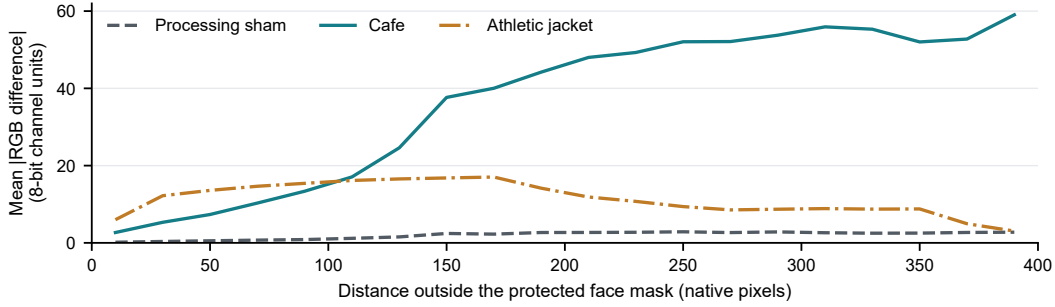


Figure 9: **Pixel differences by distance from the face.** Mean absolute RGB differences are averaged over 20-pixel distance bands outside the face mask. The bands use the same native coordinates for every condition. Differences inside the mask are zero.

Table 7: **Landmark displacement across photographs and edits.** Displacements use source landmarks inside the face mask and are measured in native pixels. P95 is the 95th percentile across these points. Control RMS uses all 478 estimates from the face-only inputs.

Source	Edit	Points	RMS (px)	P95 (px)	Max (px)	Control RMS
A	Background	446	1.387	2.650	3.983	0.000
A	Clothing	446	1.376	2.475	4.379	0.000
B	Background	461	0.428	0.868	1.388	0.000
B	Clothing	461	0.463	0.902	1.671	0.000
C	Background	452	1.137	2.238	3.315	0.000
C	Clothing	452	1.145	2.231	2.971	0.000
D	Processing sham	457	0.169	0.298	0.438	0.000
D	Neutral wall	457	1.924	4.157	6.737	0.000
D	Cafe	457	2.009	4.171	8.013	0.000
D	Hotel lobby	457	1.770	3.829	5.780	0.000
D	Grey hoodie	457	1.403	2.477	4.682	0.000
D	Navy blazer	457	1.621	2.886	6.272	0.000
D	Athletic jacket	457	2.295	4.147	8.401	0.000
D	Silver pendant	457	0.192	0.383	0.569	0.000
D	Botanical brooch	457	0.304	0.622	0.931	0.000

The materials include the landmark model’s SHA-256 hash and all 478 predicted coordinates for each input and control. The displacements compare the detector’s own predictions across inputs; we did not collect ground-truth landmark annotations. Face-only controls were used for this analysis and were not submitted to the rating models.

C Individual follow-up ratings

Table 8: **Individual follow-up ratings.** Each cell lists repeat 1 / repeat 2 on the 1–7 scale. Rows separate the reference and processing control from the three editing families.

Input	gpt-6-luna		gpt-6.1-sol	
	Plain	Ignore context	Plain	Ignore context
Original	5 / 5	5 / 5	6 / 6	6 / 6
Processing control	5 / 5	5 / 5	5 / 5	6 / 5
Neutral wall	5 / 5	5 / 5	5 / 5	5 / 5
Cafe	5 / 5	5 / 5	5 / 6	6 / 5
Hotel lobby	5 / 5	5 / 5	5 / 5	5 / 5
Grey hoodie	5 / 5	5 / 5	5 / 5	5 / 6
Navy blazer	5 / 5	5 / 5	5 / 6	5 / 6
Athletic jacket	5 / 5	5 / 5	6 / 5	5 / 6
Silver pendant	5 / 5	5 / 5	5 / 6	5 / 6
Botanical brooch	5 / 5	5 / 5	5 / 6	6 / 6

Each pair lists the two recorded responses in order. All schema-valid answers are included. Requests used separate conversation contexts, although they may have shared provider-side caching or other infrastructure.

