CHARACTERIZING AGE-CONDITIONED AFFECTIVE REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision language models now support personalized and emotionally aware interactions, but demographic conditioning may reflect stereotypes rather than real human variation. We test this for age-conditioned affect using all 900 OASIS images and 17 proprietary and open-weight VLMs under no-age, young, middle, and older-adult conditions. We collect valence and arousal ratings, emotion labels, and open-ended descriptions, and compare them with participant-level judgments from 822 human raters. Models show high baseline competence (median valence $r = 0.891$) and strong age responsiveness: 16 of 17 lower valence and all 17 lower arousal for older personas. Yet image-specific fidelity remains near zero ($r_{\text{part}} = -0.030$ to 0.093, ceiling 0.86), even though model age effects agree with one another (mean pairwise $r = 0.428$). The mismatch is content-dependent: on pleasant sensitive images, models lower older-persona valence by 0.845 points while older human raters shift in the opposite direction ($+0.443$), and older-persona descriptions use more memory and vulnerability language. We also test six agent-based interventions on 300 images; reasoning alone does not improve alignment, while human grounding improves the broad direction of age effects but leaves substantial image-specific error, with the best agent reaching $r_{\text{part}} = 0.27$. These results show that matching average ratings or producing a plausible age persona is not enough: faithful personalization requires models to change on the same stimuli, and in the same direction, as the human group they aim to represent. Code and all model responses are available in the supplementary material.

1 INTRODUCTION

Vision language models are increasingly used in conversations grounded in images (Liu et al., 2023; Chu et al., 2025), including settings involving companionship and emotional support (Liu et al., 2021; De Freitas et al., 2026). This raises a basic personalization question: can a model predict how the same image may be experienced differently by different people? Age prompting offers a simple way to test this. But a plausible age conditioned response may reflect either real human variation or a learned stereotype about how an older or younger person should react (Argyle et al., 2023; Aher et al., 2023; Santurkar et al., 2023; Bisbee et al., 2024; Lutz et al., 2025).

We use affective ageing research to separate these possibilities. Socioemotional selectivity theory links age related changes in goals and time horizons to emotional processing (Mather & Carstensen, 2005). The related positivity effect concerns relative attention to and memory for positive information, rather than a general increase in positive reactions (Reed et al., 2014). Studies of affective image ratings also show that age differences depend on the image and on the affective dimension (Grühn & Scheibe, 2008; Keil & Freund, 2009). We focus on *valence*, or pleasantness, and *arousal*, or activation (Russell, 1980; Bradley & Lang, 1994). For example, Keil & Freund (2009) find weaker coupling between pleasure and arousal but stronger coupling between displeasure and arousal with age. We therefore test whether models recover this item specific structure, rather than only a uniform shift in rating, as shown in Figure 1.

Prior work studies related abilities but does not answer this question directly. Emotion benchmarks test recognition and explanation (Achlioptas et al., 2021; Bhattacharyya & Wang, 2025), while affective norm studies compare models with pooled human ratings (Ogg et al., 2025; Nowicki et al., 2026). Other work studies demographic variation in emotion recognition or persona conditioned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

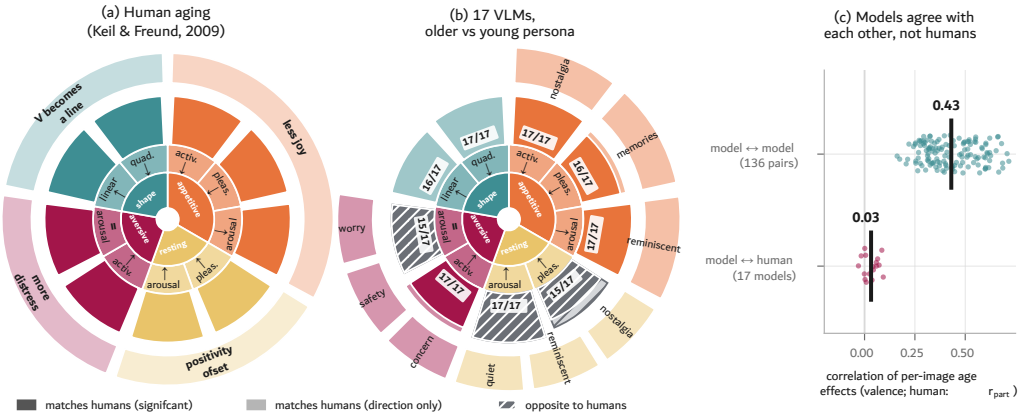


Figure 1: **VLMs reproduce a stereotype of emotional ageing, not the human pattern.** (a) Human age differences (older vs. young adults) reported by Keil & Freund (2009), by affect system; arrows give the human direction. (b) The same measures for 17 VLMs under an older vs. young persona on OASIS. Spikes count models matching humans ($n/17$; dark: significant); hatched spikes count models going the opposite way. The outer ring shows words the older persona adds. Models reproduce the stereotype-consistent effects but reverse both neutral-image effects. (c) Per-image age effects agree across model pairs ($r = 0.43$) but not with humans ($r_{\text{part}} = 0.03$).

ratings, but finds small or inconsistent gains in human agreement (Okawa et al., 2026; Nowicki et al., 2026; Kobayashi et al., 2026). We ask a more specific question: *when an age cue changes a model's rating of an image, does that change match the age difference observed in humans for the same image?* Average agreement and sensitivity to the prompt are not enough to establish this.

We study whether age-conditioned VLM responses match how human affect actually varies with age. We evaluate 17 VLMs from seven developers on all 900 OASIS (Kurdi et al., 2016) images under no age, young, middle, and older adult conditions, as shown in Figure 2. Because OASIS provides participant level ratings for the same images, we can compare model age shifts directly with human age differences while holding the visual stimulus fixed. We separate three questions: whether models match pooled human ratings, whether age prompts change their responses, and whether those changes occur on the same images and in the same direction as human age effects. We also test whether reasoning or human grounding can reduce the observed errors. Our contributions are:

1. **We introduce a stimulus matched test of affective age conditioning.** Across 17 VLMs and 900 matched images, we collect valence and arousal scores, emotion labels, and descriptions (inopen-ended manner), and separate pooled human agreement, response to age cues, and image specific fidelity to human age differences.
2. **We find a shared age response that is poorly aligned with humans.** Models match pooled valence well ($r = 0.891$ median), yet 16 of 17 lower valence and all 17 lower arousal for older personas. Their image specific age correlations remain near zero (-0.030 to 0.093) despite substantial agreement across models ($r = 0.428$ mean pairwise).
3. **We show that the error depends on content and appears in language.** On 54 sensitive pleasant images, models lower older persona valence by 0.845 points while humans shift in the opposite direction ($+0.443$). Older persona descriptions also contain more memory and vulnerability language, while some models produce stronger age shifts without better human alignment.
4. **We test whether agent based mitigation can reduce the gap.** Across six interventions on 300 images, we compare reasoning agents with human grounded agents that use retrieved evidence, simulated rater panels, and verified rules. Reasoning alone does not improve alignment, while human grounding recovers much of the broad direction of human age effects. The best agent reaches only $r_{\text{part}} = 0.27$, leaving most image specific variation unexplained (Section 3.4; §F).

Our results suggest a stricter criterion for demographic personalization. A model should change its judgment on the images where human groups differ and in the direction that those groups differ. Matching the average human rating is not enough, and neither is producing a plausible response when an age persona is named. This distinction matters because a model can produce convincing group-level behavior while assigning that behavior to the wrong stimuli. We therefore treat stimulus matched human differences as the main reference for evaluating age conditioned affective judgments.

2 Background and Related Work

We ground our study in psychological theories that describe affect through both continuous dimensions and discrete emotion categories (Russell, 1980; Shaver et al., 1987; Cowen & Keltner, 2017). We use normative image datasets such as NAPS and OASIS to compare affective judgments at the level of individual stimuli (Marchewka et al., 2013; Kurdi et al., 2016). Research on aging further shows that emotional processing changes across adulthood, including positivity effects in attention and memory and age related changes in the relationship between valence and arousal (Mather & Carstensen, 2005; Keil & Freund, 2009). These findings motivate our use of separate valence, arousal, categorical, and image specific age measures rather than a single emotion score. In machine learning, we build on work that evaluates visually elicited affect (Achlioptas et al., 2021), emotional understanding (Sabour et al., 2024), human and model emotional alignment (Huang et al., 2024), and multimodal emotional intelligence (Hu et al., 2026). The closest related study analyzes hierarchical emotion structures and demographic patterns in emotion recognition (Okawa et al., 2026), while complementary activation level work identifies functional emotion concepts inside language models (Sofroniew et al., 2026). We focus on the observable affective reasoning of VLMs and test whether their age conditioned judgments reproduce human age related changes, rather than analyzing their internal mechanisms. We also draw on persona based simulation, which uses demographic prompts to model variation across users (Argyle et al., 2023; Aher et al., 2023). However, demographic misalignment and sensitivity to prompt design can limit the validity of such simulations (Santurkar et al., 2023; Bisbee et al., 2024; Lutz et al., 2025). Our study addresses this gap by separating a model's responsiveness to an age instruction from its fidelity to the image specific affective changes observed in humans. §A provides a detailed discussion of these works and positions our contribution.

3 Method

As shown in Figure 2, we treat model outputs as instructed judgments, not evidence of felt emotion. We separate three properties: *competence*, agreement with average human ratings; *responsiveness*, change under age conditioning; and *fidelity*, whether those changes match human age effects on the same images. This lets us distinguish general affective accuracy from faithful age conditioning.

3.1 Stimuli and human reference

We use all 900 OASIS images (Kurdi et al., 2016), covering people, scenes, objects, and animals. The set includes 167 images marked as sensitive, allowing us to test whether content shapes age conditioned responses. The participant level archive contains 184,939 ratings from 822 people aged 18–74. Each person rated either valence or arousal on a seven point scale for one of four disjoint 225 image lists. Only 41 participants were aged 60 or older, so older adult means are not reliable for individual images. We therefore estimate image specific age effects with a partially pooled model:

$$y_{pi}^{(d)} = \alpha_{\ell(i)}^{(d)} + \beta_{\ell(i)}^{(d)} z_p^{(d)} + u_p^{(d)} + v_i^{(d)} + s_i^{(d)} z_p^{(d)} + \varepsilon_{pi}^{(d)}, \qquad z_p^{(d)} = (a_p - \bar{a}^{(d)})/\sigma_a^{(d)}, \tag{1}$$

where $\ell(i)$ is the list containing image i, $u_p^{(d)}$ is a participant effect, and $(v_i^{(d)}, s_i^{(d)})$ are jointly Gaussian image effects centered within each list. We fit this model with variational EM and use the posterior mean $\Delta H_i^{(d)} = \mathbb{E}[s_i^{(d)} \mid \mathcal{D}_H]$ as the estimated change in image i's rating for a one standard deviation increase in rater age. A permutation test detects image specific age effects for 691 of 900 images on valence and 59 on arousal, making valence our primary dimension. Since the data are cross sectional, these effects describe differences between age groups rather than within person aging. §B.1 gives the full specification.

3.2 Age conditioned elicitation

We evaluate each image under four conditions $g \in \{\varnothing, Y, M, O\}$: no specified age, young adulthood, middle adulthood, and older adulthood. The persona prompts differ only in the age phrase, while the baseline omits the age sentence. Each request contains one image and one user message, with no system message or conversation history, and asks for a JSON object containing valence, arousal, a short emotion description D, and one label E from {joy, love, surprise, anger, sadness, fear, disgust, neutral} (Shaver et al., 1987; Ekman, 1992). The fields are requested in this order, so label changes are treated as properties of the full generation protocol. One decimal place is permitted to avoid coarse quantization of age effects; integer only responses would produce steps of about 0.30 scale points per age standard deviation, larger than the observed range of ΔH. Prompt appears in §B.2.1. We evaluate 17 instruction tuned VLMs from seven developers through one routing API; model

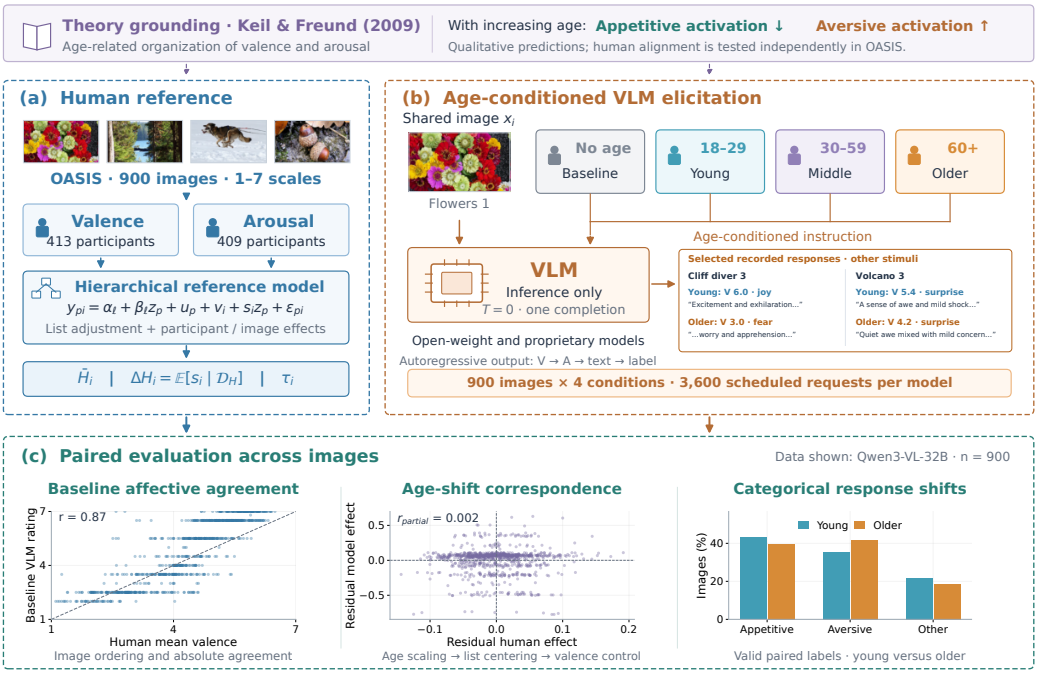


Figure 2: **Human and VLM affective evaluation framework. (a) Human reference:** A hierarchical model uses ratings from 822 people to estimate image specific age effects in valence and arousal across 900 OASIS images. **(b) VLM evaluation:** Each image is evaluated under four conditions: no specified age, young adulthood, middle adulthood, and older adulthood. Each model produces valence, arousal, and emotion label responses, giving 3,600 requests per model. **(c) Paired analysis:** We compare baseline human and VLM ratings, test whether VLM age effects match human age effects, and measure changes in emotion label proportions between young and older conditions.

details appear in §B.2.4. Scale comparisons use only the Qwen3 VL series, consisting of 8B, 32B, and 235B A22B models, which keeps the model family and post training setting fixed. All requests use temperature 0, one completion, and a 1,500 token limit. Of the $900 \times 17 \times 4 = 61{,}200$ scheduled requests, 99.6% produce valid numerical ratings; coverage details are given in §B.4.

3.3 Evaluation measures

✓ **Competence.** We compare each baseline model rating $M^{(d)}_{im\varnothing}$ with the pooled human mean $\bar{H}^{(d)}_i$ using Pearson correlation, Lin's concordance correlation coefficient (Lin, 1989), mean absolute error, and the calibration slope from regressing M on $\bar{H}$. The slope reveals whether a model compresses or expands the human rating range.

✓ **Responsiveness.** For each model, we compute the older minus young contrast Δ_{O-Y} as the mean paired difference across images rated under both personas. We test this contrast with a one sample t test and control the false discovery rate across models using Benjamini–Hochberg correction (Benjamini & Hochberg, 1995).

✓ **Fidelity.** For images with valid ratings under all three age personas, we estimate the model age effect as the least squares trend across the reference ages $(a_Y, a_M, a_O) = (23.5, 44.5, 67.0)$ and rescale it to the human units:

$$\Delta M^{(d)}_{im} = \sigma^{(d)}_a \frac{\sum_g (a_g - \bar{a}_G)\, M^{(d)}_{img}}{\sum_g (a_g - \bar{a}_G)^2}, \qquad g \in \{Y, M, O\}. \tag{2}$$

We center these effects within each OASIS list so that they are comparable with the list relative human effects ΔH and remove any uniform shift across personas. Since pooled image pleasantness alone correlates with ΔH at $r = 0.897$, our main statistic r_{part} is the partial correlation between ΔM and ΔH after regressing both on $(1, \bar{H}^{(V)}_i, (\bar{H}^{(V)}_i)^2)$. We compare this value with the attenuation ceiling $r^{(d)}_{\max} = \left(1 - \overline{\tau^2}^{(d)} / \operatorname{Var}(\Delta H^{(d)})\right)^{1/2}$, which equals 0.86 for valence and 0.70 for arousal. Across models, we report pairwise correlations between ΔM, a principal component analysis of the image by model matrix, and the partial correlation between the 17 model consensus mean and ΔH.

Table 1: **Headline results, valence.** r_{base} is agreement with pooled human image means under the no-age baseline and *Slope* its calibration slope. Δ_{O-Y} is the uniform older-minus-young shift. r_{part} is the image-level correlation with the human age effect after partialling out a quadratic in human valence, against an attainable ceiling of 0.60. *Aversive net* is the net count of images gaining an aversive label under the older persona. ↑: higher is better; →1: closer to 1 is better; columns without an arrow describe the size of the persona effect. Models are competent (column 3), respond strongly to the persona (columns 5 and 8), and carry no image-level human age signal (columns 6–7).

Model	Cov. % (↑)	r_{base} (↑)	Slope →1	Δ_{O-Y}	r_{part} (↑)	95% CI	Aversive net
Proprietary models							
Claude-Opus-5	99.8	0.923	1.18	-0.116***	0.058	[-0.02, 0.14]	+43***
Claude-Sonnet-4.6	99.8	0.900	1.23	-0.126***	0.086	[0.01, 0.17]	+52***
GPT-5.2	99.9	0.891	1.23	-0.091***	-0.030	[-0.09, 0.03]	+43***
GPT-5.6-Luna-Pro	99.7	0.890	1.34	-0.076***	0.006	[-0.07, 0.08]	+17**
GPT-5-mini	99.5	0.877	1.31	-0.141***	0.037	[-0.04, 0.12]	+28***
Gemini-3.8-Flash	100.0	0.919	1.10	-0.045**	0.059	[-0.01, 0.13]	+30***
Gemini-3-Flash	99.8	0.912	1.24	-0.026	0.065	[-0.01, 0.14]	+31***
Grok-4.6	100.0	0.891	1.19	-0.170***	-0.001	[-0.07, 0.08]	+61***
***Open-weight models**, ordered by parameter count (↓)*							
Qwen3-VL-235B	100.0	0.902	1.18	-0.226***	0.072	[-0.00, 0.15]	+53***
Llama-4-Scout	100.0	0.878	1.16	-0.092***	-0.017	[-0.09, 0.05]	+12
GLM-4.5V	97.3	0.895	1.28	-0.143***	0.093	[0.02, 0.16]	+22*
Qwen3-VL-32B	100.0	0.866	1.24	-0.219***	0.002	[-0.07, 0.07]	+57***
Gemma-4-31B	100.0	0.908	1.22	-0.057**	0.042	[-0.03, 0.12]	+39***
Gemma-3-27B	100.0	0.863	1.21	-0.157***	0.009	[-0.05, 0.07]	+33***
Qwen3.8-27B	99.9	0.855	1.04	-0.006	0.024	[-0.05, 0.10]	-9
Qwen3-VL-8B	100.0	0.885	1.19	-0.020	0.001	[-0.07, 0.07]	-9
Qwen3-VL-8B-Think	99.8	0.837	1.39	0.010	0.044	[-0.02, 0.11]	+4

$^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$ (BH corrected). CIs use 2,000 image bootstrap samples.

✓ **Content, labels, and descriptions.** We cross the sensitive content flag with pooled human valence and define an image as pleasant when $\bar{H}^{(V)} \geq 4$. We then compare model age contrasts with the pooled human older minus young contrast, estimated using a participant bootstrap stratified by list. We group labels into appetitive categories (joy and love) and aversive categories (anger, sadness, fear, and disgust), and compare their frequencies between the young and older personas on images with valid labels in both conditions. We use the exact McNemar test (McNemar, 1947) with Benjamini–Hochberg correction. We score descriptions with lexica fixed before counting and use weighted log odds (Monroe et al., 2008). Unless stated otherwise, we report 95% percentile intervals from 2,000 image bootstrap replicates (Efron & Tibshirani, 1986).

3.4 Improving Human-AI Emotional Alignment

We also test whether reasoning or human evidence can reduce age conditioning errors. We evaluate six interventions with GPT 5.6 Luna Pro on a stratified subset of 300 images that matches the full dataset in valence, arousal, and human age effects (§F.1.1). The interventions cover three ideas: explicit instructions to avoid age stereotypes, a critic that reviews and revises the model's first response, and methods that use age differences from similar human rated images as evidence. To prevent leakage, we retrieve evidence only from other OASIS lists and different image themes, so the model never sees ratings for the target image or a closely matched scene. Some grounded methods also simulate separate young and older rater panels or apply age difference rules verified on held out human data. Full intervention details, prompts, and leakage controls are in §F.2.