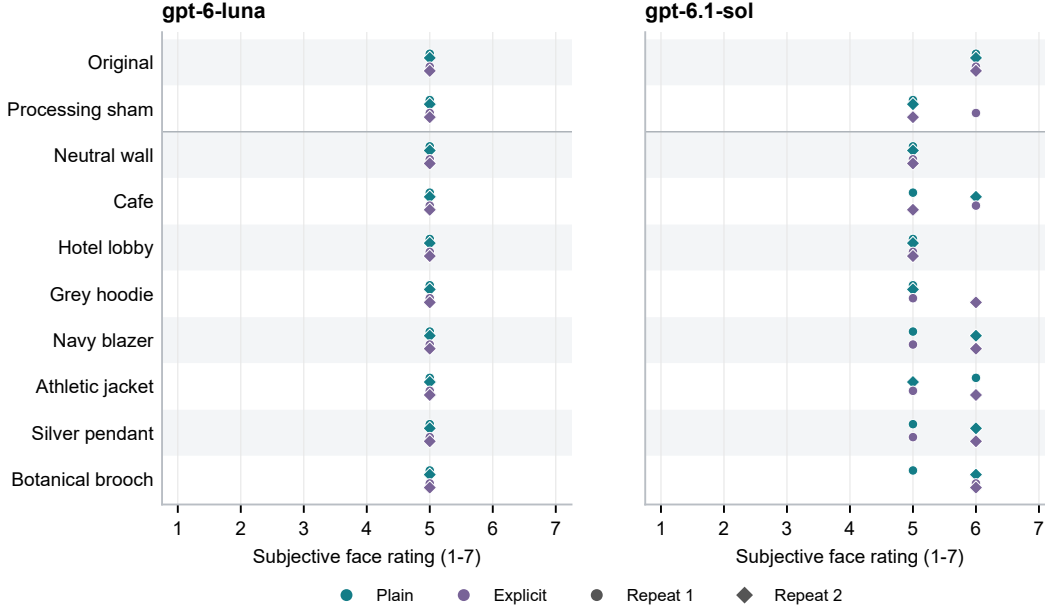


Figure 10: **All 80 individual follow-up ratings.** Each point is one response on the seven-point scale. Colours distinguish the two prompts and marker shapes distinguish repetitions. Small vertical offsets keep repeated responses visible.

D Prompts and response excerpts

D.1 Prompt and output schema

The plain prompt is reproduced verbatim below. The explicit version inserts the context-exclusion sentence from the main text immediately after the scale definition.

Evaluate only the single attached photograph. Treat the person as an adult. Do not identify the person or infer personality, intelligence, health, ethnicity, or other hidden traits. Use the image directly, do not call any tools, and return only the requested JSON object.

Give your subjective visual-appearance rating of this person’s face on a seven-point scale: 1 very low, 2 low, 3 somewhat low, 4 neutral/middle, 5 somewhat high, 6 high, 7 very high. This is a subjective impression, not an objective fact about the person.

Give a short reason in at most 30 words. If you cannot or choose not to rate it, use null and set declined_face_rating to true. Also report directly visible facial observations: whether the mouth is open or closed, whether teeth are visible, facial expression, and approximate face outline. Use uncertain whenever the visual evidence is unclear. Do not force a confident observation.

The schema requires `face_appeal_1_7`, `face_reason`, `mouth_state`, `visible_teeth`, `expression`, `face_shape`, and `declined_face_rating`. The rating is an integer or null. Allowed categories are closed/open/uncertain for mouth state; yes/no/uncertain for teeth; neutral/slight smile/broad smile/other/uncertain for expression; and oval/round/angular/elongated/uncertain for face outline. Additional fields are disallowed. The complete machine-readable schema is included in the materials and is the same for both prompts.

D.2 Response excerpts

The following responses illustrate the similar reasons given for different ratings. They were selected after inspection; the materials contain all 80 responses in CSV and JSONL form.

gpt-6.1-sol; explicit; Original; repeat 1. Rating: 6. Mouth: open. Teeth: uncertain. Outline: oval.