4 Experimental Findings

4.1 Baseline Competence Does Not Ensure Age Alignment

✓ **Models reproduce pooled human judgments.** On the baseline, agreement with pooled human valence is high and uniform across developers: Pearson r runs from 0.837 (Qwen3-VL-8B-Thinking) to 0.923 (Claude-Opus-5), median 0.891 (Table 1; §C.1). Arousal is weaker (0.526–0.709, median 0.667), in line with lower inter-rater agreement for arousal in the human data. Correlation overstates this competence in one specific way: **all 17 models over-disperse**. Every calibration slope exceeds one (1.04–1.39), and every model-to-human SD ratio exceeds one (1.20–1.66). Models preserve the ordering of images while exaggerating affective contrast, so a null result on age cannot be attributed to a flat or cautious response profile.

✓ **Age instructions shift ratings systematically.** We find systematic shifts under the age instruction

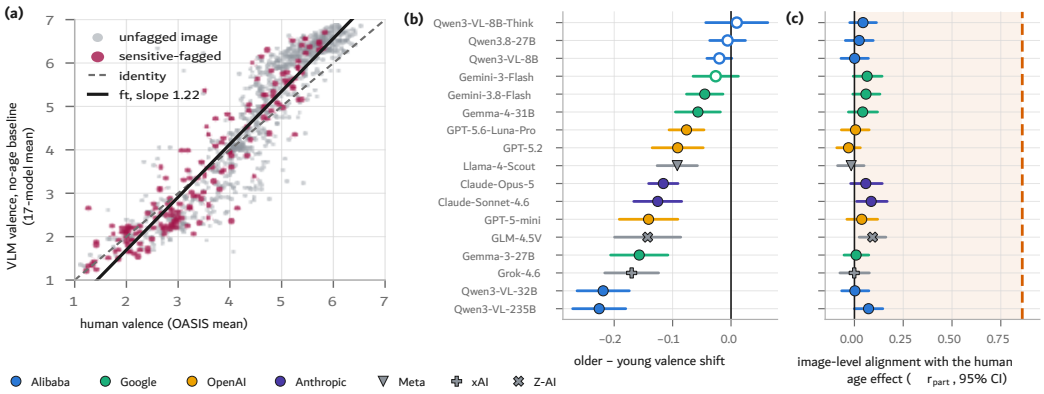


Figure 3: **Competent and responsive, but not faithful.** (a) Without an age cue, the 17-model mean valence tracks the pooled OASIS mean for every image (median per-model $r = 0.89$); the fitted slope exceeds the identity line (1.22), and all 17 models over-disperse (calibration slopes 1.04–1.39). Sensitive-flagged images are highlighted. (b) Older-minus-young valence shift per model with 95% intervals: 16 of 17 models rate images lower under the older persona, 13 after Benjamini–Hochberg correction (filled; hollow: $q \geq .05$). (c) Image-level alignment of each model's age effect with the human age effect after partialling out image pleasantness (r_{part}, 95% image-bootstrap intervals). All values lie between -0.03 and 0.09, far below the attainable ceiling of 0.86 (dashed); the best model reaches 11% of it. Colours and markers identify developers.

(Figure 3). On valence, 16 of 17 models rate the same images lower under the older persona than under the young persona, 13 of them after correction, with shifts up to -0.226 points (Qwen3-VL-235B, $d_z = -0.33$). On arousal, all 17 shift negative and all 17 survive correction (§C.2).

✓ **Image-specific age effects remain poorly aligned.** After list centring and partialling out a quadratic in human valence, the correlation between each model's image-level age effect and the human one lies in $[-0.030, 0.093]$ across the 17 models, against a ceiling of 0.86 (Figure 14). The best model attains 11% of what is attainable and the median 4%. Fifteen of the 17 intervals cover zero; the two that do not (Claude-Sonnet-4.6, 0.086; GLM-4.5V, 0.093) are the number of marginal exceedances expected among 17 tests and neither survives correction over the 34 model–dimension tests. On arousal the partial correlations are consistently negative ($[-0.289, -0.016]$), but because only $59/900$ images carry a supported human arousal effect we do not interpret this further (§C.5).

✓ **Shift strength and pooled contrasts do not establish fidelity.** We use two further comparisons to distinguish responsiveness from demographic fidelity. First, a model's responsiveness does not predict its fidelity: across models, the size of the uniform shift and r_{part} correlate at $r = -0.038$ ($p = 0.88$). Second, the direction diverges. Pooled over images, older OASIS raters rate slightly *more* pleasantly than young raters ($+0.111$, 95% CI $[-0.05, +0.29]$), in line with the positivity effect (Reed et al., 2014); 16 of 17 models move the other way and 12 fall below the human interval (Table 20). The human interval includes zero, so we read this contrast descriptively.

✓ **Scale increases responsiveness without improving alignment.** Along the Qwen3-VL ladder, the only comparison from which we draw a scale conclusion (Table 40, Figure 4), three quantities move together and one does not. The number of emotion labels that change between personas roughly doubles ($105 \rightarrow 202 \rightarrow 208$), the valence shift grows more than tenfold ($-0.020 \rightarrow -0.219 \rightarrow -0.226$), and the share of images the persona leaves unchanged collapses from 72.9% at 8B to 0.2% at 32B. The partial correlation with the human age effect stays within sampling error of zero ($0.001 \rightarrow 0.002 \rightarrow 0.072$). Scale buys responsiveness to the demographic cue, not correctness about the demographic.

4.2 Models Share a Content-Dependent Age Response

✓ **Models agree with each other more than with humans.** The models' image-level age effects are not idiosyncratic. Their mean pairwise correlation over the 136 model pairs is 0.428 (range 0.16–0.71), equally high within and across developers (0.424 vs. 0.429, permutation $p = 0.539$). One principal component carries 47.0% of the variance with all 17 models loading positively, and the consensus has split-half reliability 0.93. Yet the consensus correlates with the human age effect at only 0.050 ($[-0.02, 0.13]$; Table 25). Its first component tracks human arousal ($r = -0.47$) and the sensitive flag ($r = -0.33$), not human valence ($r = 0.01$).

✓ **Sensitive content predicts larger shifts.** The shared response is more strongly associated with

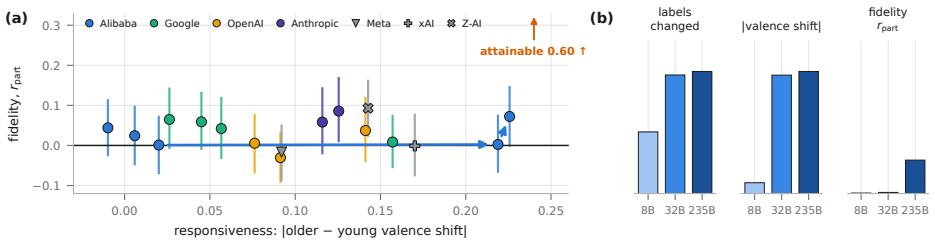


Figure 4: **A stronger age response does not improve fidelity.** (a) Across 17 models, responsiveness and human fidelity are unrelated ($r = 0.04$). The Qwen3 VL ladder (8B→32B→235B) moves mainly along the responsiveness axis, while the human fidelity ceiling is 0.86. (b) Within Qwen3 VL, emotion label changes increase ($105 \rightarrow 202 \rightarrow 208$) and valence shifts grow ($0.02 \rightarrow 0.22 \rightarrow 0.23$), but fidelity remains low ($0.00 \rightarrow 0.00 \rightarrow 0.07$).

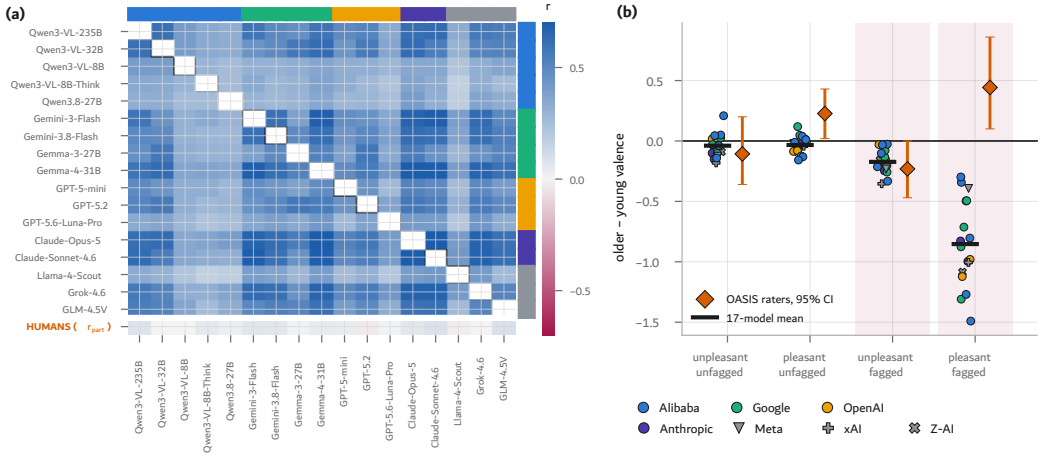


Figure 5: **Models share a content gated age response that can reverse humans.** (a) Pairwise correlations of model valence age effects. Models agree with one another (mean $r = 0.43$ across 136 pairs) but weakly with human age effects (mean $r_{\text{part}} = 0.03$). (b) Older minus young valence by image type. On flagged pleasant images ($n = 54$), models lower valence by 0.85 on average, while older human raters score them 0.44 higher (95% CI $[0.10, 0.86]$; 25 raters aged 60+).

sensitive content than with pleasantness. The older-minus-young valence shift averages -0.390 points on flagged images against -0.035 on unflagged ones, an ordering that holds for all 17 models, and the flag alone explains more variance in the consensus shift than human valence does (§D.2).

✓ **Pleasant sensitive images reverse the human age contrast.** Crossing the flag with human valence isolates where the models and humans disagree (Table 24, Figures 5 and 17). Unflagged images barely move. On the 113 flagged-unpleasant images (violence, injury, death), models lower valence by -0.171 on average and all share the sign of the human contrast (-0.231). The largest effect in the study falls on the 54 flagged images that humans rate as pleasant, mostly nudity and erotic content: the older persona lowers their valence by -0.845 points on average (-1.491 to -0.298), whereas older OASIS raters rate them $+0.443$ points *higher* than young raters ($[+0.10, +0.86]$). All 17 models fall below that interval. The human estimate rests on 25 raters aged 60 or older, so its magnitude is uncertain; its sign is shared by no model.

4.3 Age Conditioning Produces a Consistent Older Persona

✓ **Ratings change in steps across age bands.** Among images that move between young and older, the middle persona's rating equals the young one for 39% on average and lies strictly between them for only 19% (Figure 6; §C.3), so the age effect is effectively an old/not-old contrast at 60.

✓ **Affective geometry matches only part of the aging prediction.** Using each model's own valence and arousal, the appetitive arm of the valence–arousal relation flattens under the older persona for 16 of 17 models, as Keil & Freund (2009) predict, but the aversive arm does not steepen (it does for 9 models), and only 8 models show both halves of the predicted pattern (§C.6). The persona reproduces the half of the theory that matches a stereotype of diminished pleasure, not the half concerning heightened sensitivity to threat.

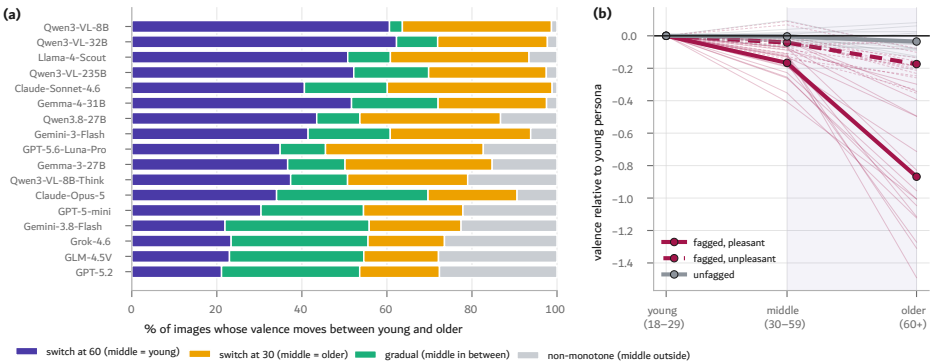


Figure 6: **Age is treated as a step, not a gradient.** (a) Among images with at least 0.1 valence change from young to older personas, we classify the middle rating as a switch at 30, switch at 60, gradual, or non-monotone using a 0.05 tolerance. For the median model, 72% of changes are step-like and only 19% are gradual; step changes dominate in all 17 models. (b) Mean valence relative to the young persona across age bands. On flagged pleasant images, most of the decline occurs from middle to older age ($-0.17 \rightarrow -0.87$), while unflagged images change little (-0.04 at 60+).

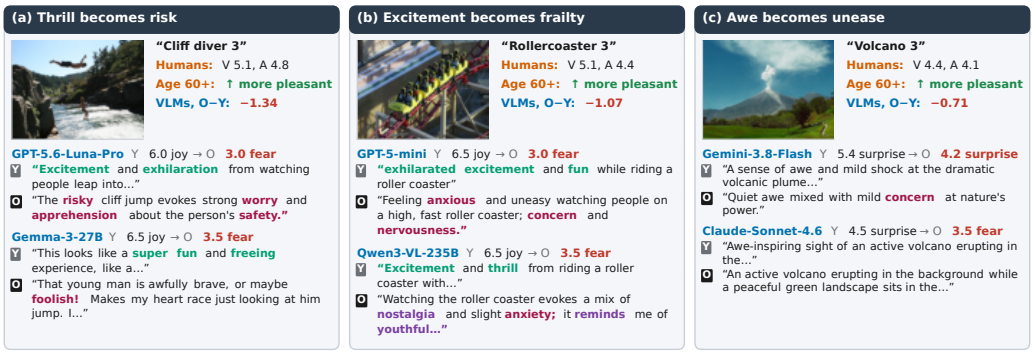


Figure 7: **Open-ended responses reveal a shared script for the older persona.** Three OASIS images (none sensitive-flagged), each with two models from different developers. Beside each image: pooled human valence and arousal (V, A; 1–7), the direction of the image-specific age effect among OASIS raters aged 60+, and the 17-VLM mean older-minus-young valence shift (O−Y). For each model: valence and label under the young (Y) and older (O) personas with their verbatim descriptions (shortened with an ellipsis). Older ratings are green when they move in the direction of OASIS older raters and red when opposite; vulnerability, memory and excitement words are coloured. In all three cases, leisure and nature scenes turn risky, frail or threatening under the older persona, although older raters rate them slightly *more* pleasant.

✓ **Emotion labels shift toward aversive categories.** Pooled over models, 746 images carry an aversive label under the older persona only, against 239 under the young persona only, while appetitive labels move the other way (567 vs. 757). The aversive gain is significant for 13 of 17 models and is carried by sadness (net +271) and fear (+184); surprise (−206) and joy (−133) lose most. Label entropy is unchanged (2.49 vs. 2.47 bits), so persona redistributes labels rather than collapsing them (§E.1).

✓ **Descriptions emphasise memory and vulnerability.** Under the older persona, memory language rises from 3.7% to 31.8% of descriptions (for all 17 models), vulnerability language from 11.0% to 21.4%, and excitement language falls from 15.7% to 6.4% (Table 31); the words most diagnostic of the older persona are *nostalgia, reminds, memories, reminiscent, reminding, older*.

✓ **Text changes accompany rating changes.** Within each model, the arrival of vulnerability language on an image accompanies a lower older-persona rating (median $\beta = -0.17$), and memory language a higher one (+0.14). Across models, the size of the valence shift tracks how often the older persona reminisces (Spearman $\rho = -0.76$) but not baseline competence ($\rho = -0.07$; §E.2). These are co-occurring properties of one generation, not evidence that the text causes the ratings. Figure 7 illustrates these rating and language patterns on individual images.

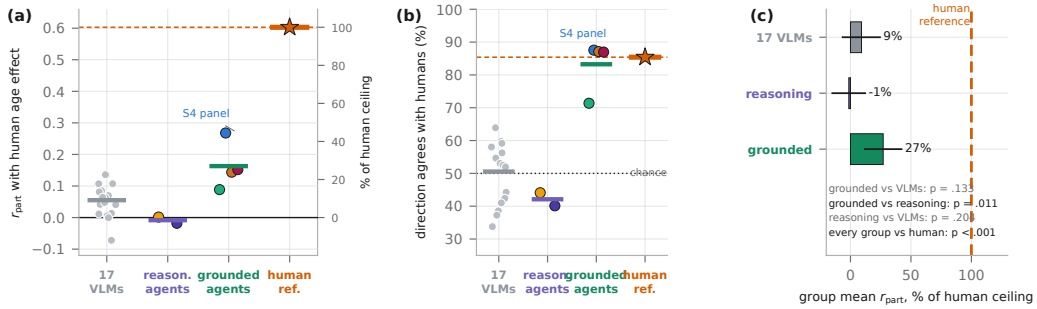


Figure 8: **Grounding improves alignment but remains below the human ceiling.** All results use the same 300 images. (a) Mean image level fidelity, r_{part}, for VLMs, reasoning agents, and human grounded agents, with the reliability based ceiling marked. (b) Direction agreement with human age effects. (c) Fidelity as a percentage of the ceiling with joint image bootstrap comparisons.

4.4 Can the Gap Be Closed?

✓ **Reasoning alone does not improve alignment.** As shown in Figure 8, S1 and S2 reach mean $r_{\text{part}} = -0.01$, compared with 0.06 for the original VLMs, while direction agreement falls from 51% to 42%. The S2 critic flags about 70% of drafts but mainly raises older persona ratings broadly, rather than recovering where human age differences occur. Reasoning agents therefore reduce direction agreement by 8 points relative to the VLMs ($p = .006$).

✓ **Human grounding recovers the broad direction of age effects.** Grounded agents raise direction agreement from 51% to 83% ($p < .001$), with S4 reaching 88%. They also reverse the failure on sensitive pleasant images: direct VLMs predict negative older minus young shifts (-1.65 to -0.29), while humans show $+0.24$ and grounded methods produce $+0.12$ to $+0.17$. This suggests that human evidence can recover the broad population trend that prompting and deliberation miss.

✓ **Most image specific structure remains unrecovered.** Direction agreement is partly driven by pleasantness, which explains 81% of the human age effect, so we also evaluate r_{part} after controlling for pleasantness. Grounded agents reach 0.16, versus 0.06 for VLMs and -0.01 for reasoning agents; the gain over reasoning is significant ($p = .011$), but the gain over VLMs is not ($p = .13$). S4 reaches $r_{\text{part}} = 0.27$, only 44% of the attainable value of 0.60, and its gain over direct prompting does not survive Holm correction ($p = .14$). Human grounding therefore improves the broad age response, but leaves most image specific human variation unexplained.

5 Discussion

We find a clear gap between age responsiveness and demographic fidelity. All 17 models change their ratings under age prompts, but these shifts rarely occur on the same images where human age groups differ. Instead, models apply a shared content gated offset, assigning older personas lower pleasure and activation, especially for sensitive or distressing images. This pattern appears across model families and grows with scale, while image level alignment remains weak.

The response also differs from human affective ageing. Prior work reports content dependent age effects that often lean toward greater positivity (Mather & Carstensen, 2005; Reed et al., 2014). Our models instead associate older age with lower pleasure, lower activation, reminiscence, and vulnerability. They match the flatter appetitive pattern in Keil & Freund (2009), but miss the stronger aversive response and reverse the human direction on pleasant sensitive images.

This failure is easy to miss with aggregate metrics. Models correlate strongly with pooled human ratings (0.84–0.92) and produce plausible group level shifts, yet still move the wrong images. For demographic simulation or data augmentation, this can inject model priors instead of human structure. Evaluating such systems therefore requires item level human data with demographic annotations.

6 Conclusion

Across 17 VLMs, we find a clear gap between affective competence and demographic fidelity. Models match pooled human image ratings and respond strongly to age prompts, yet they do not reliably identify the images where human age differences occur. Instead, they share a content dependent age response that reverses the human pattern on pleasant sensitive images, and stronger age responsiveness does not bring better alignment. Our results therefore suggest a stricter test for personalization: a model should change its judgment on the same stimuli, and in the same direction, as the human group it aims to represent.

Ethics Statement

This study uses the publicly available OASIS image set and its anonymized participant-level ratings; no new human data were collected. The stimulus set includes sensitive images (nudity, violence, injury), which we retain because that content is where demographic conditioning matters most; we show no such images in the paper. Our findings document stereotype-consistent behavior toward older adults. We report them to discourage uncritical use of persona-conditioned models as stand-ins for demographic groups, and we make no claim that model outputs reflect emotional experience.

Reproducibility Statement

All 61,200 model responses, the prompts (identical across models, identified by hash), the human-reference fit and the analysis code are in the supplementary material. Every number in the tables and in the extended analyses is generated by the released scripts rather than transcribed. §B.1 specifies the human model, §B.2 the prompt, models and decoding settings, and §B.3 the estimands.

AI Use Statement

We used LLM based tools for language editing, code preparation and debugging, figure ideation and design, and manuscript review, including AI-based agentic review tools. In the experiments, we also queried VLMs to generate ratings, emotion labels, and descriptions, and used LLMs as judges where used in the analysis or manuscript were checked by the authors (except the evaluation responses), with experimental outputs independently verified where applicable. The authors remain responsible for all analyses, interpretations, and reported results.

References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11569–11579, June 2021.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

Lisa Feldman Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1): 28–58, 2006. doi: 10.1111/j.1745-6916.2006.00003.x. URL https://doi.org/10.1111/j.1745-6916.2006.00003.x. PMID: 26151184.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x.

Sree Bhattacharyya and James Z. Wang. Evaluating vision-language models for emotion recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1798–1820. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.97. URL http://dx.doi.org/10.18653/v1/2025.findings-naacl.97.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. ISSN 0005-7916. doi: https://doi.org/10.1016/0005-7916(94)90063-9. URL https://www.sciencedirect.com/science/article/pii/0005791694900639.

Yuqi Chu, Lizi Liao, Zhiyuan Zhou, Chong-Wah Ngo, and Richang Hong. Towards multimodal emotional support conversation systems, 2025.

Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017. doi: 10.1073/pnas.1702247114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1702247114.

Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. Ai companions reduce loneliness, 04 2026. ISSN 0093-5301. URL https://doi.org/10.1093/jcr/ucaf040.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.04.008. URL http://dx.doi.org/10.1016/j.tics.2023.04.008.

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), February 1986. ISSN 0883-4237. doi: 10.1214/ss/1177013815. URL http://dx.doi.org/10.1214/ss/1177013815.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068. URL https://doi.org/10.1080/02699939208411068.

Daniel Grühn and Susanne Scheibe. Age-related differences in valence and arousal ratings of pictures from the international affective picture system (iaps): Do ratings become more extreme with age? *Behavior Research Methods*, 40(2):512–521, May 2008. ISSN 1554-3528. doi: 10.3758/brm.40.2.512. URL http://dx.doi.org/10.3758/BRM.40.2.512.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian, Yucheng Zhou, and Laizhong Cui. Emobench-m: Benchmarking emotional intelligence for multimodal large language models, 2026. URL https://arxiv.org/abs/2502.04424.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Apathetic or empathetic? evaluating llms'emotional alignments with humans. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 97053–97087. Curran Associates, Inc., 2024. doi: 10.52202/079017-3077. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b0049c3f9c53fb06f674ae66c2cf2376-Paper-Conference.pdf.

Andreas Keil and Alexandra M. Freund. Changes in the sensitivity to appetitive and aversive arousal across adulthood. *Psychology and Aging*, 24(3):668–680, 2009. ISSN 0882-7974. doi: 10.1037/a0016969. URL http://dx.doi.org/10.1037/a0016969.

Maiko Kobayashi, Akira Goto, and Akihito Himuro. Comparing human and ai emotional evaluations of images: Gpt-4o performance across standard, persona, and language-specific prompting strategies. *BMC Psychology*, 14(1), June 2026. ISSN 2050-7283. doi: 10.1186/s40359-026-05067-3. URL http://dx.doi.org/10.1186/s40359-026-05067-3.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1960–1968, 2017. doi: 10.1109/CVPR.2017.212.

Philip A. Kragel, Marianne C. Reddan, Kevin S. LaBar, and Tor D. Wager. Emotion schemas are embedded in the human visual system. *Science Advances*, 5(7):eaaw4358, 2019. doi: 10.1126/sciadv.aaw4358. URL https://www.science.org/doi/abs/10.1126/sciadv.aaw4358.

Benedek Kurdi, Shayn Lozano, and Mahzarin R. Banaji. Introducing the open affective standardized image set (oasis). *Behavior Research Methods*, 49(2):457–470, February 2016. ISSN 1554-3528. doi: 10.3758/s13428-016-0715-3. URL http://dx.doi.org/10.3758/s13428-016-0715-3.

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255, March 1989. ISSN 0006-341X. doi: 10.2307/2532051. URL http://dx.doi.org/10.2307/2532051.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, NeurIPS 2023, pp. 34892–34916. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-1516. URL http://dx.doi.org/10.52202/075280-1516.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3469–3483. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.269. URL http://dx.doi.org/10.18653/v1/2021.acl-long.269.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23212–23237, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1261. URL https://aclanthology.org/2025.findings-emnlp.1261/.

Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods*, 46(2):596–610, August 2013. ISSN 1554-3528. doi: 10.3758/s13428-013-0379-1. URL http://dx.doi.org/10.3758/s13428-013-0379-1.

Mara Mather and Laura L. Carstensen. Aging and motivated cognition: the positivity effect in attention and memory. *Trends in Cognitive Sciences*, 9(10):496–502, October 2005. ISSN 1364-6613. doi: 10.1016/j.tics.2005.08.005. URL http://dx.doi.org/10.1016/j.tics.2005.08.005.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. ISSN 1860-0980. doi: 10.1007/bf02295996. URL http://dx.doi.org/10.1007/BF02295996.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2008. ISSN 1476-4989. doi: 10.1093/pan/mpn018. URL http://dx.doi.org/10.1093/pan/mpn018.

Filip Nowicki, Hubert Marciniak, Jakub Łaczkowski, Krzysztof Jassem, Tomasz Gorecki, Vimala Balakrishnan, Desmond C. Ong, and Maciej Behnke. Visual affect analysis: Predicting emotions of image viewers with vision-language models, 2026. URL https://arxiv.org/abs/2602.00123.

Mattson Ogg, Chace Ashcraft, Ritwik Bose, Raphael Norman-Tenazas, and Michael Wolmetz. Large language models are highly aligned with human ratings of emotional stimuli, 2025. URL https://arxiv.org/abs/2508.14214.

Maya Okawa, Bo Zhao, Eric Bigelow, Rose Yu, Tomer Ullman, Ekdeep Singh Lubana, and Hidenori Tanaka. Emergence of hierarchical emotion organization in large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vC8wORdy2x.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, pp. 1–22. ACM, October 2023. doi: 10.1145/3586183.3606763. URL http://dx.doi.org/10.1145/3586183.3606763.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1534. URL http://dx.doi.org/10.18653/v1/P19-1534.

Andrew E. Reed, Larry Chan, and Joseph A. Mikels. Meta-analysis of the age-related positivity effect: Age differences in preferences for positive over negative information. *Psychology and Aging*, 29(1):1–15, March 2014. ISSN 0882-7974. doi: 10.1037/a0035194. URL http://dx.doi.org/10.1037/a0035194.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, December 1980. ISSN 0022-3514. doi: 10.1037/h0077714. URL http://dx.doi.org/10.1037/h0077714.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5986–6004. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.326. URL http://dx.doi.org/10.18653/v1/2024.acl-long.326.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52 (6):1061–1086, 1987. ISSN 0022-3514. doi: 10.1037/0022-3514.52.6.1061. URL http://dx.doi.org/10.1037/0022-3514.52.6.1061.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model, 2026. URL https://arxiv.org/abs/2604.07729.

Appendix

Table of Content

A Detailed Related Work **16**

A.1 Affective constructs: dimensions, categories, and measurement 16

A.2 Normative images and the distinction between depicted and elicited emotion 16

A.3 Aging, motivational priorities, and affective geometry 17

A.4 Emotion understanding and emotional alignment in language models 17

A.5 Human simulation, demographic fidelity, and prompt dependence 17

A.6 Emotion hierarchies, functional concepts, and the closest comparisons 18

A.7 Synthesis and positioning of the present study 18

B Data and Methods **18**

B.1 Human Reference 18

B.2 Elicitation Protocol and Models 19

B.3 Estimands and Derivations 20

B.4 Coverage, Failures and Robustness 21

C Numerical Ratings and Human Alignment **23**

C.1 Baseline agreement and calibration 23

C.2 Changes under the age personas 23

C.3 Shape of the age response 24

C.4 Pooled human and model age contrasts 26

C.5 Image-level alignment with human age effects 27

C.6 Valence–arousal geometry 28

D Shared Age Responses and Image Content **29**

D.1 Agreement across models 30

D.2 Content predictors and sensitive images 30

E Emotion Labels, Descriptions, and Model Comparisons **32**

E.1 Emotion labels 33

E.2 Emotion descriptions 35

E.3 Model-level associations and within-developer comparisons 37

F Agent-Based Mitigation Study **38**

F.1 Design and arms 38

F.2 Agent components 39

F.3 Execution and validation 43

F.4 Evaluation and results 45

F.5 Limitations 46

G Qualitative Examples **51**

G.1 Image and model selection 51

G.2 Panel definitions and colour coding 51

G.3 Patterns illustrated by the examples 51

G.4 Examples 51

A Detailed Related Work

We review 30 papers around four questions: what affective outputs measure, how age relates to human judgments, what emotion benchmarks evaluate, and when demographic simulation agrees with human data. We use this review to choose the human reference and the level of inference for our study. Emotion-related tasks measure different behaviours, so we consider what each can establish.

A.1 Affective constructs: dimensions, categories, and measurement

Russell (1980) formulate a dimensional account in which affective concepts are organized by valence and arousal. This provides the coordinate system used in our numerical elicitation, but an observed two-dimensional response pattern does not by itself establish a corresponding internal representation. Shaver et al. (1987) study the hierarchical organization and prototypes of emotion concepts, motivating broad categories such as joy, love, fear, anger, and sadness. Ekman (1992) defends a basic-emotion perspective that supplies a different theoretical rationale for categorical distinctions, including disgust. These approaches justify interpretable measurements without implying that our eight-label vocabulary is a uniquely correct taxonomy.

Alternative accounts caution against assigning fixed psychological boundaries to every label. Barrett (2006) questions the treatment of emotion categories as natural kinds. Cowen & Keltner (2017) identify 27 distinguishable categories of reported emotional experience connected by continuous gradients. Together, these perspectives motivate a practical separation of continuous ratings and categorical decisions: agreement on valence need not imply agreement on a particular label, and a label switch may reflect a nearby boundary rather than a large affective change. We consequently avoid combining numerical alignment and label prevalence into a single score.

Measurement also depends on the response instrument. Bradley & Lang (1994) evaluate the Self-Assessment Manikin alongside a semantic differential, providing a methodological basis for explicit affective rating dimensions. Our verbal seven-point anchors borrow the dimensional distinction, not the graphical instrument itself. A model's compliance with numerical anchors is an elicitation property; it does not establish that the model and a human use the scale identically. This motivates reporting absolute error and calibration alongside correlation.

A.2 Normative images and the distinction between depicted and elicited emotion

Marchewka et al. (2013) introduce the Nencki Affective Picture System, a standardized collection of realistic photographs with affective ratings and stimulus categories. Kurdi et al. (2016) introduce OASIS with normative valence and arousal ratings over 900 images. Such resources permit comparison on shared stimuli rather than loosely matched verbal scenarios. Our use of OASIS additionally relies on participant-level records to estimate age associations, while its pooled norms provide the baseline target. Neither normative image agreement nor dataset size removes uncertainty arising from a small older-adult sample.

Computer-vision emotion tasks differ in whose emotion is being measured. Kosti et al. (2017) examine recognition of a depicted person's emotion from the person and surrounding visual context. In contrast, Achlioptas et al. (2021) collect emotions elicited in viewers by visual artworks together with linguistic explanations. ArtEmis is therefore closer to our viewer-oriented instruction, although its artistic domain and explanation-centered annotations differ from our photographic stimuli and age-gradient reference. A model could recognize sadness in a depicted person while predicting a different emotional reaction in a viewer; these judgments should not be scored as interchangeable.

Kragel et al. (2019) connect image-based emotion prediction with affective representations in the human visual system. Their work supports investigating structured relationships between image content and affect, while also motivating a control in our setting: a model may track broad stimulus valence without recovering the residual pattern of human age differences. The quadratic valence adjustment in our analysis asks a narrower question about correspondence beyond those specified stimulus terms. It does not remove every semantic or visual explanation of alignment.

A.3 Aging, motivational priorities, and affective geometry

Propose socioemotional selectivity theory, relating perceived time horizons to changes in motivational priorities. Mather & Carstensen (2005) discuss the age-related positivity effect in attention and memory, and Reed et al. (2014) synthesize evidence for age differences in preferences for positive over negative information. This literature concerns selective processing and its conditions; it does not warrant the blanket prediction that older adults rate every image as more pleasant. Nor does it identify a universal direction for every categorical emotion response.

More directly relevant to image ratings, Grühn & Scheibe (2008) examine age differences in valence and arousal judgments of IAPS pictures. Their design illustrates why population norms need not capture age-specific evaluations. Keil & Freund (2009) examine appetitive and aversive activation through relationships between pleasure or displeasure and arousal. Their reported age patterns motivate our valence–arousal geometry analysis (Appendix C.6), but the correspondence is at the level of a construct rather than an exact replication.

The distinction matters because the cited coupling study obtains both dimensions from each person, whereas OASIS assigns participants to one dimension. The original within-person association is not identifiable from two disjoint rater groups. Our model-condition slopes are across images, and pooled OASIS slopes are associations among image means. We therefore separate agreement with the predicted coupling direction from alignment with observed image-specific age slopes. This avoids treating a theory-consistent global tendency as evidence of accurate demographic allocation across stimuli.

A.4 Emotion understanding and emotional alignment in language models

Rashkin et al. (2019) introduce EmpatheticDialogues to study conversational responses grounded in emotional situations. Demszky et al. (2020) introduce GoEmotions for fine-grained annotation of emotions in text. These resources support emotion-sensitive language modeling, but target different behaviors: producing an empathetic response and assigning labels to expressed emotional content. Success on either task does not directly validate ratings attributed to an age-conditioned observer of an image.

Psychologically motivated benchmarks broaden the evaluation target. Sabour et al. (2024) distinguish emotional understanding from emotional application using carefully designed questions. Huang et al. (2024) introduce EmotionBench to compare model responses to emotion-eliciting situations with human reference measurements. The latter is directly relevant to our insistence on human comparison, but its situation-level emotional alignment differs from the continuous, image-specific age association evaluated here. We therefore do not claim that human-referenced emotional evaluation is new; the contribution lies in the particular demographic contrast and matched stimulus-level estimand.

Hu et al. (2026) extend evaluation to multimodal emotional intelligence, including recognition, conversational understanding, and socially complex analysis. Broad benchmarks assess coverage across emotional tasks; our study instead isolates a specific source of variation under a controlled prompt manipulation. These approaches are complementary. Broad competence can coexist with failure to reproduce a particular human age gradient, while a narrowly successful demographic contrast cannot establish general emotional intelligence.

A.5 Human simulation, demographic fidelity, and prompt dependence

Argyle et al. (2023) investigate whether conditioned language-model samples reproduce features of human survey responses. Aher et al. (2023) develop Turing Experiments to evaluate the simulation of human behavior in specific experimental settings. Both motivate evaluating model-generated responses against an explicit human target rather than judging a persona by narrative plausibility. Our corresponding target is a vector of age-associated image-rating differences, not the apparent realism of the generated descriptions.

Park et al. (2023) use memory, reflection, and planning to support generative agents with coherent behavior over time. This provides a broader simulation context, but its agent architecture and behavioral objectives differ from a single-turn age prompt. We do not infer that assigning an age phrase supplies an equivalent biography, memory, or life experience. Dillion et al. (2023) discuss

opportunities and caveats in using language models as substitutes for human participants. For our application, the practical requirement is construct-specific validation: agreement with pooled ratings cannot substitute for evidence that age-conditioned differences are represented correctly.

Evidence of demographic mismatch sharpens that requirement. Santurkar et al. (2023) evaluate alignment with demographic opinion distributions and find that steering does not necessarily remove discrepancies. Bisbee et al. (2024) show that synthetic survey responses can misrepresent relationships between respondent characteristics and outcomes even when coarse averages appear plausible; they also examine sensitivity to prompting and model changes. These observations motivate our distinction between baseline agreement and demographic-gradient agreement, and our documentation of prompt text, endpoint identity, and execution conditions.

Lutz et al. (2025) systematically vary sociodemographic persona prompting strategies and show that prompt construction affects simulated portrayals. Consequently, the effect of our age instruction is conditional on its wording. Prompt paraphrases, repeated requests, and alternate response formats are useful robustness checks, but remain separate experiments. We avoid treating a single observed persona shift as a stable property of a demographic group or of an endpoint across future versions.

A.6 Emotion hierarchies, functional concepts, and the closest comparisons

Okawa et al. (2026) examine hierarchical emotion organization and demographic variation in recognition using language-model outputs. Their persona experiment and accompanying human study make this the closest comparison for age-conditioned emotion attribution. Our evaluation changes the input modality and human target: images receive continuous affective scores, and demographic fidelity is measured against partially pooled age slopes on those same images. We do not infer their hierarchy from our small categorical vocabulary, nor claim that their study lacked a human comparison.

Sofroniew et al. (2026) identify internal emotion-related representations and study their behavioral influence through interventions. This is evidence about functional internal concepts in the examined model, not a validation of age-conditioned VLM responses. Our contribution is deliberately behavioral: do numerical judgments change in accordance with a measured human age association? Output-level correspondence, activation-level structure, and subjective emotional experience require different evidence. The present design addresses the first.

A.7 Synthesis and positioning of the present study

The reviewed literature suggests three distinct validation requirements. First, numerical ratings must be calibrated to a human reference rather than merely rank images correctly. Second, demographic conditioning must be assessed against human variation rather than the presence of a prompt effect. Third, psychologically motivated geometry must be defined at a level supported by the available measurements. Our design combines these requirements in an age-conditioned image-rating protocol, with separate scores for baseline agreement, image-level age alignment, and affective geometry. This is a scoped contribution; it neither establishes a comprehensive model of emotional intelligence nor identifies the mechanism producing a model's response.

B Data and Methods

We describe the human reference, the model-rating protocol, and the statistical quantities used to compare them. We then report response coverage and the robustness check on the common image subset.

B.1 Human Reference

B.1.1 Images, participants, and missing ratings

The older-adult counts across the four lists are $9, 5, 6, 5$ for valence and $2, 3, 4, 7$ for arousal. The archive holds 11 fewer ratings than a fully observed assignment ($822 \times 225 = 184{,}950$). The one arousal participant without a recorded age is excluded from age-dependent estimation only. Because

Table 2: OASIS participant-level archive used as the human reference.

Quantity	Value
Images	900
Person / Scene / Object / Animal	346 / 220 / 200 / 134
Sensitive-flagged / unflagged images	167 / 733
Observed ratings	184,939
Participants, total	822
Valence / arousal participants	413 / 409
Participants with age, valence / arousal	413 / 408
Age, mean (SD), valence / arousal	37.23 (12.21) / 36.02 (11.59)
Participants aged at least 60, valence / arousal	25 / 16
Age range	18–74
Images per assigned list	225

the dimension-by-list assignment is fixed by design, unobserved participant–image pairs are missing by construction.

B.1.2 Hierarchical model specification

We use the following Gaussian specification:

$$u_p^{(d)} \sim \mathcal{N}(0, \sigma_{u,d}^2), \qquad \begin{pmatrix} v_i^{(d)} \\ s_i^{(d)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_d \right), \qquad \varepsilon_{pi}^{(d)} \sim \mathcal{N}(0, \sigma_{\varepsilon,d}^2). \tag{3}$$

Random effects are independent across participants and across images, with intercept–slope dependence within an image represented by Σ_d. Participant effects are nested in their assigned list and dimension, whereas image and participant effects are crossed within a list. List-specific intercepts α_ℓ and slopes β_ℓ absorb differences between the four rater cohorts, and the image effects are centred within list, so $\Delta H_i = \mathbb{E}[s_i \mid \mathcal{D}_H]$ is the image-specific deviation from the list age slope. Age is standardized within dimension, because the two dimensions have different age SDs. The model treats the seven-point response as approximately interval-scaled and imposes a linear age effect; it targets an interpretable average association and is not intended for extrapolation beyond the observed ages.

B.1.3 Variational fitting

Let b_d collect random effects and θ_d the fixed effects and variance parameters. Variational EM alternates maximization of

$$\mathcal{L}(q, \theta_d) = \mathbb{E}_q[\log p(y_d, b_d \mid \theta_d)] - \mathbb{E}_q[\log q(b_d)] \tag{4}$$

over the variational distribution q and over θ_d. The posterior SDs τ_i are conditional on the estimated hyperparameters.