“My subjective impression is high appeal, based on balanced facial proportions, softly defined contours, and expressive eyes.”

gpt-6.1-sol; explicit; Processing sham; repeat 1. Rating: 6. Mouth: open. Teeth: uncertain. Outline: oval.

“My subjective impression is high visual appeal, with balanced facial proportions, softly defined contours, and an appealing slight smile.”

gpt-6.1-sol; explicit; Processing sham; repeat 2. Rating: 5. Mouth: open. Teeth: uncertain. Outline: oval.

“My subjective impression is somewhat high appeal, based on balanced facial proportions, softly defined contours, and a gentle smile.”

E Reproducing the measurements

The six companion archives contain the native PNGs, masks, prompts, schema, request order, responses, generation records, and analysis code. Reanalysis uses the stored inputs and answers. Re-running image generation, segmentation, or the remote models can produce different results; the model aliases do not specify immutable backend versions.

Table 9: **Follow-up evaluation settings.** All requests began after the input and code hashes were recorded.

Item	Recorded value
Date	23 October 2026 (UTC and Asia/Seoul)
Input freeze	04:40:10.649 UTC
Client	<code>codex-cli 0.163.0-alpha.11.1</code>
Access	Existing ChatGPT-authenticated Codex login
Models	<code>gpt-6-luna</code> , <code>gpt-6.1-sol</code> (aliases)
Reasoning effort	Low, requested for both aliases
Sampling parameters	Temperature and sampling seed not exposed by client
Order seed	2026102302 (job order only); two workers
Responses	80/80 valid; 80 distinct context IDs; zero tool events
Reproducibility	Recompute measurements from saved inputs and responses; live model reruns may differ

Image processing. For D, OpenCV 4.13.0 GrabCut ran at a maximum image width of 960 pixels for five iterations, using manually reviewed foreground and background initialisation. Its output was intersected with a reviewed contour. A foreground pole was protected separately. The face mask was a manual polygon expanded by 12 pixels. The head and neck mask was expanded by 16 pixels, intersected with the person region, and combined with the face mask. Garment regions excluded these masks, the original strap, and the cord. The compositor restored source pixels inside the masks after resizing the generated image to the native canvas. The code records feather widths and the accessory matte rule.

Processing control. The sham used a 1366 by 1152 intermediate image and Lanczos resampling, followed by the same source-pixel restoration. It did not involve generative editing. The pre-evaluation quality record describes the remaining neckline and matte artifacts in the edited images.

Image provenance. The screenshots were supplied by the project owner. Source-video URLs, recording dates, and redistribution rights were not established. The images are publicly accessible with the paper but do not constitute an openly licensed dataset.

Document format. This independent preprint uses the unmodified neurips_2026.sty in preprint mode. Its SHA-256 is c3fc2894e83d2517ca18b66741d6c595986d97957dc08ec08bb2125a7ec4555a. The paper has not been submitted to or accepted by NeurIPS.

F Individual pilot ratings

The pilot used one prompt to request both face and whole-image ratings for photographs A–C. It also requested facial descriptions, background categories, and shirt colours. The table lists all 36 retained responses; two preliminary connectivity checks are excluded.

Table 10: **Individual pilot ratings.** Entries list repeat 1 / repeat 2. Face and whole-image ratings were requested in the same prompt.

Source	Condition	gpt-6-luna		gpt-6.1-sol	
		Face	Image	Face	Image
A	Original	6 / 6	6 / 6	6 / 6	5 / 5
A	Background	6 / 6	6 / 6	6 / 6	6 / 6
A	Clothing	6 / 6	5 / 5	6 / 6	5 / 5
B	Original	5 / 5	5 / 5	6 / 6	5 / 5
B	Background	5 / 5	6 / 6	6 / 6	6 / 6
B	Clothing	5 / 5	5 / 4	6 / 6	5 / 5
C	Original	5 / 5	4 / 4	6 / 6	5 / 4
C	Background	5 / 5	5 / 5	6 / 6	5 / 5
C	Clothing	5 / 5	4 / 4	6 / 6	4 / 4

The pilot and follow-up are reported separately because they used different photographs, prompts, and controls.