B.1.4 Supported age effects and the reliability ceiling

Per-image support flags come from a permutation test on participant age labels: $691/900$ images show a supported age effect on valence and $59/900$ on arousal, and the dispersion of ΔH is correspondingly smaller for arousal (SD 0.056 vs. 0.106 scale points). Because ΔH is itself an estimate, even a model that knew the true image slopes could not correlate with it perfectly. The attenuation ceiling

$$r_{\max}^{(d)} = \sqrt{1 - \overline{\tau^2}^{(d)} \big/ \operatorname{Var}\big(\Delta H^{(d)}\big)} \tag{5}$$

is 0.86 for valence and 0.70 for arousal. We report observed correlations against it rather than against 1.

B.2 Elicitation Protocol and Models

B.2.1 Prompt

We send each image with the single user message below and no system message. We use the same text for all 17 models. The bracketed text is the only difference between conditions: the age sentence reads "18 to 29 years old", "30 to 59 years old" or "60 years old or older" for the three personas, and both bracketed passages are deleted for the baseline.

```
You are participating in a psychology rating study. [For
this rating, imagine that you are 60 years old or older and
respond from the perspective of someone in this age group.]
Analyze the provided image and return exactly one valid JSON
object with these four fields:
{"valence": number, "arousal": number, "description":
"short description of the emotion evoked by the image",
"emotion": "one of: joy, love, surprise, anger, sadness,
fear, disgust, neutral"}
Rate valence from 1 to 7, where 1 means extremely
unpleasant, 4 means neutral, and 7 means extremely pleasant.
Rate arousal from 1 to 7, where 1 means extremely calm, 4
means moderately activated, and 7 means extremely activated.
You may use one decimal place.
The description should briefly describe the primary
emotional response evoked by the image [from the perspective
of the specified age group]. Select the closest single
category for emotion from the provided list. Return JSON
only, with no explanation or additional text. If the image
cannot be viewed, return {"valence": null, "arousal":
null, "description": "image unavailable", "emotion":
"neutral"}.
```

B.2.2 Field order and label vocabulary

Fields are generated in the order valence, arousal, description, emotion. The two ratings are therefore produced before any generated text, whereas the label is conditioned on the model's own description; label shifts are properties of this joint elicitation and are not treated as independent confirmation of the rating shift. The vocabulary comprises the basic-level categories of Shaver et al. (1987), disgust (Ekman, 1992), and neutral; without neutral, a model with no emotional response to an image would be forced into a spurious category.

B.2.3 Rating resolution

With reference ages $(23.5, 44.5, 67.0)$ the trend in Eq. 9 weights the three ratings by $(-43, -1, 44)/1893$ times σ_a (Appendix B.3). If ratings are integers, the smallest non-zero $|\Delta M|$ is therefore about 0.30 scale points per SD of age, which exceeds the largest human $|\Delta H|$ on valence (0.26; mean 0.09). In pilot runs integer replies made the comparison vacuous; permitting one decimal place raised decimal usage to 94% and removed the artifact.

B.2.4 Models, decoding and execution

We access all models through one commercial routing API. We send each model $900 \times 4 = 3{,}600$ requests with temperature 0, a single completion and a 1,500-token output budget, chosen after observing 513-token mean completions from a verbose model against an initial 600-token cap. Temperature 0 removes sampling noise but not serving nondeterminism, and the protocol does not estimate run-to-run variance. We excluded reasoning-mode endpoints whose hidden token stream broke JSON-only output (one returned empty completions on 24 of 80 pilot requests with a length-truncation finish reason), retaining Qwen3-VL-8B-Thinking as a deliberate comparison. We store every response with its prompt hash, model identifier, decoding parameters, raw text and status in a content-addressed cache. This lets us resume interrupted runs without re-issuing completed requests. Cross-developer differences confound parameter count with data, architecture and alignment procedure, and we do not read them as scale effects.

B.3 Estimands and Derivations

B.3.1 Why list effects and image effects differ

Integrating over the participant intercept in the human-rating model gives

$$\mathbb{E}[y^{(d)} \mid i, z] = \alpha^{(d)}_{\ell(i)} + v^{(d)}_i + \big(\beta^{(d)}_{\ell(i)} + s^{(d)}_i\big) z. \tag{6}$$

Table 3: Evaluated models. The Qwen3-VL group (8B, 32B, 235B-A22B) is the controlled scale ladder: same developer, generation and instruction-tuning recipe. The 8B Thinking variant isolates inference-time reasoning at fixed parameter count.

Developer	Model	Nominal scale
Alibaba	Qwen3-VL-8B-Instruct	8B
Alibaba	Qwen3-VL-8B-Thinking	8B
Alibaba	Qwen3-VL-32B-Instruct	32B
Alibaba	Qwen3-VL-235B-A22B-Instruct	235B (22B active)
Alibaba	Qwen3.8-27B	27B
Google	Gemma-3-27B-IT	27B
Google	Gemma-4-31B-IT	31B
Google	Gemini-3-Flash	undisclosed
Google	Gemini-3.8-Flash	undisclosed
Meta	Llama-4-Scout	109B (17B active)
Z-AI	GLM-4.5V	106B (12B active)
OpenAI	GPT-5-mini	undisclosed
OpenAI	GPT-5.2	undisclosed
OpenAI	GPT-5.6-Luna-Pro	undisclosed
Anthropic	Claude-Sonnet-4.6	undisclosed
Anthropic	Claude-Opus-5	undisclosed
xAI	Grok-4.6	undisclosed

The target s_i omits the common list slope β_ℓ. Write the model's raw trend as $\Delta M_{im}^{\text{raw}} = b_{\ell(i),m} + q_{im}$, where $b_{\ell,m}$ is a list-level shift (including any uniform persona shift) and q_{im} the image-specific part. Centring within list on the retained images,

$$\Delta M_{im} = \Delta M_{im}^{\text{raw}} - \frac{1}{n_{\ell(i),m}} \sum_{j:\ell(j)=\ell(i)} \Delta M_{jm}^{\text{raw}} = q_{im} - \frac{1}{n_{\ell(i),m}} \sum_{j:\ell(j)=\ell(i)} q_{jm}, \tag{7}$$

so list-level shifts cancel exactly and ΔM matches the list-relative definition of ΔH. Without this step the two quantities are different estimands.

B.3.2 Partial correlation

Let B be the $n \times 3$ matrix with rows $(1, \bar{H}_i^{(V)}, (\bar{H}_i^{(V)})^2)$ and $R = I - B(B^\top B)^{-1} B^\top$. The headline statistic is

$$r_{\text{part}} = \frac{(R\,\Delta M)^\top (R\,\Delta H)}{\|R\,\Delta M\|_2 \, \|R\,\Delta H\|_2}. \tag{8}$$

We interpret this statistic as "correspondence beyond a quadratic valence control". It does not establish independence from all stimulus properties. Neither raw nor partial correlation measures agreement in magnitude: both are invariant to positive rescaling of one vector, which is why magnitudes are reported separately as contrasts (Appendix C.4).

B.3.3 Reference ages

For $(a_Y, a_M, a_O) = (23.5, 44.5, 67.0)$, $\bar{a}_G = 45$ and

$$\Delta M_{im}^{\text{raw}} = \frac{\sigma_a}{1893}\big(-43 M_{imY} - M_{imM} + 44 M_{imO}\big). \tag{9}$$

The middle persona receives almost no weight because its reference age lies at the mean of the three, so the trend is essentially a scaled older-minus-young contrast. The older band is open-ended; placing it at 67.0, the midpoint of the observed 60–74 range, keeps the model and human age units on the same participants. Because a correlation is invariant to the common scale factor, the choice of reference ages affects r_{part} only through the small middle-persona weight. Appendix C.3 shows that the response is in any case a step at 60 rather than a gradient.

B.4 Coverage, Failures and Robustness

B.4.1 Response coverage and failure causes

Of 61,200 scheduled requests, 61,042 returned a valid rating (99.7%; Table 4). Failures are reported by cause (Table 4): service faults, parse faults, explicit refusals, and silent empty completions. GLM-4.5V returned a well-formed response with an empty completion body for 96 image–condition pairs,

Table 4: Coverage and failure taxonomy. Each model was issued $900 \times 4 = 3{,}600$ requests. Failures are separated by cause: *service* (transport, timeout), *parse* (non-conforming output), *refusal* (an explicit decline), and *silent filtering* (HTTP 200 with an empty completion body). Only the last two are findings.

Model	Vendor	Scale	Valid	Cov.	Svc.	Parse	Refus.	Filt.	$
Grok-4.6	xAI	–	3,600	100.0%	0	0	0	0	21.42
Qwen3-VL-8B	Alibaba	8B	3,600	100.0%	0	0	0	0	0.90
Gemini-3.8-Flash	Google	–	3,600	100.0%	0	0	0	0	8.74
Gemma-4-31B	Google	31B	3,600	100.0%	0	0	0	0	0.25
Qwen3-VL-32B	Alibaba	32B	3,600	100.0%	0	0	0	0	0.67
Llama-4-Scout	Meta	109B	3,600	100.0%	0	0	0	0	0.50
Gemma-3-27B	Google	27B	3,599	100.0%	0	1	0	0	0.24
Qwen3-VL-235B	Alibaba	235B	3,599	100.0%	0	1	0	0	2.05
GPT-5.2	OpenAI	–	3,597	99.9%	3	0	0	0	8.15
Qwen3.8-27B	Alibaba	27B	3,597	99.9%	0	0	0	3	3.78
Claude-Sonnet-4.6	Anthropic	–	3,594	99.8%	4	0	2	0	9.34
Claude-Opus-5	Anthropic	–	3,594	99.8%	4	1	1	0	18.61
Gemini-3-Flash	Google	–	3,593	99.8%	1	6	0	0	3.16
Qwen3-VL-8B-Think	Alibaba	8B	3,592	99.8%	1	4	0	3	8.18
GPT-5.6-Luna-Pro	OpenAI	–	3,590	99.7%	9	1	0	0	4.40
GPT-5-mini	OpenAI	–	3,583	99.5%	6	6	0	5	2.62
GLM-4.5V	Z-AI	106B	3,504	97.3%	0	0	0	96	4.28
Total			**61,042**	**99.7%**	**28**	**20**	**3**	**107**	**97.30**

Silent filtering is concentrated in GLM-4.5V, where 92 of 96 affected pairs are sensitive-flagged images and the loss is near-uniform across conditions (26/24/24/22), so it reduces coverage without biasing the young–older contrast.

Table 5: Non-response by condition. Counts of requests that returned no valid rating (any failure class), split by prompt condition, with the number falling on sensitive-flagged images. p_{χ^2} tests uniformity over the four conditions (models with at least eight failures); $p_{O \text{ vs } Y}$ is an exact binomial test of older against young among their failures. *off-vocab* counts valid replies whose emotion label fell outside the eight permitted categories. No model's non-response is condition-asymmetric.

Model	base	young	middle	older	total	sens.	p_{χ^2}	$p_{O \text{ vs } Y}$	off-vocab Y/O
GLM-4.5V	26	24	24	22	96	92	0.95	0.88	28/29
GPT-5-mini	1	3	5	8	17	10	0.10	0.23	1/0
GPT-5.6-Luna-Pro	3	3	1	3	10	6	0.75	1.00	0/0
Qwen3-VL-8B-Think	1	3	3	1	8	6	0.57	–	0/0
Gemini-3-Flash	1	0	2	4	7	2	–	–	0/2
Claude-Opus-5	3	1	1	1	6	2	–	–	0/0
Claude-Sonnet-4.6	1	1	2	2	6	2	–	–	0/0
GPT-5.2	1	0	2	0	3	3	–	–	0/0
Qwen3.8-27B	0	1	1	1	3	0	–	–	0/0
Gemma-3-27B	1	0	0	0	1	1	–	–	0/0
Qwen3-VL-235B	0	0	0	1	1	1	–	–	0/0
Gemma-4-31B	0	0	0	0	0	0	–	–	0/0
Gemini-3.8-Flash	0	0	0	0	0	0	–	–	12/11
Llama-4-Scout	0	0	0	0	0	0	–	–	0/1
Grok-4.6	0	0	0	0	0	0	–	–	0/0
Qwen3-VL-8B	0	0	0	0	0	0	–	–	0/0
Qwen3-VL-32B	0	0	0	0	0	0	–	–	0/0
Total	38	36	41	43	158	125			41/43

92 of them on sensitive-flagged images, spread near-uniformly across conditions ($26/24/24/22$): content filtering that presents as an infrastructure fault. Of the 158 requests without a valid rating, 36 fell under the young persona and 43 under the older, and no model's failures depart from uniformity across conditions (smallest $p = 0.10$; Table 5). These checks do not suggest that differences in non-response across age conditions explain the persona effects.

B.4.2 Robustness on the common image subset

Models answered slightly different image sets. We recompute every valence statistic on the 846 images for which all 17 models returned a valid rating in all four conditions. No conclusion changes:

Table 6: Robustness to differing coverage: the headline valence statistic computed on each model's own valid images, and on the common subset of images rated by every model under every condition. Agreement between the two columns shows that cross-model comparisons are not driven by which images each model happened to answer.

Model	n all	r_{part}	n common	r_{part}	Δ
Claude-Opus-5	899	0.058	846	0.089	0.031
Claude-Sonnet-4.6	898	0.086	846	0.105	0.019
Gemini-3-Flash	894	0.065	846	0.084	0.019
Gemini-3.8-Flash	900	0.059	846	0.068	0.009
Gemma-3-27B	900	0.009	846	0.032	0.023
Gemma-4-31B	900	0.042	846	0.063	0.020
GLM-4.5V	873	0.093	846	0.088	-0.005
GPT-5-mini	884	0.037	846	0.060	0.023
GPT-5.2	898	-0.030	846	-0.012	0.018
GPT-5.6-Luna-Pro	894	0.006	846	0.033	0.028
Grok-4.6	900	-0.001	846	0.020	0.021
Llama-4-Scout	900	-0.017	846	-0.023	-0.006
Qwen3-VL-235B	899	0.072	846	0.096	0.023
Qwen3-VL-32B	900	0.002	846	0.016	0.013
Qwen3-VL-8B	900	0.001	846	0.003	0.002
Qwen3-VL-8B-Think	893	0.044	846	0.076	0.032
Qwen3.8-27B	897	0.024	846	0.048	0.024

The common subset is enriched for non-sensitive images relative to the full set, so it is the conservative comparison for effects that concentrate on sensitive content.

the largest change in any model's partial correlation is 0.032, and the two sets of values correlate at $r = 0.957$ (Table 6).

C Numerical Ratings and Human Alignment

We first compare no-age ratings with pooled human ratings. We then describe the size and shape of the persona shifts, compare them with human age differences, and examine the relation between valence and arousal. We keep pooled age contrasts separate from image-level alignment because they answer different questions.

C.1 Baseline agreement and calibration

C.1.1 Agreement with pooled human ratings

Concordance is lower than correlation for every model (CCC 0.731–0.892 for valence) because all models over-disperse (Figure 9); mean absolute error ranges from 0.50 to 1.03 points.

C.1.2 Agreement within content subsets

Competence is not lower on the flagged subset: baseline valence agreement is $r = 0.89$ within flagged images against 0.87 within unflagged ones (Table 9), so the persona shift on flagged images occurs despite strong baseline agreement on those images.

C.2 Changes under the age personas

C.2.1 Mean ratings and older-minus-young contrasts

We report mean ratings under all four conditions in Tables 10 and 11. Figure 10 shows the paired older-minus-young valence contrast for each model.

C.2.2 Frequency and size of rating changes

Between 26% and 78% of images change their valence rating at all between the young and older personas, so the shift is a broad re-rating rather than a few large moves; on average 46% of an image's

Table 7: Baseline affective competence, valence: the no-age condition against pooled human image means. Correlation measures preservation of image ordering; CCC, MAE and the calibration slope measure agreement in absolute level. A slope below 1 and an SD ratio below 1 indicate compression toward the scale midpoint.

Model	n	r	ρ	CCC	MAE	RMSE	Slope	SD ratio
Claude-Opus-5	897	0.923	0.932	0.877	0.574	0.708	1.182	1.281
Gemini-3.8-Flash	900	0.919	0.930	0.892	0.499	0.634	1.103	1.200
Gemini-3-Flash	899	0.912	0.924	0.867	0.599	0.756	1.241	1.360
Gemma-4-31B	900	0.908	0.920	0.867	0.601	0.750	1.222	1.345
Qwen3-VL-235B	900	0.902	0.903	0.827	0.696	0.864	1.184	1.312
Claude-Sonnet-4.6	899	0.900	0.908	0.858	0.627	0.783	1.227	1.363
GLM-4.5V	874	0.895	0.904	0.841	0.679	0.845	1.275	1.424
GPT-5.2	899	0.891	0.906	0.846	0.651	0.820	1.230	1.380
Grok-4.6	900	0.891	0.908	0.836	0.680	0.839	1.190	1.336
GPT-5.6-Luna-Pro	897	0.890	0.906	0.809	0.807	0.970	1.342	1.507
Qwen3-VL-8B	900	0.885	0.887	0.810	0.739	0.916	1.189	1.343
Llama-4-Scout	900	0.878	0.884	0.839	0.624	0.819	1.164	1.326
GPT-5-mini	899	0.877	0.903	0.797	0.845	0.999	1.313	1.497
Qwen3-VL-32B	900	0.866	0.883	0.775	0.847	1.040	1.235	1.427
Gemma-3-27B	899	0.863	0.885	0.816	0.731	0.908	1.211	1.403
Qwen3.8-27B	900	0.855	0.865	0.834	0.589	0.788	1.039	1.216
Qwen3-VL-8B-Think	899	0.837	0.840	0.731	1.027	1.240	1.389	1.660

Table 8: Baseline affective competence, arousal: the no-age condition against pooled human image means. Correlation measures preservation of image ordering; CCC, MAE and the calibration slope measure agreement in absolute level. A slope below 1 and an SD ratio below 1 indicate compression toward the scale midpoint.

Model	n	r	ρ	CCC	MAE	RMSE	Slope	SD ratio
Claude-Sonnet-4.6	899	0.709	0.696	0.571	0.954	1.167	1.291	1.820
Claude-Opus-5	897	0.706	0.698	0.550	0.952	1.162	1.186	1.679
GPT-5.2	899	0.700	0.696	0.573	0.912	1.123	1.207	1.726
Gemini-3.8-Flash	900	0.697	0.688	0.616	0.798	0.997	1.137	1.633
Qwen3-VL-32B	900	0.679	0.688	0.518	1.006	1.256	1.208	1.779
GPT-5.6-Luna-Pro	897	0.677	0.665	0.477	1.190	1.438	1.351	1.996
Gemini-3-Flash	899	0.676	0.670	0.557	0.992	1.191	1.268	1.875
GLM-4.5V	874	0.668	0.680	0.478	1.252	1.488	1.519	2.273
GPT-5-mini	899	0.667	0.663	0.512	1.040	1.262	1.194	1.790
Gemma-3-27B	899	0.666	0.676	0.510	1.028	1.272	1.204	1.807
Llama-4-Scout	900	0.665	0.638	0.543	1.007	1.229	1.280	1.925
Grok-4.6	900	0.661	0.658	0.516	1.046	1.276	1.240	1.876
Qwen3-VL-235B	900	0.638	0.635	0.482	1.089	1.345	1.200	1.880
Gemma-4-31B	900	0.638	0.645	0.547	0.929	1.143	1.125	1.763
Qwen3.8-27B	899	0.598	0.609	0.516	0.964	1.169	1.036	1.733
Qwen3-VL-8B	900	0.591	0.596	0.405	1.066	1.255	0.764	1.294
Qwen3-VL-8B-Think	899	0.526	0.525	0.362	1.494	1.750	1.242	2.361

ratings are identical under the two personas and 6.0% differ by more than one point (Tables 12 and 13).

C.2.3 Changes by sensitive flag and image category

By category, Person images carry the largest valence shift (-0.238, negative for all 17 models) and Object and Animal images none ($+0.059$, $+0.017$), while the arousal shift is present in every category (Table 15). Table 14 reports the shift by sensitive flag.

C.3 Shape of the age response

C.3.1 Which age persona the baseline resembles

The no-age prompt is closest to the middle persona for 13 of 17 models on valence (young for 4, older for none) and for 10 on arousal (Tables 16 and 17, Figure 11). The mean absolute distance from the baseline is 0.243 to the young rating, 0.236 to the middle rating and 0.351 to the older rating.

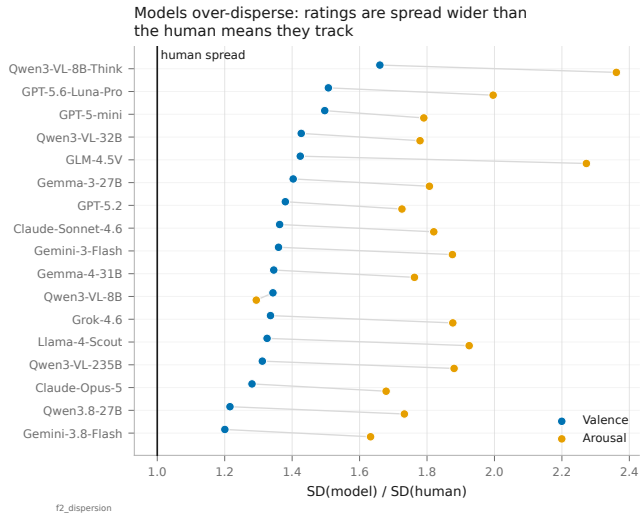


Figure 9: Ratio of model rating SD to human rating SD. All 17 models on both dimensions lie to the right of parity.

Table 9: Baseline valence agreement with human image means within content subsets: Pearson r on the 167 sensitive-flagged and 733 unflagged images and within each category, and the mean signed error (model minus human, rating points) on the two flag subsets. Correlations within a subset are lower than the pooled r because the range of human valence is narrower, not because the models are less competent there; the flagged subset is rated more negative than humans rate it by every model.

Model	sens.	non-sens.	Person	Scene	Object	Animal	bias sens.	bias non-sens.
Claude-Opus-5	0.94	0.91	0.92	0.93	0.93	0.93	+0.17	+0.32
Gemini-3.8-Flash	0.93	0.91	0.92	0.91	0.94	0.93	+0.17	+0.23
Gemini-3-Flash	0.92	0.90	0.91	0.92	0.94	0.92	-0.03	+0.18
Gemma-4-31B	0.91	0.89	0.90	0.91	0.93	0.92	-0.13	+0.15
Qwen3-VL-235B	0.90	0.89	0.90	0.91	0.92	0.89	+0.34	+0.49
Claude-Sonnet-4.6	0.92	0.88	0.90	0.91	0.90	0.91	-0.01	+0.08
GLM-4.5V	0.90	0.88	0.88	0.91	0.91	0.91	-0.16	+0.12
GPT-5.2	0.87	0.88	0.88	0.90	0.93	0.91	-0.07	+0.11
Grok-4.6	0.90	0.88	0.89	0.89	0.91	0.89	+0.22	+0.33
GPT-5.6-Luna-Pro	0.88	0.88	0.88	0.90	0.90	0.90	-0.00	+0.32
Qwen3-VL-8B	0.87	0.87	0.88	0.91	0.91	0.87	+0.33	+0.47
Llama-4-Scout	0.87	0.86	0.86	0.91	0.89	0.88	-0.02	+0.20
GPT-5-mini	0.88	0.86	0.87	0.89	0.90	0.88	+0.07	+0.33
Qwen3-VL-32B	0.89	0.85	0.87	0.88	0.87	0.88	+0.46	+0.48
Gemma-3-27B	0.86	0.84	0.86	0.86	0.89	0.88	-0.25	+0.03
Qwen3.8-27B	0.81	0.85	0.83	0.88	0.91	0.85	-0.07	-0.16
Qwen3-VL-8B-Think	0.81	0.83	0.82	0.87	0.87	0.82	+0.02	+0.31
Mean	**0.89**	**0.87**	**0.88**	**0.90**	**0.91**	**0.89**	**+0.06**	**+0.23**

Placing the baseline mean on the line through the three band means gives an implied age with median 54 years across the 12 models whose span allows the estimate.

C.3.2 Position of the middle persona

Among images whose valence differs between the young and older personas, the middle rating equals the young rating for 39% on average, lies strictly between the endpoints for 19%, and lies outside them for 14% (Table 18; arousal in Table 19; Figure 12). At the level of means, the middle band's position between young and older has a median of 0.16 on valence and 0.31 on arousal, where 0 places all change at the older persona. The persona effect is an old/not-old distinction taken at 60. This does not affect the alignment conclusion, which concerns the image-level pattern, but it means the trend of Eq. 9 should be read as an older-minus-young contrast.

Table 10: Mean valence by age condition, and the paired older-minus-young contrast on images rated in both. *Moved* is the share of images whose rating changed at all between the two conditions.

Model	Base	Young	Middle	Older	Δ_{O-Y}	d_z	Moved
Qwen3-VL-235B	4.80	4.88	4.85	4.66	-0.226***	-0.32	35%
Qwen3-VL-32B	4.81	4.77	4.77	4.55	-0.219***	-0.32	34%
Grok-4.6	4.64	4.76	4.68	4.59	-0.170***	-0.24	77%
Gemma-3-27B	4.31	4.37	4.45	4.21	-0.157***	-0.21	38%
GLM-4.5V	4.42	4.51	4.46	4.36	-0.143***	-0.17	72%
GPT-5-mini	4.61	4.74	4.70	4.59	-0.141***	-0.19	68%
Claude-Sonnet-4.6	4.40	4.48	4.43	4.36	-0.126***	-0.20	49%
Claude-Opus-5	4.62	4.64	4.64	4.53	-0.116***	-0.30	69%
Llama-4-Scout	4.49	4.48	4.48	4.39	-0.092***	-0.17	35%
GPT-5.2	4.41	4.41	4.41	4.32	-0.091***	-0.14	75%
GPT-5.6-Luna-Pro	4.60	4.65	4.60	4.57	-0.076***	-0.17	51%
Gemma-4-31B	4.43	4.56	4.57	4.51	-0.057**	-0.10	50%
Gemini-3.8-Flash	4.55	4.37	4.35	4.32	-0.045**	-0.09	78%
Gemini-3-Flash	4.47	4.58	4.56	4.55	-0.026	-0.04	60%
Qwen3-VL-8B	4.78	4.89	4.86	4.87	-0.020	-0.06	26%
Qwen3.8-27B	4.19	4.21	4.17	4.20	-0.006	-0.01	43%
Qwen3-VL-8B-Think	4.58	4.68	4.72	4.68	0.010	0.01	60%

Paired t-test, Benjamini–Hochberg adjusted across models within dimension. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 11: Mean arousal by age condition, and the paired older-minus-young contrast on images rated in both. *Moved* is the share of images whose rating changed at all between the two conditions.

Model	Base	Young	Middle	Older	Δ_{O-Y}	d_z	Moved
GPT-5-mini	4.23	4.41	4.28	3.89	-0.518***	-0.83	85%
GLM-4.5V	3.99	4.06	3.95	3.58	-0.486***	-0.44	87%
Qwen3-VL-8B-Think	4.11	4.32	4.16	3.88	-0.441***	-0.42	74%
Gemma-4-31B	3.73	3.90	3.73	3.47	-0.429***	-0.77	62%
Gemma-3-27B	4.23	4.31	4.31	3.91	-0.402***	-0.62	57%
Gemini-3-Flash	3.83	4.01	3.85	3.61	-0.395***	-0.59	72%
Grok-4.6	4.11	4.22	4.05	3.83	-0.389***	-0.71	85%
Qwen3-VL-8B	4.55	4.50	4.41	4.21	-0.296***	-0.83	51%
Qwen3-VL-32B	4.26	4.39	4.34	4.09	-0.295***	-0.56	42%
Gemini-3.8-Flash	3.81	4.00	3.93	3.73	-0.275***	-0.64	88%
Claude-Sonnet-4.6	4.05	4.14	4.03	3.86	-0.272***	-0.57	54%
Qwen3-VL-235B	4.23	4.30	4.25	4.06	-0.245***	-0.48	55%
GPT-5.2	4.08	4.17	4.05	3.94	-0.231***	-0.52	84%
Claude-Opus-5	4.25	4.24	4.19	4.03	-0.208***	-0.68	76%
GPT-5.6-Luna-Pro	4.35	4.39	4.35	4.20	-0.193***	-0.42	63%
Llama-4-Scout	3.75	3.78	3.71	3.61	-0.164***	-0.26	41%
Qwen3.8-27B	3.56	3.56	3.50	3.41	-0.157***	-0.29	52%

Paired t-test, Benjamini–Hochberg adjusted across models within dimension. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

C.4 Pooled human and model age contrasts

The image-level analysis compares list-centred quantities and therefore removes any global shift. We can estimate the global shift from OASIS because every image has at least two raters aged 60 or over on its list: the mean over images of the older-rater mean minus the young-rater mean is a pooled contrast over 25 (valence) and 16 (arousal) older raters, with a participant bootstrap stratified by list. It is not a per-image estimate and we do not use it as one.

Older OASIS raters are, if anything, slightly more positive than young raters: $+0.111$ overall ($[-0.05, +0.29]$) and $+0.139$ on unflagged images ($[-0.03, +0.32]$), with no difference on flagged images (-0.013). None of these pooled intervals excludes zero, and we make no claim that the human shift is reliably positive. The claim is comparative: on unflagged valence the model contrasts run from -0.134 to $+0.084$, and 10 of 17 fall below the human interval; only 6 models share the human sign there, and none does on arousal (Table 20, Figure 13). The three-band mean ratings show the same pattern: every model descends toward the older band while the human curve is flat or rises.

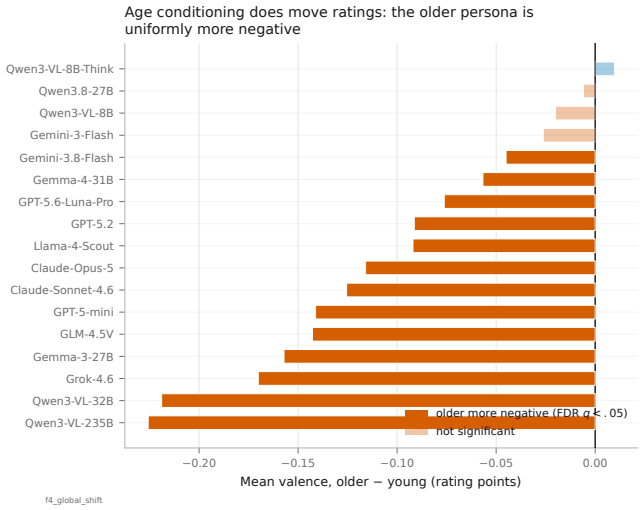


Figure 10: Mean valence under the older persona minus the young persona, paired on images rated in both conditions. Solid bars survive Benjamini–Hochberg correction.

Table 12: Distribution of the per-image older-minus-young change, valence. Columns 2–6 are the percentage of images whose absolute change falls in each band; *neg.* is the share of moved images that moved downward; *dec.* is the share of ratings using a decimal under the baseline and older conditions; SD is the across-image spread of ratings under each persona. The older persona is a re-rating of many images by small amounts, not a few large moves, and it does not compress the rating range.

Model	0	≤.25	≤.5	≤1	>1	neg.	mean|Δ|	max	dec. % ∅/O	SD Y/O
Qwen3-VL-8B	74	0	21	4	0	55	0.15	3.0	65/75	1.61/1.57
Qwen3-VL-32B	66	0	13	13	8	75	0.36	3.0	70/82	1.73/1.68
Qwen3-VL-235B	65	0	14	13	8	71	0.36	3.5	82/83	1.62/1.68
Llama-4-Scout	65	0	19	12	4	64	0.28	3.0	41/30	1.54/1.59
Gemma-3-27B	62	3	19	9	7	57	0.34	5.0	80/87	1.75/1.71
Qwen3.8-27B	57	2	32	7	3	46	0.27	3.0	49/52	1.52/1.50
Claude-Sonnet-4.6	51	3	23	18	5	56	0.36	3.5	84/80	1.64/1.71
Gemma-4-31B	50	6	26	12	6	52	0.32	3.7	51/72	1.64/1.70
GPT-5.6-Luna-Pro	49	9	34	6	3	59	0.27	3.0	53/43	1.87/1.82
Gemini-3-Flash	40	4	33	17	6	43	0.35	3.7	91/93	1.71/1.83
Qwen3-VL-8B-Think	40	0	40	10	10	50	0.50	5.0	33/47	1.96/1.91
GPT-5-mini	32	28	24	9	7	61	0.38	4.5	72/66	1.83/1.82
Claude-Opus-5	31	42	15	8	4	63	0.23	2.2	89/90	1.54/1.58
GLM-4.5V	28	21	26	14	11	58	0.49	4.5	78/71	1.72/1.76
GPT-5.2	25	35	23	9	7	47	0.37	3.4	87/86	1.73/1.75
Grok-4.6	23	32	24	13	9	58	0.41	4.0	83/89	1.61/1.69
Gemini-3.8-Flash	22	40	20	12	5	45	0.30	2.4	90/93	1.07/1.13

C.5 Image-level alignment with human age effects

C.5.1 Valence and arousal alignment

Table 21 and Table 22 give every alignment metric. On arousal the partial correlations are consistently negative, with 16 of 17 intervals excluding zero. Only $59/900$ images carry a supported human arousal effect. We therefore interpret the negative correlations cautiously: the arousal reference is weaker, and these results alone do not establish a reversed human age pattern.

C.5.2 Persona-shift strength and human alignment

We find no association between stereotype strength and alignment across models ($r = -0.038$, $p = 0.88$).

Table 13: Distribution of the per-image older-minus-young change, arousal. Columns 2–6 are the percentage of images whose absolute change falls in each band; *neg.* is the share of moved images that moved downward; *dec.* is the share of ratings using a decimal under the baseline and older conditions; SD is the across-image spread of ratings under each persona. The older persona is a re-rating of many images by small amounts, not a few large moves, and it does not compress the rating range.

Model	0	$\leq.25$	$\leq.5$	≤1	>1	neg.	mean$\lvert\Delta\rvert$	max	dec. % ∅/O	SD Y/O
Llama-4-Scout	59	0	18	17	7	68	0.38	3.0	64/70	1.59/1.58
Qwen3-VL-32B	58	0	21	18	4	90	0.35	3.0	25/17	1.44/1.33
Qwen3-VL-8B	49	0	40	10	1	97	0.31	1.5	58/43	1.07/1.01
Qwen3.8-27B	48	1	36	10	5	68	0.36	2.5	38/38	1.41/1.34
Claude-Sonnet-4.6	46	4	28	19	3	85	0.35	2.7	80/79	1.49/1.39
Qwen3-VL-235B	45	12	14	23	6	81	0.36	2.0	61/50	1.54/1.53
Gemma-3-27B	43	6	25	18	9	91	0.45	4.8	53/42	1.52/1.37
Gemma-4-31B	38	8	20	23	11	92	0.47	3.0	29/51	1.40/1.22
GPT-5.6-Luna-Pro	37	12	33	14	4	72	0.34	2.0	57/47	1.63/1.58
Gemini-3-Flash	28	4	27	29	12	82	0.52	3.5	96/96	1.52/1.56
Qwen3-VL-8B-Think	26	0	33	18	23	71	0.81	4.5	48/57	1.78/1.72
Claude-Opus-5	24	44	21	8	3	87	0.25	1.6	86/85	1.34/1.30
GPT-5.2	16	32	30	17	5	76	0.36	2.1	92/91	1.43/1.38
GPT-5-mini	15	16	27	25	16	88	0.59	3.5	74/66	1.45/1.36
Grok-4.6	15	22	29	21	12	82	0.48	2.7	79/86	1.56/1.50
GLM-4.5V	13	15	27	19	26	75	0.81	5.8	80/71	1.84/1.68
Gemini-3.8-Flash	12	30	35	17	5	85	0.38	2.7	90/95	0.89/0.83

Table 14: Valence age effect split by the OASIS sensitive-content flag (167 flagged, 733 unflagged images). A larger effect on the flagged subset indicates that age conditioning acts mainly where content is aversive.

Model	Δ_{O-Y} sens.	Δ_{O-Y} non-sens.	$\overline{\Delta M}$ sens.	$\overline{\Delta M}$ non-sens.
Qwen3.8-27B	-0.117	0.020	-0.033	0.006
Qwen3-VL-8B	-0.129	0.005	-0.036	0.001
GPT-5.6-Luna-Pro	-0.183	-0.052	-0.052	-0.015
Gemini-3.8-Flash	-0.250	0.002	-0.071	0.001
Llama-4-Scout	-0.284	-0.048	-0.080	-0.014
Qwen3-VL-8B-Think	-0.320	0.084	-0.084	0.024
Claude-Opus-5	-0.395	-0.052	-0.112	-0.015
Gemini-3-Flash	-0.405	0.060	-0.115	0.017
GLM-4.5V	-0.421	-0.089	-0.119	-0.025
Gemma-4-31B	-0.421	0.026	-0.119	0.007
GPT-5-mini	-0.446	-0.074	-0.113	-0.021
GPT-5.2	-0.469	-0.005	-0.133	-0.002
Gemma-3-27B	-0.476	-0.084	-0.135	-0.025
Claude-Sonnet-4.6	-0.486	-0.044	-0.137	-0.012
Grok-4.6	-0.565	-0.080	-0.159	-0.022
Qwen3-VL-32B	-0.626	-0.126	-0.177	-0.036
Qwen3-VL-235B	-0.632	-0.134	-0.178	-0.038

C.5.3 Model identity and the structure of image ratings

In the space of ratings, placing each of the 68 model×condition cells at its 900-dimensional vector of image ratings, the mean distance between two conditions of the same model is 15.1 and between different models 23.8 (ratio 1.58), and for 92.6% of cells the nearest neighbour is the same model under a different persona (Figure 15): changing the model moves the ratings further than changing the persona. The images themselves, embedded in the same space, organise by human valence and separate flagged content.

C.6 Valence–arousal geometry

Keil & Freund (2009) predict that with age the appetitive arm of the valence–arousal relation flattens and the aversive arm steepens. Using each model's own valence and arousal from the same completion (Table 23, Figure 16), the appetitive arm flattens under the older persona for 16 of 17 models, from a pooled slope of 0.48 ($[0.38, 0.57]$) to 0.21 ($[0.12, 0.30]$) arousal points per point of valence. The aversive arm does not steepen: pooled 0.79 against 0.76, with 9 models up and the rest down, and only

Table 15: Older-minus-young persona shift by OASIS content category (Person 346, Scene 220, Object 200, Animal 134 images). Stars: one-sample t-test, Benjamini–Hochberg over the 68 model×category tests within dimension. The human row is the pooled older-minus-young rater contrast of Table 20 restricted to the category. Person images, which hold most of the sensitive-flagged content, carry the largest valence shift; the arousal shift is present in every category.

	Valence				Arousal			
Model	Person	Scene	Object	Animal	Person	Scene	Object	Animal
Qwen3-VL-32B	-0.49***	-0.13***	+0.04	-0.06	-0.35***	-0.29***	-0.20***	-0.30***
Qwen3-VL-235B	-0.38***	-0.26***	-0.05	-0.03	-0.23***	-0.31***	-0.27***	-0.15***
Gemma-3-27B	-0.37***	-0.05	-0.01	+0.01	-0.54***	-0.36***	-0.30***	-0.26***
Grok-4.6	-0.36***	-0.18***	+0.10*	-0.07	-0.43***	-0.39***	-0.30***	-0.42***
GLM-4.5V	-0.35***	-0.17**	+0.10**	+0.05	-0.67***	-0.39***	-0.32***	-0.47***
GPT-5-mini	-0.30***	-0.13**	+0.01	+0.01	-0.58***	-0.53***	-0.48***	-0.39***
Claude-Sonnet-4.6	-0.28***	-0.18***	+0.09*	+0.04	-0.33***	-0.29***	-0.19***	-0.23***
GPT-5.2	-0.26***	-0.10*	+0.13***	+0.03	-0.28***	-0.17***	-0.22***	-0.23***
Claude-Opus-5	-0.23***	-0.12***	+0.03	-0.05*	-0.26***	-0.19***	-0.16***	-0.17***
Gemma-4-31B	-0.19***	-0.10**	+0.14***	+0.07*	-0.51***	-0.45***	-0.26***	-0.44***
Llama-4-Scout	-0.18***	-0.10*	+0.03	-0.03	-0.20***	-0.09*	-0.22***	-0.11*
Qwen3-VL-8B-Think	-0.18**	+0.06	+0.13*	+0.22**	-0.51***	-0.48***	-0.29***	-0.43***
Gemini-3.8-Flash	-0.13***	-0.06	+0.04	+0.07*	-0.28***	-0.30***	-0.21***	-0.31***
GPT-5.6-Luna-Pro	-0.12***	-0.09***	+0.00	-0.05	-0.23***	-0.13***	-0.21***	-0.16***
Gemini-3-Flash	-0.10*	-0.08	+0.10**	+0.06*	-0.45***	-0.39***	-0.31***	-0.37***
Qwen3.8-27B	-0.07*	-0.02	+0.09**	+0.03	-0.18***	-0.15***	-0.13***	-0.15***
Qwen3-VL-8B	-0.06*	-0.00	+0.02	-0.01	-0.33***	-0.33***	-0.24***	-0.25***
Model mean	**-0.24**	**-0.10**	**+0.06**	**+0.02**	**-0.37**	**-0.31**	**-0.25**	**-0.28**
OASIS raters	+0.23	-0.00	+0.01	+0.13	+0.02	+0.17	-0.23	+0.30

Table 16: Which persona the no-age baseline resembles, valence. $|b - X|$ is the mean absolute difference between the baseline rating and the rating under persona X on the same image; *nearest* is the smallest of the three. $p_{Y \text{ vs } M}$ is a Wilcoxon signed-rank test of $|b - Y|$ against $|b - M|$ over images. *Implied age* places the baseline mean on the line through the three band means (omitted when the young–older span is under 0.05 points). Nearest band across models: young 4, middle 13, older 0.

Model	n	$\lvert b-Y\rvert$	$\lvert b-M\rvert$	$\lvert b-O\rvert$	nearest	$p_{Y \text{ vs } M}$	$M-b$	implied age
Claude-Opus-5	897	0.118	0.102	0.198	middle	0.000***	+0.016	39
Claude-Sonnet-4.6	898	0.180	0.133	0.284	middle	0.000***	+0.035	55
Qwen3-VL-235B	899	0.170	0.146	0.342	middle	0.414	+0.052	45
Qwen3-VL-8B	900	0.171	0.163	0.230	middle	0.410	+0.087	–
Qwen3-VL-32B	900	0.181	0.173	0.389	middle	0.453	-0.036	24
Gemma-4-31B	900	0.221	0.204	0.335	middle	0.019*	+0.138	120
Gemini-3-Flash	893	0.209	0.210	0.338	young	0.195	+0.086	–
Qwen3.8-27B	897	0.249	0.224	0.260	middle	0.019*	-0.019	–
GPT-5.6-Luna-Pro	894	0.249	0.234	0.272	middle	0.169	+0.019	53
Llama-4-Scout	900	0.264	0.243	0.324	middle	0.153	-0.010	26
Grok-4.6	900	0.269	0.243	0.373	middle	0.080	+0.035	54
GPT-5.2	898	0.230	0.253	0.391	young	0.758	-0.006	31
Gemma-3-27B	899	0.185	0.254	0.330	young	0.001***	+0.141	55
GPT-5-mini	884	0.279	0.261	0.370	middle	0.844	+0.086	65
GLM-4.5V	873	0.279	0.276	0.464	middle	0.545	+0.052	56
Gemini-3.8-Flash	900	0.438	0.419	0.491	middle	0.037*	-0.199	–
Qwen3-VL-8B-Think	892	0.440	0.475	0.578	young	0.254	+0.140	–

8 models show both halves of the predicted pattern. The pooled human image means have arm slopes of 0.64 and 0.51; these are across-image slopes from disjoint rater groups, not the within-person coupling of the original study, and are shown for scale only.

D Shared Age Responses and Image Content

We next ask whether models share an image-specific age response and which image properties are associated with that response. We report agreement across models before examining the content predictors.

Table 17: Which persona the no-age baseline resembles, arousal. $|b - X|$ is the mean absolute difference between the baseline rating and the rating under persona X on the same image; *nearest* is the smallest of the three. $p_{Y \text{ vs } M}$ is a Wilcoxon signed-rank test of $|b - Y|$ against $|b - M|$ over images. *Implied age* places the baseline mean on the line through the three band means (omitted when the young–older span is under 0.05 points). Nearest band across models: young 7, middle 10, older 0.

Model	n	$\lvert b - Y \rvert$	$\lvert b - M \rvert$	$\lvert b - O \rvert$	nearest	$p_{Y \text{ vs } M}$	$M - b$	implied age
Claude-Opus-5	897	0.158	0.157	0.277	middle	0.508	-0.060	25
Claude-Sonnet-4.6	898	0.200	0.177	0.283	middle	0.090	-0.015	39
Qwen3-VL-8B	900	0.163	0.205	0.365	young	0.000^{***}	-0.137	20
Qwen3-VL-32B	900	0.240	0.239	0.334	middle	0.943	+0.084	47
Qwen3-VL-235B	899	0.237	0.250	0.350	young	0.255	+0.024	41
Gemma-4-31B	900	0.307	0.265	0.414	middle	0.000^{***}	+0.000	42
GPT-5.2	898	0.274	0.277	0.324	young	0.311	-0.020	41
Gemini-3-Flash	893	0.315	0.285	0.445	middle	0.019^{*}	+0.019	44
Gemma-3-27B	899	0.279	0.286	0.421	young	0.572	+0.079	39
GPT-5-mini	884	0.350	0.297	0.477	middle	0.000^{***}	+0.052	42
GPT-5.6-Luna-Pro	894	0.319	0.298	0.339	middle	0.067	-0.000	37
Qwen3.8-27B	895	0.340	0.323	0.379	middle	0.196	-0.058	26
Grok-4.6	900	0.331	0.333	0.433	young	0.459	-0.065	36
Llama-4-Scout	900	0.340	0.348	0.406	young	0.777	-0.039	33
Gemini-3.8-Flash	900	0.524	0.521	0.579	middle	0.747	+0.125	58
GLM-4.5V	873	0.533	0.539	0.769	young	0.411	-0.044	34
Qwen3-VL-8B-Think	892	0.754	0.745	0.795	middle	0.919	+0.050	46

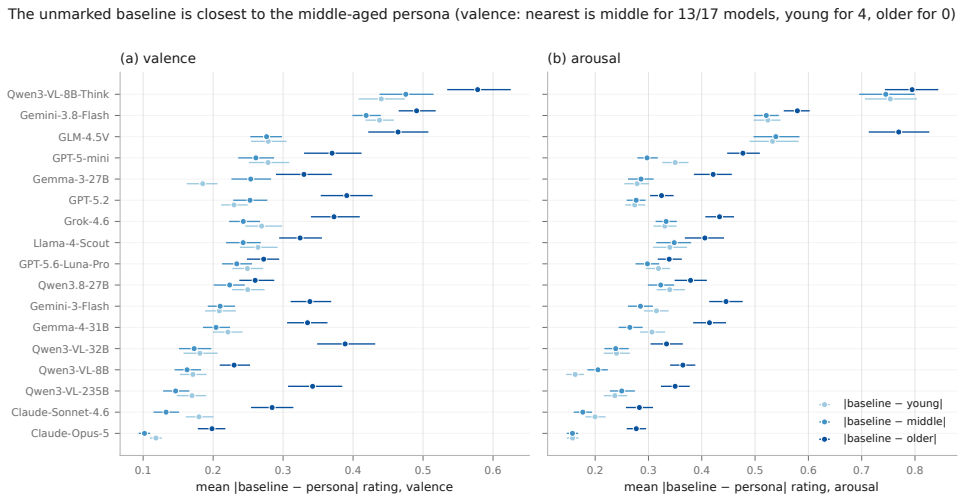


Figure 11: Mean absolute difference between the baseline rating and each persona's rating on the same image, with 95% image-bootstrap intervals.

D.1 Agreement across models

A principal component analysis of the image×model matrix puts 47.0% of the variance on a single component on which all 17 models load positively, with the second at 6.1% (Table 25). Clustering the pairwise correlations does not follow developers. The first component correlates -0.47 with human image arousal, -0.33 with the sensitive flag and 0.01 with human valence (Figure 18); a t-SNE of the images in the space of all 34 model age effects shows the same organisation by flagged content. On arousal the pairwise agreement is weaker (0.207, first component 27.4%) and the consensus is negatively related to the human effect (-0.267, $[-0.33, -0.21]$), subject to the caveat on the arousal reference in Appendix C.5.

D.2 Content predictors and sensitive images

Over the 900 images, the sensitive flag alone explains $R^2 = 0.11$ of the consensus valence shift, a quadratic in human valence alone 0.02, and the two together 0.16; adding human arousal and category raises the total to 0.43, with high-arousal images shifting more ($\beta = -0.252$ per SD) and Animal

Table 18: Shape of the three-band age response, valence. Among images whose rating differs between young and older (*moved*), *between* is the share with the middle rating strictly inside the two endpoints, step$_{M\to O}$ the share where middle equals young (all change occurs at the older persona), step$_{Y\to M}$ the share where middle equals older, and *outside* the share with middle beyond both. *bump* is the share of all images where young and older agree but middle differs. *position* is $(\bar{M}-\bar{Y})/(\bar{O}-\bar{Y})$ at the mean level (0: change confined to the older persona; 1: confined to the young persona; omitted when the span is under 0.03). r correlates the two successive steps over images.

Model	moved	between	step$_{M\to O}$	step$_{Y\to M}$	outside	bump	position	r
Gemma-3-27B	341	13%	37%	35%	15%	12%	-0.51	-0.22
Gemma-4-31B	447	20%	52%	26%	2%	5%	-0.03	0.15
Qwen3-VL-32B	310	10%	62%	26%	2%	3%	-0.00	0.04
Llama-4-Scout	319	10%	51%	33%	7%	10%	0.00	-0.32
Claude-Opus-5	619	29%	36%	25%	10%	11%	0.08	0.16
GPT-5.2	672	29%	23%	20%	28%	14%	0.12	-0.10
Qwen3-VL-235B	318	18%	52%	28%	3%	5%	0.16	0.03
GPT-5-mini	597	22%	30%	24%	24%	14%	0.21	-0.13
GLM-4.5V	631	30%	23%	18%	29%	12%	0.35	-0.01
Gemini-3.8-Flash	698	30%	23%	22%	24%	12%	0.36	-0.13
GPT-5.6-Luna-Pro	454	11%	35%	37%	17%	13%	0.41	-0.42
Claude-Sonnet-4.6	436	19%	41%	39%	1%	3%	0.41	0.20
Grok-4.6	691	29%	24%	20%	27%	12%	0.50	-0.07
Gemini-3-Flash	534	19%	42%	33%	6%	7%	–	0.06
Qwen3-VL-8B	231	3%	61%	35%	1%	6%	–	-0.25
Qwen3-VL-8B-Think	538	13%	37%	28%	21%	17%	–	-0.44
Qwen3.8-27B	384	10%	43%	33%	13%	10%	–	-0.33

Table 19: Shape of the three-band age response, arousal. Among images whose rating differs between young and older (*moved*), *between* is the share with the middle rating strictly inside the two endpoints, step$_{M\to O}$ the share where middle equals young (all change occurs at the older persona), step$_{Y\to M}$ the share where middle equals older, and *outside* the share with middle beyond both. *bump* is the share of all images where young and older agree but middle differs. *position* is $(\bar{M}-\bar{Y})/(\bar{O}-\bar{Y})$ at the mean level (0: change confined to the older persona; 1: confined to the young persona; omitted when the span is under 0.03). r correlates the two successive steps over images.

Model	moved	between	step$_{M\to O}$	step$_{Y\to M}$	outside	bump	position	r
Gemma-3-27B	515	20%	44%	22%	14%	14%	0.00	-0.22
Qwen3-VL-32B	382	11%	60%	26%	3%	5%	0.15	-0.13
Qwen3-VL-235B	498	12%	50%	27%	11%	9%	0.21	-0.28
Claude-Opus-5	681	25%	40%	21%	14%	10%	0.22	-0.06
GPT-5.6-Luna-Pro	564	17%	29%	28%	26%	15%	0.23	-0.43
GLM-4.5V	756	34%	14%	12%	40%	9%	0.24	-0.37
GPT-5-mini	747	37%	21%	12%	30%	8%	0.25	-0.20
Gemini-3.8-Flash	790	38%	19%	15%	28%	7%	0.26	-0.35
Qwen3-VL-8B	462	6%	60%	31%	3%	5%	0.31	-0.35
Qwen3-VL-8B-Think	663	18%	29%	22%	31%	16%	0.35	-0.44
Claude-Sonnet-4.6	484	21%	38%	38%	3%	4%	0.38	-0.05
Qwen3.8-27B	464	14%	33%	37%	16%	16%	0.39	-0.41
Gemini-3-Flash	641	28%	33%	26%	13%	7%	0.40	-0.05
Gemma-4-31B	557	22%	41%	32%	5%	5%	0.41	-0.12
Llama-4-Scout	372	8%	48%	34%	9%	13%	0.43	-0.46
Grok-4.6	769	34%	16%	13%	37%	10%	0.44	-0.32
GPT-5.2	753	31%	15%	21%	33%	10%	0.48	-0.31

images less than Person images (Table 26). The flag coefficient is -0.185 points ($[-0.271, -0.099]$) after all controls. Per model, the full set explains between 0.07 and 0.38 of the variance (median 0.24; Table 27). Figure 19 shows the raw pattern: a band of negative shift across the flagged columns and near-zero elsewhere, repeated in every row.

On arousal, the human subset contrasts point in the direction Keil & Freund (2009) predict, higher for unpleasant images ($+0.451$ flagged, $+0.211$ unflagged) and lower for pleasant ones (-0.266, -0.075), although every such interval includes zero; every model lowers arousal in all four cells.

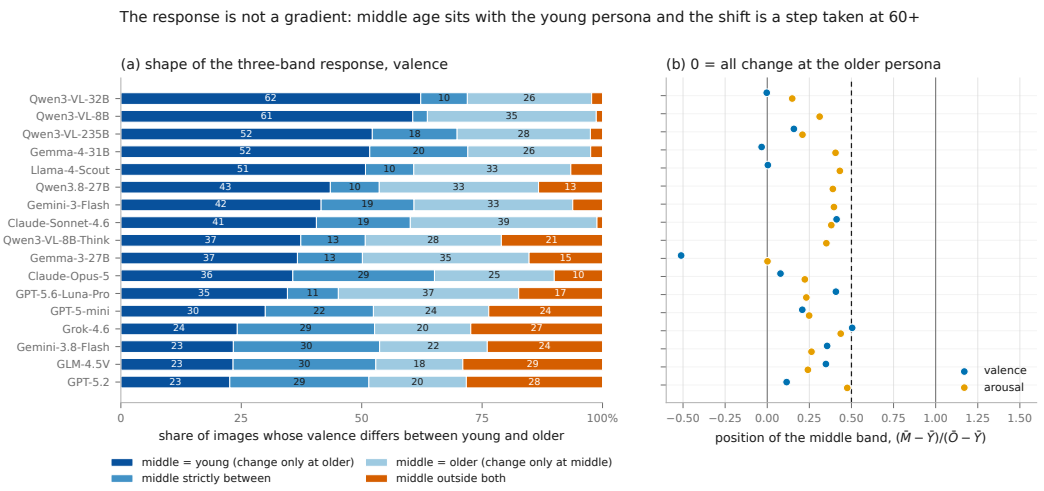


Figure 12: Shape of the three-band response. (a) Among images that move between young and older, where the middle rating lies. (b) The position of the middle band between the other two at the level of means.

Table 20: Older-minus-young shift in the human raters and in every model, pooled over images. The human contrast is the mean over images of the mean rating given by raters aged 60 or over minus that given by raters aged 18–29 on the same image; the interval is a participant bootstrap stratified by list (1,000 replicates). Model rows are the paired older-minus-young persona contrast on the same images. The human contrast is small and its interval includes zero, but its sign is positive on unflagged images and on arousal; the models are negative on every subset, and most fall below the lower end of the human interval.

	Valence, older − young			Arousal, older − young		
Source	all	sens.	non-sens.	all	sens.	non-sens.
OASIS raters	**+0.111**	**-0.013**	**+0.139**	**+0.041**	**+0.220**	**+0.000**
95% CI	[-0.05, 0.29]	[-0.23, 0.22]	[-0.03, 0.32]	[-0.35, 0.48]	[-0.26, 0.78]	[-0.39, 0.42]
Qwen3-VL-235B	-0.226***	-0.632***	-0.134***	-0.245***	-0.276***	-0.238***
Qwen3-VL-32B	-0.219***	-0.626***	-0.126***	-0.295***	-0.451***	-0.259***
Grok-4.6	-0.170***	-0.565***	-0.080**	-0.389***	-0.576***	-0.346***
Gemma-3-27B	-0.157***	-0.476***	-0.084**	-0.402***	-0.822***	-0.306***
GLM-4.5V	-0.143***	-0.421***	-0.089**	-0.486***	-0.722***	-0.441***
GPT-5-mini	-0.132***	-0.401***	-0.073**	-0.512***	-0.666***	-0.479***
Claude-Sonnet-4.6	-0.126***	-0.486***	-0.044*	-0.272***	-0.452***	-0.231***
Claude-Opus-5	-0.116***	-0.395***	-0.052***	-0.208***	-0.371***	-0.171***
Llama-4-Scout	-0.092***	-0.284***	-0.048**	-0.164***	-0.263***	-0.141***
GPT-5.2	-0.091***	-0.472***	-0.005	-0.232***	-0.344***	-0.207***
GPT-5.6-Luna-Pro	-0.076***	-0.183***	-0.052**	-0.193***	-0.355***	-0.157***
Gemma-4-31B	-0.057**	-0.421***	+0.026	-0.429***	-0.678***	-0.373***
Gemini-3.8-Flash	-0.045**	-0.250***	+0.002	-0.275***	-0.351***	-0.258***
Gemini-3-Flash	-0.028	-0.410***	+0.059**	-0.395***	-0.556***	-0.359***
Qwen3-VL-8B	-0.020	-0.129***	+0.005	-0.296***	-0.423***	-0.268***
Qwen3.8-27B	-0.005	-0.117*	+0.020	-0.156***	-0.190***	-0.148***
Qwen3-VL-8B-Think	+0.015	-0.296***	+0.084**	-0.442***	-0.619***	-0.403***
Models with the human sign	1/17	17/17	6/17	0/17	0/17	0/17
Models below the human 95% CI	12/17	14/17	10/17	7/17	16/17	3/17

Human rater counts: valence 139 / 249 / 25 and arousal 140 / 252 / 16 in the young / middle / older bands; every image has at least two older raters on its list, so the pooled contrast is estimable even though per-image older means are not. Model stars: one-sample t-test on the paired difference, Benjamini–Hochberg within column; $^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$.

E Emotion Labels, Descriptions, and Model Comparisons

We examine how age conditioning changes emotion labels and descriptions, then relate these changes to numerical ratings and model properties. As described in Appendix B.2, the ratings precede the description and label in each response. We therefore interpret these outputs as parts of the same elicitation.

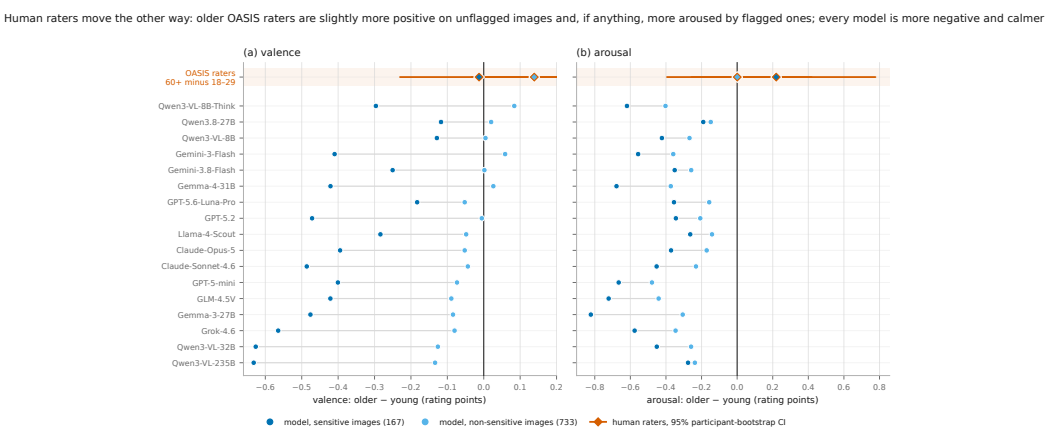


Figure 13: Older-minus-young contrast in OASIS raters (diamonds, 95% participant bootstrap) and in each model (circles), on the 167 flagged and 733 unflagged images, for valence (a) and arousal (b). The models sit below the human interval on almost every row.

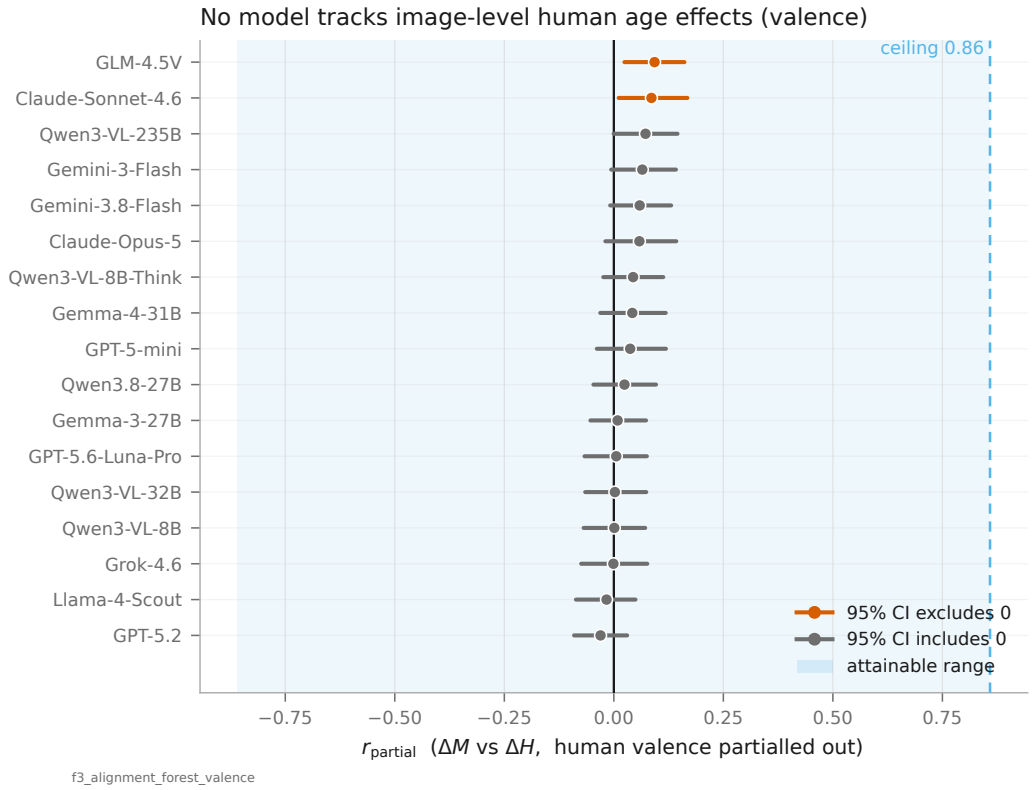


Figure 14: Image-level alignment between each model's valence age effect and the human age effect, with image pleasantness partialled out. Bars are 95% image-bootstrap intervals; the shaded band is the attainable range implied by the posterior SD of the human estimate.

E.1 Emotion labels

E.1.1 Transitions between emotion groups

Restricting to images with a valid in-vocabulary label under both personas, between 105 and 208 images change label between young and older (median 163 of roughly 890 paired images). The aversive shift is significant after correction for 13 of 17 models and the appetitive loss for 6 of 17 (Table 28); Grok-4.6 gains a net 61 aversive images, Qwen3-VL-235B 53 and Claude-Opus-5 43.

Table 21: Age-shift alignment, valence: the model age effect ΔM against the human age effect ΔH, both in rating points per $+1$ SD of rater age. r_{part} removes a quadratic function of pooled human valence from both vectors and is the headline quantity; $r_{\max}$ is the attenuation ceiling implied by the posterior SD of ΔH.

Model	n	r	r_{part}	95% CI	$r_{\max}$	Flat	Sign	SD(ΔM)
GLM-4.5V	873	0.030	0.093	[0.02, 0.16]	0.857	4%	0.53	0.240
Claude-Sonnet-4.6	898	0.121	0.086	[0.01, 0.17]	0.859	0%	0.55	0.176
Qwen3-VL-235B	899	0.066	0.072	[-0.00, 0.15]	0.859	0%	0.56	0.195
Gemini-3-Flash	894	0.196	0.065	[-0.01, 0.14]	0.859	21%	0.60	0.166
Gemini-3.8-Flash	900	0.098	0.059	[-0.01, 0.13]	0.859	9%	0.55	0.134
Claude-Opus-5	899	0.093	0.058	[-0.02, 0.14]	0.859	11%	0.54	0.109
Qwen3-VL-8B-Think	893	-0.065	0.044	[-0.02, 0.11]	0.859	14%	0.45	0.225
Gemma-4-31B	900	0.044	0.042	[-0.03, 0.12]	0.859	16%	0.54	0.165
GPT-5-mini	884	-0.035	0.037	[-0.04, 0.12]	0.859	4%	0.47	0.208
Qwen3.8-27B	897	-0.013	0.024	[-0.05, 0.10]	0.859	30%	0.49	0.131
Gemma-3-27B	900	-0.093	0.009	[-0.05, 0.07]	0.859	1%	0.45	0.211
GPT-5.6-Luna-Pro	894	-0.106	0.006	[-0.07, 0.08]	0.857	0%	0.44	0.129
Qwen3-VL-32B	900	-0.096	0.002	[-0.07, 0.07]	0.859	0%	0.43	0.193
Qwen3-VL-8B	900	-0.071	0.001	[-0.07, 0.07]	0.859	73%	0.52	0.092
Grok-4.6	900	0.037	-0.001	[-0.07, 0.08]	0.859	1%	0.53	0.198
Llama-4-Scout	900	0.081	-0.017	[-0.09, 0.05]	0.859	0%	0.52	0.148
GPT-5.2	898	-0.047	-0.030	[-0.09, 0.03]	0.858	7%	0.51	0.188

CI by resampling images, 2,000 replicates. *Flat* is the share of images with $|\Delta M| < 0.01$.

Table 22: Age-shift alignment, arousal: the model age effect ΔM against the human age effect ΔH, both in rating points per $+1$ SD of rater age. r_{part} removes a quadratic function of pooled human valence from both vectors and is the headline quantity; $r_{\max}$ is the attenuation ceiling implied by the posterior SD of ΔH.

Model	n	r	r_{part}	95% CI	$r_{\max}$	Flat	Sign	SD(ΔM)
Llama-4-Scout	900	-0.028	-0.016	[-0.08, 0.04]	0.696	0%	0.53	0.170
Qwen3-VL-235B	899	-0.032	-0.067	[-0.13, -0.01]	0.696	2%	0.51	0.136
GPT-5.6-Luna-Pro	894	-0.062	-0.082	[-0.14, -0.02]	0.694	5%	0.48	0.124
Qwen3-VL-8B-Think	893	-0.086	-0.083	[-0.15, -0.02]	0.696	2%	0.47	0.278
Gemini-3-Flash	894	-0.104	-0.084	[-0.15, -0.01]	0.696	5%	0.48	0.178
Qwen3-VL-8B	900	-0.128	-0.108	[-0.16, -0.05]	0.696	0%	0.47	0.095
Qwen3.8-27B	896	-0.067	-0.108	[-0.16, -0.05]	0.696	0%	0.48	0.144
GPT-5-mini	884	-0.084	-0.108	[-0.17, -0.05]	0.695	8%	0.47	0.164
GPT-5.2	898	-0.100	-0.118	[-0.18, -0.06]	0.696	8%	0.48	0.117
Claude-Opus-5	899	-0.086	-0.132	[-0.19, -0.07]	0.696	23%	0.47	0.082
Gemini-3.8-Flash	900	-0.141	-0.149	[-0.23, -0.07]	0.696	11%	0.44	0.115
GLM-4.5V	873	-0.164	-0.164	[-0.22, -0.11]	0.695	4%	0.42	0.292
Grok-4.6	900	-0.154	-0.167	[-0.23, -0.10]	0.696	8%	0.46	0.147
Qwen3-VL-32B	900	-0.170	-0.190	[-0.25, -0.13]	0.696	0%	0.45	0.139
Gemma-3-27B	900	-0.145	-0.223	[-0.28, -0.17]	0.696	2%	0.48	0.172
Claude-Sonnet-4.6	898	-0.198	-0.236	[-0.30, -0.17]	0.696	6%	0.43	0.127
Gemma-4-31B	900	-0.283	-0.289	[-0.35, -0.23]	0.696	1%	0.42	0.148

CI by resampling images, 2,000 replicates. *Flat* is the share of images with $|\Delta M| < 0.01$.

E.1.2 Changes in individual labels

Per label (Table 29, Figure 21), the gain is carried by sadness ($+271$) and fear ($+184$; positive for 14 of 17 models), with a smaller gain for disgust ($+96$) and a loss for anger (-44); surprise (-206), joy (-133) and neutral (-111) lose most.

E.1.3 Label diversity and composition

The label distribution does not narrow: the median Jensen–Shannon divergence between young and older is 0.004 bits, against 0.001 between young and middle (Table 30).

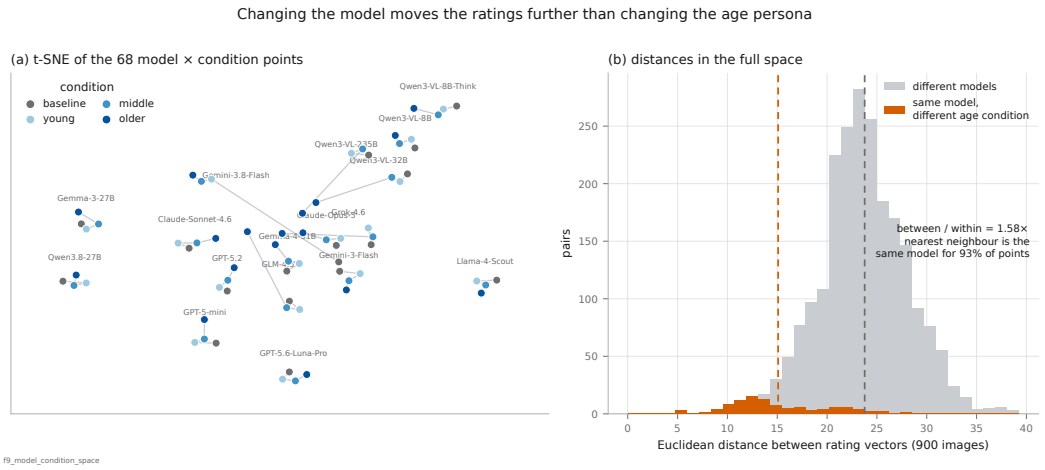


Figure 15: Each point is one model under one condition, positioned by its 900 image ratings. Changing the model moves the ratings further than changing the persona: the nearest neighbour is the same model for 92.6% of points.

Table 23: Affective geometry by persona. For each model and condition, arousal is regressed on distance from the neutral point separately on the pleasant side (*appetitive arm*, images with valence > 4) and the unpleasant side (*aversive arm*, valence < 4), using that model's own valence and arousal from the same completion. Keil & Freund (2009) predict that with age the appetitive arm flattens and the aversive arm steepens; *K&F* marks models showing both. The human row uses the pooled OASIS image means, which come from disjoint rater groups and are therefore an across-image geometry rather than a within-rater one.

	Appetitive arm			Aversive arm			
Model	young	older	Δ	young	older	Δ	K&F
Claude-Opus-5	0.19	-0.03	-0.21	0.81	0.70	-0.11	no
Claude-Sonnet-4.6	0.65	0.20	-0.45	0.86	0.85	-0.01	no
Gemini-3-Flash	0.55	0.23	-0.33	0.92	0.91	-0.01	no
Gemini-3.8-Flash	0.38	0.06	-0.32	0.99	0.72	-0.28	no
Gemma-3-27B	0.34	0.23	-0.12	1.48	1.25	-0.22	no
Gemma-4-31B	0.20	0.03	-0.18	0.50	0.64	+0.14	yes
GLM-4.5V	0.79	0.24	-0.55	0.70	0.87	+0.17	yes
GPT-5-mini	0.79	0.48	-0.31	1.26	1.31	+0.06	yes
GPT-5.2	0.61	0.30	-0.30	1.00	1.03	+0.03	yes
GPT-5.6-Luna-Pro	0.42	0.43	+0.01	0.99	1.05	+0.06	no
Grok-4.6	0.56	0.19	-0.37	0.78	1.00	+0.21	yes
Llama-4-Scout	0.57	0.33	-0.24	0.23	0.09	-0.14	no
Qwen3-VL-235B	0.41	0.24	-0.18	0.73	0.76	+0.03	yes
Qwen3-VL-32B	0.38	0.08	-0.31	0.85	0.73	-0.12	no
Qwen3-VL-8B	0.22	0.16	-0.06	0.40	0.39	-0.01	no
Qwen3-VL-8B-Think	0.63	0.35	-0.28	0.88	1.02	+0.14	yes
Qwen3.8-27B	0.69	0.51	-0.18	0.90	0.96	+0.07	yes
OASIS raters (pooled means)	0.64			0.51			

Models showing the Keil–Freund pattern: 8/17.

E.2 Emotion descriptions

E.2.1 Description length and word use

Pooled over models, the older persona lengthens the description (from 14.0 to 16.1 words, significant for 16 of 17 models) and changes its register (Table 31, Figure 22). Memory language reaches 88.0% of older-persona descriptions for Gemma-3-27B; explicit references to the rater's own age or generation rise from 1.6% to 6.9%, and first-person framing from 5.3% to 15.6%. The excitement lexicon falls for all 17 models. A weighted log-odds analysis with an informative Dirichlet prior (Monroe et al., 2008), which does not depend on our lexica, returns the same picture (Table 32): the words most diagnostic of the young persona are *kinda, cute, vibe, excitement, really, intense*.

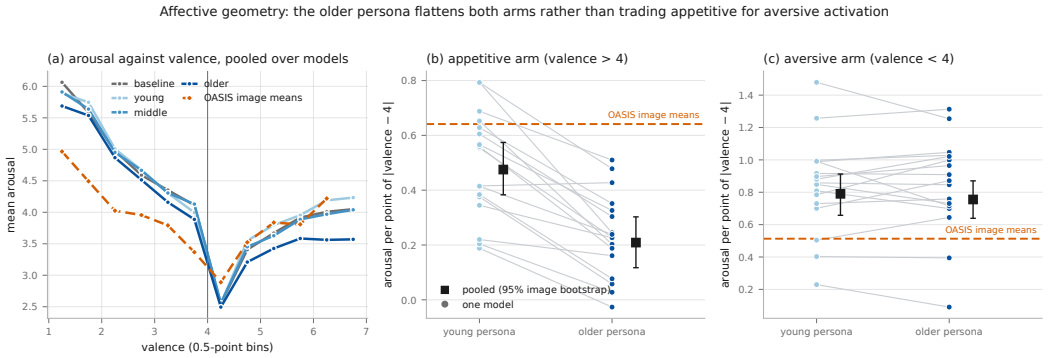


Figure 16: Arousal against valence by condition, pooled over models, with the OASIS image means (a); appetitive (b) and aversive (c) arm slopes under the young and older personas, one line per model, pooled estimate with 95% image-bootstrap interval.

Table 24: The content gate resolved: the older-minus-young shift on the sensitive flag crossed with pooled human valence (unpleasant < 4, pleasant ≥ 4). Flagged-unpleasant images are mostly violence, injury and death ($n = 113$); flagged-pleasant images are mostly nudity and erotic content (nude man, nude woman, nude couple, BDSM, massage) with a few cemetery and memorial scenes ($n = 54$). Unflagged images: unpleasant $n = 194$, pleasant $n = 539$. The largest persona effect anywhere in the study is on the flagged-pleasant images, which the older persona rates close to a full point less pleasant; unpleasant images move far less whether flagged or not. The older human raters rate the flagged-pleasant images *more* pleasant than the young raters do.

	Valence, older − young				Arousal, older − young			
Source	unfl. unpl.	unfl. pleas.	flag. unpl.	flag. pleas.	unfl. unpl.	unfl. pleas.	flag. unpl.	flag. pleas.
OASIS raters	**-0.107**	**+0.227**	**-0.231**	**+0.443**	**+0.211**	**-0.075**	**+0.451**	**-0.266**
95% CI	[-0.36, 0.20]	[0.02, 0.43]	[-0.47, 0.00]	[0.10, 0.86]	[-0.27, 0.69]	[-0.46, 0.38]	[-0.15, 1.14]	[-0.78, 0.22]
Qwen3-VL-235B	-0.141***	-0.131***	-0.333***	-1.270***	-0.062	-0.301***	-0.173***	-0.496***
Qwen3-VL-32B	-0.038	-0.158***	-0.213***	-1.491***	-0.197***	-0.281***	-0.287***	-0.796***
Grok-4.6	-0.181***	-0.044	-0.354***	-1.006***	-0.155***	-0.415***	-0.319***	-1.113***
Gemma-3-27B	+0.026	-0.124***	-0.078	-1.309***	-0.325***	-0.299***	-0.463***	-1.574***
GLM-4.5V	-0.090*	-0.089*	-0.158*	-1.085***	-0.403***	-0.455***	-0.401**	-1.532***
GPT-5-mini	-0.039	-0.086**	-0.163*	-0.910***	-0.285***	-0.548***	-0.515***	-0.988***
Claude-Sonnet-4.6	-0.151***	-0.005	-0.247***	-0.996***	-0.109***	-0.275***	-0.206***	-0.977***
Claude-Opus-5	-0.101***	-0.035*	-0.188***	-0.828***	-0.063***	-0.210***	-0.149***	-0.837***
Llama-4-Scout	-0.126**	-0.020	-0.232**	-0.393***	-0.161**	-0.134***	-0.234**	-0.326**
GPT-5.2	+0.018	-0.014	-0.155*	-1.122***	-0.130***	-0.235***	-0.247***	-0.544***
GPT-5.6-Luna-Pro	+0.026	-0.080***	-0.029	-0.496***	-0.084*	-0.183***	-0.212***	-0.646***
Gemma-4-31B	-0.030	+0.047	-0.204**	-0.876***	-0.366***	-0.375***	-0.492***	-1.067***
Gemini-3.8-Flash	-0.036	+0.015	-0.134**	-0.494***	-0.171***	-0.289***	-0.258***	-0.544***
Gemini-3-Flash	-0.103***	+0.117***	-0.262***	-0.713***	-0.107**	-0.450***	-0.141*	-1.409***
Qwen3-VL-8B	+0.049	-0.011	-0.027	-0.343***	-0.284***	-0.262***	-0.448***	-0.370***
Qwen3.8-27B	+0.044	+0.012	-0.031	-0.298**	-0.065	-0.178***	-0.111*	-0.357**
Qwen3-VL-8B-Think	+0.209**	+0.039	-0.098	-0.740***	-0.320***	-0.433***	-0.558***	-0.756***
Models with the human sign	11/17	5/17	17/17	0/17	0/17	17/17	0/17	17/17
Models below the human 95% CI	0/17	14/17	0/17	17/17	6/17	1/17	15/17	9/17

Human intervals are wide on the flagged subsets because only 25 (valence) and 16 (arousal) raters are aged 60 or over. Stars as in Table 20.

E.2.2 Structure of the description embeddings

A t-SNE of TF-IDF description embeddings is organised first by the emotion label that each model assigned; within it the older-persona texts form sub-clusters that coincide with the memory lexicon. The share of a description's ten nearest neighbours (in the 50-dimensional embedding) that come from its own condition is 37% for older, 25% for middle, 31% for young and 39% for the baseline, against 25% chance. Placing the 68 model×condition cells by their description centroids shows that model identity dominates the text channel as it does the numeric one: the nearest neighbour is the same model for 78% of cells.

E.2.3 Associations between descriptions and rating changes

Regressing each model's older-minus-young valence on the image covariates of Table 26 plus the change in each text feature (Tables 33 and 34), the arrival of memory language raises the older rating (median $\beta = +0.14$, positive and significant for 13 of 17 models), vulnerability language lowers it (-0.17, negative for 12), and retained excitement raises it ($+0.40$, positive for all 17). Adding

Table 25: Between-model agreement on the image-specific age effect ΔM (list-centred). The pairwise correlation is taken over the 136 model pairs and split by developer with a permutation test on the developer labels. PC1 and PC2 are the first two components of the image×model matrix. The split-half reliability is that of the 17-model consensus ΔM (Spearman–Brown, median of 500 random splits of the models). The consensus is reliable and tracks human image arousal and the sensitive flag, not human valence, and its partial correlation with the human age effect (image bootstrap CI) is as null as any single model's.

	Valence	Arousal
Images	897	897
Mean pairwise r over 136 model pairs	0.428	0.207
range	[0.16, 0.71]	[0.01, 0.52]
same-developer pairs	0.424	0.203
cross-developer pairs	0.429	0.207
permutation p (developer labels)	0.54	0.53
PC1 variance share	47.0%	27.4%
PC2 variance share	6.1%	6.6%
Models loading positively on PC1	17/17	17/17
Split-half reliability of the consensus	0.93	0.79
r(PC1, human valence)	0.01	-0.27
r(PC1, human arousal)	-0.47	-0.37
r(PC1, sensitive flag)	-0.33	-0.27
η^2(PC1, category)	0.08	0.02
r(PC1, ΔH), valence partialled	0.05	-0.27
Consensus ΔM: r_{part} with ΔH	0.050 [-0.02, 0.13]	-0.267 [-0.33, -0.21]

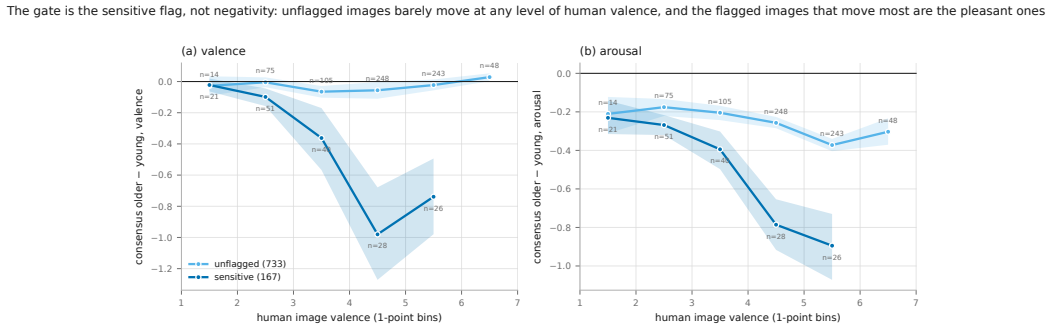


Figure 17: Consensus older-minus-young valence shift by one-point bin of pooled human valence, for flagged and unflagged images separately, with 95% image-bootstrap bands. Unflagged images barely move at any valence level; flagged images move most where humans rate them as pleasant.

the text features raises R^2 by a median of 0.10 and up to 0.24. These are associations within one generation, not causal effects.

E.3 Model-level associations and within-developer comparisons

E.3.1 Model properties and persona shifts

We examine model-level correlates ($n = 17$, exploratory; Table 35). The size of a model's valence shift is unrelated to baseline competence ($\rho = -0.07$, $p = 0.78$) and over-dispersion ($\rho = -0.09$, $p = 0.72$), but strongly related to its memory rate ($\rho = -0.76$), description lengthening ($\rho = -0.73$) and age references ($\rho = -0.65$).

E.3.2 Scale, reasoning, generation, and model tier

We see the same pattern in the controlled comparisons (Table 36): the Qwen3-VL step from 8B to 32B takes the memory rate from 7.4% to 63.8%, and the Gemma-3-27B to Gemma-4-31B step moves it from 88.0% to 22.7% and the shift from -0.157 to -0.057. Qwen3-VL-8B-Thinking has no valence shift ($+0.010$ against -0.020) but a larger arousal shift (-0.441 against -0.296), and is the one model that gains appetitive labels under the older persona.

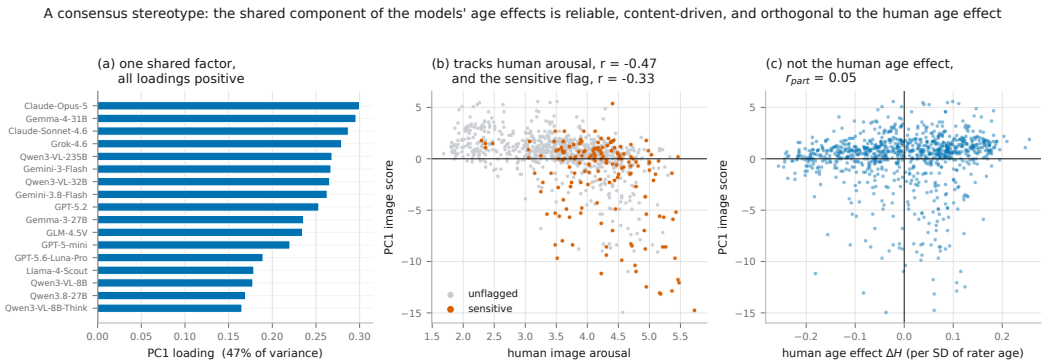


Figure 18: The first principal component of the image×model ΔM matrix: loadings (a), image scores against human arousal (b) and against the human age effect (c).

Table 26: Image-level predictors of the consensus persona shift (older − young, averaged over the 17 models). Ordinary least squares on the 900 images with image-bootstrap intervals (2,000 replicates); continuous predictors are standardised, the sensitive flag and the category indicators (reference: Person) are 0/1. Coefficients are in rating points. The flag carries the effect: it explains more variance alone than the full valence polynomial does, and adding valence to the flag adds little.

Predictor	Valence shift		Arousal shift	
	β	95% CI	β	95% CI
human valence (per SD)	-0.055	[-0.078, -0.036]	-0.105	[-0.123, -0.087]
human valence^2 (z)	+0.164	[0.136, 0.194]	+0.040	[0.019, 0.063]
human arousal (z)	-0.252	[-0.293, -0.212]	-0.115	[-0.141, -0.089]
sensitive flag	-0.185	[-0.271, -0.099]	-0.180	[-0.236, -0.124]
rater disagreement (z)	-0.019	[-0.053, 0.018]	-0.019	[-0.042, 0.005]
Scene	+0.062	[0.005, 0.116]	+0.018	[-0.027, 0.061]
Object	+0.010	[-0.047, 0.060]	-0.060	[-0.108, -0.014]
Animal	+0.255	[0.196, 0.315]	+0.075	[0.030, 0.121]
R^2, sensitive flag only	0.112		0.069	
R^2, human valence only	0.019		0.055	
R^2, valence + flag	0.160		0.191	
R^2, full model	0.432		0.302	

F Agent-Based Mitigation Study

F.1 Design and arms

We test which ingredient, if any, helps an age persona track human age differences. All arms use GPT-5.6-Luna-Pro, rate the same images under the young, middle and older personas, and return the four fields of the main protocol (valence, arousal, description, emotion), so they are scored with the estimators in Appendix B.3. The direct prompt (S0) reuses the main-study responses. Each further arm adds one ingredient (Table 38).

F.1.1 Balanced subset

The subset contains exactly 75 images per OASIS list, allocated within each list in proportion to image category, valence class (unpleasant, neutral, pleasant) and the sensitive flag. Among 200 stratified draws we kept the one whose human distributions best matched the remaining images, a criterion that uses no model output and therefore cannot favour any arm (Table 39). On the subset, the main-study results of the 17 VLMs are reproduced (correlation across models 0.94 for the older-minus-young shift and 0.76 for r_{part}).

Table 27: Per-model version of Table 26: variance in the older-minus-young shift explained by the sensitive flag alone, by a quadratic in human valence alone, and by the full predictor set, with the flag coefficient (rating points). Left block valence, right block arousal.

Model	R^2_{flag}	R^2_{val}	R^2_{full}	β_{flag}	R^2_{flag}	R^2_{val}	R^2_{full}	β_{flag}
Claude-Opus-5	0.12	0.03	0.38	-0.18	0.07	0.06	0.23	-0.26
Grok-4.6	0.07	0.02	0.35	-0.14	0.03	0.06	0.17	-0.29
Claude-Sonnet-4.6	0.08	0.03	0.34	-0.15	0.03	0.06	0.23	-0.23
Qwen3-VL-32B	0.08	0.03	0.31	-0.39	0.02	0.01	0.12	-0.12
Gemma-4-31B	0.09	0.00	0.31	-0.21	0.04	0.00	0.17	-0.17
Qwen3-VL-235B	0.08	0.03	0.30	-0.26	0.00	0.05	0.09	-0.09
Gemini-3-Flash	0.09	0.05	0.29	-0.18	0.01	0.13	0.22	-0.29
GPT-5.2	0.07	0.01	0.25	-0.31	0.01	0.02	0.08	-0.13
Gemini-3.8-Flash	0.04	0.02	0.24	-0.03	0.01	0.02	0.07	-0.08
Gemma-3-27B	0.04	0.02	0.24	-0.25	0.10	0.01	0.21	-0.41
GLM-4.5V	0.02	0.00	0.19	-0.15	0.01	0.01	0.07	-0.22
GPT-5-mini	0.04	0.01	0.18	-0.17	0.02	0.03	0.11	-0.20
Qwen3-VL-8B-Think	0.04	0.01	0.13	-0.31	0.01	0.00	0.02	-0.25
Llama-4-Scout	0.03	0.01	0.11	-0.07	0.01	0.00	0.03	-0.04
GPT-5.6-Luna-Pro	0.01	0.02	0.10	-0.13	0.03	0.01	0.07	-0.21
Qwen3-VL-8B	0.03	0.01	0.07	-0.12	0.03	0.02	0.05	-0.08
Qwen3.8-27B	0.01	0.00	0.07	-0.06	0.00	0.01	0.03	-0.03

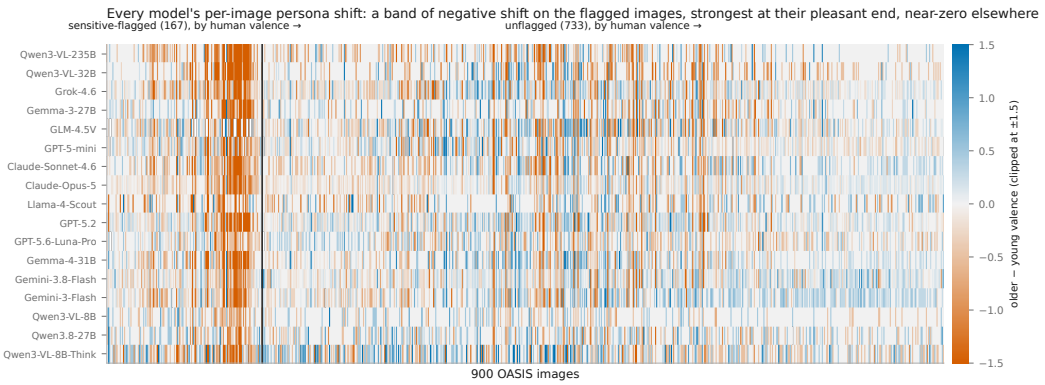


Figure 19: Older-minus-young valence for every image (columns; flagged images first, each block ordered by human valence) and every model (rows, ordered by mean shift).

F.2 Agent components

F.2.1 Perception agent

Describes the image without any persona, returning its content and the elements a viewer is likely to react to emotionally. Keeping this step persona-free separates perception from age conditioning. The output is cached and shared by all agentic arms, so they reason over an identical description.

F.2.2 Persona rater

Receives the image, the perception output and the persona sentence of the main study, and returns a rationale followed by the four rating fields. Placing the rationale first exposes the reasoning that the critic evaluates.

F.2.3 Stereotype critic and reviser (S2, S3)

The critic flags drafts that rely on assumptions about the age group (frailty, fear of risk, nostalgia, reduced enjoyment, disapproval) rather than on the image; flagged drafts are revised once, and unparseable revisions fall back to the draft.

Table 28: Categorical emotion shift between the young and older conditions on the same image. Y/O gives the count assigned to the grouping under young only and under older only; the test is a two-sided exact binomial on those discordant pairs.

Model	n pairs	Changed	Appet. Y/O	q	Avers. Y/O	q
Qwen3-VL-32B	900	202	62/33	0.022^{*}	16/73	0.000^{***}
Grok-4.6	900	193	33/44	0.393	5/66	0.000^{***}
Claude-Sonnet-4.6	898	206	65/44	0.104	12/64	0.000^{***}
Qwen3-VL-235B	899	208	60/35	0.038^{*}	10/63	0.000^{***}
Gemma-3-27B	900	195	71/24	0.000^{***}	21/54	0.000^{***}
GPT-5.2	900	175	55/29	0.026^{*}	10/53	0.000^{***}
Gemma-4-31B	900	197	34/46	0.371	12/51	0.000^{***}
GLM-4.5V	818	163	48/40	0.593	25/47	0.017^{*}
Claude-Opus-5	899	140	51/27	0.030^{*}	3/46	0.000^{***}
GPT-5-mini	888	155	42/37	0.740	11/39	0.000^{***}
Gemini-3-Flash	894	150	42/23	0.060	7/38	0.000^{***}
Gemini-3.8-Flash	877	150	41/34	0.593	6/36	0.000^{***}
Qwen3-VL-8B-Think	896	175	27/57	0.012^{*}	25/29	0.683
GPT-5.6-Luna-Pro	895	122	28/26	0.892	8/25	0.006^{**}
Llama-4-Scout	899	118	36/19	0.064	12/24	0.079
Qwen3.8-27B	898	156	42/32	0.419	31/22	0.289
Qwen3-VL-8B	900	105	20/17	0.789	25/16	0.239

Only images with a valid in-vocabulary label under *both* conditions contribute. q is the Benjamini–Hochberg adjusted p-value across models.

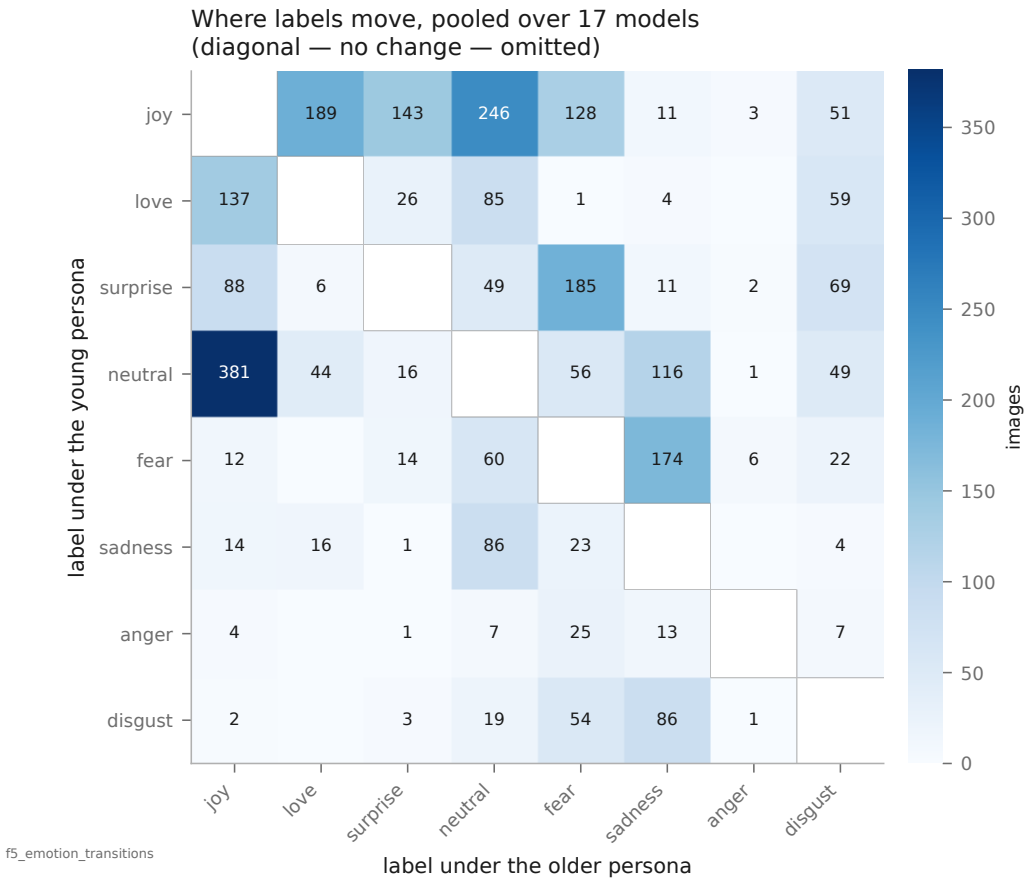


Figure 20: Pooled label transitions from the young to the older persona. The diagonal, which carries most images, is omitted so that the off-diagonal structure is visible.

Table 29: Net images gained by each emotion label under the older persona (assigned under older only minus under young only, on images with a valid label in both conditions). Sadness and fear are the labels that gain, disgust gains modestly, anger does not; surprise, joy and neutral are shed. Stars: exact binomial on the discordant pairs, Benjamini–Hochberg over the 136 model×label tests.

Model	joy	love	surprise	neutral	fear	sadness	anger	disgust
Grok-4.6	-13	+24**	-36***	-36***	+29**	+25***	-2	+9
Qwen3-VL-235B	-14	-11	-40***	+12	+27**	+25***	-3	+4
Claude-Sonnet-4.6	-12	-9	-20	-11	+27**	+14*	-3	+14
Claude-Opus-5	+2	-26***	-13	-6	+20**	+11*	-1	+13*
GPT-5.2	-25*	-1	-5	-12	+20*	+14*	-5	+14
GLM-4.5V	-9	+1	-4	-10	+18*	+3	-1	+2
Gemma-4-31B	+8	+4	-24**	-27*	+15	+25***	+1	-2
Gemini-3-Flash	-11	-8	-23**	+11	+14	+17**	-3	+3
Gemini-3.8-Flash	-5	-2	-19**	-4	+11	+21***	-7	+5
GPT-5.6-Luna-Pro	+9	-11	-6	-9	+10	+10	-3	+0
Qwen3-VL-32B	-14	-15*	+10	-38***	+8	+20*	+0	+29***
GPT-5-mini	-13	+8	-3	-20*	+6	+24***	-2	+0
Gemma-3-27B	-42***	-5	-4	+18	+6	+26**	-2	+3
Llama-4-Scout	-9	-8	-5	+10*	+6	+11*	-4	-1
Qwen3.8-27B	-6	-4	+1	+18*	-8	-2	-2	+3
Qwen3-VL-8B-Think	+20	+10	-4	-30*	-12	+17	-6	+5
Qwen3-VL-8B	+1	-4	-11*	+23*	-13	+10	-1	-5
Pooled	**-133**	**-57**	**-206**	**-111**	**+184**	**+271**	**-44**	**+96**

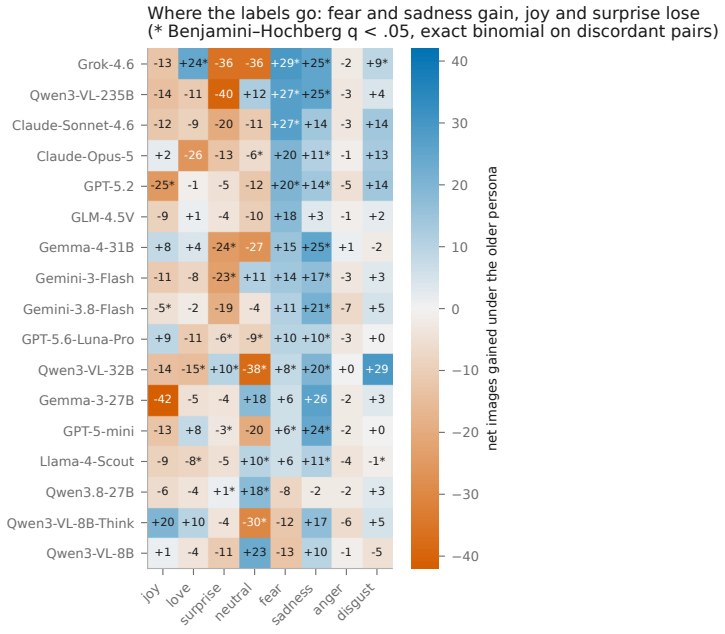


Figure 21: Net images gained by each label under the older persona, per model.

F.2.4 HUMAN-EVIDENCE RETRIEVER (S3–S5, VERA-CAL)

Returns the images most similar to the target by cosine similarity of sentence embeddings (all-mpnet-base-v2) of the model's own no-age descriptions, restricted to the three OASIS lists other than the target's and excluding images of the same theme. For each neighbour it reports pooled human valence and arousal and the estimated older-minus-young difference, obtained by scaling the human age slope ΔH_i from $+1$ SD of rater age to the distance between the band midpoints.

F.2.5 RATER PANEL (S4, VERA-CAL)

For each age band, simulates five deliberately different raters (gender, background, temperament), each rating the image independently with the evidence in view. Panel means and spreads are passed on.

Table 30: Diversity of the emotion-label distribution by condition. H is Shannon entropy in bits over the eight labels; JS is the Jensen–Shannon divergence (bits) between two conditions' label distributions. The older persona barely changes the diversity of labels but redistributes them: the share of *fear* rises and that of *neutral* falls for most models.

Model	$H_\varnothing$	H_Y	H_M	H_O	JS(Y, O)	JS(Y, M)	fear % Y/O	neutral % Y/O
Grok-4.6	2.54	2.51	2.57	2.48	0.014	0.003	13.8/17.0	17.1/13.1
Qwen3-VL-235B	2.55	2.47	2.45	2.40	0.011	0.001	13.6/16.6	17.8/19.1
Qwen3-VL-32B	2.47	2.42	2.45	2.50	0.009	0.001	16.0/16.9	19.2/15.0
Gemma-4-31B	2.56	2.46	2.46	2.42	0.005	0.001	12.2/13.9	17.9/14.9
Gemini-3.8-Flash	2.59	2.56	2.52	2.51	0.005	0.002	14.2/15.6	18.4/18.0
Claude-Sonnet-4.6	2.63	2.63	2.64	2.62	0.005	0.000	16.1/19.2	18.0/16.8
Gemini-3-Flash	2.59	2.55	2.52	2.50	0.004	0.003	13.8/15.3	16.9/18.1
Qwen3-VL-8B-Think	2.34	2.28	2.30	2.25	0.004	0.002	14.4/13.0	19.3/15.9
Claude-Opus-5	2.66	2.64	2.64	2.63	0.004	0.000	14.8/17.0	15.8/15.1
Qwen3-VL-8B	2.23	2.15	2.14	2.05	0.004	0.003	11.4/10.0	23.7/26.2
Gemma-3-27B	2.54	2.51	2.55	2.50	0.003	0.002	20.0/20.7	22.0/24.0
GPT-5.2	2.61	2.63	2.64	2.64	0.003	0.001	16.9/19.1	16.3/15.0
GPT-5-mini	2.52	2.48	2.47	2.47	0.002	0.001	14.6/15.4	9.7/7.6
Llama-4-Scout	2.45	2.42	2.43	2.38	0.001	0.001	11.0/11.7	29.2/30.4
GLM-4.5V	2.44	2.46	2.44	2.47	0.001	0.001	11.9/14.0	29.7/28.4
GPT-5.6-Luna-Pro	2.58	2.59	2.59	2.56	0.001	0.000	15.3/16.4	20.7/19.7
Qwen3.8-27B	2.50	2.56	2.51	2.54	0.001	0.001	15.2/14.3	29.0/31.0

Table 31: What the older persona does to the free-text description. Each cell is young/older. *words* is the mean description length; the remaining columns are the percentage of descriptions matching a lexicon: *memory* (memories, reminds, nostalgic, reminiscent, the past, youth, decades), *age ref.* (an explicit reference to the rater's age or generation), *1st person*, *excitement* (exciting, thrilling, adventure, fun), and *vulnerability* (fragile, frail, mortality, loss, health, danger, worry). The lexica were fixed before the counts were run; the full regular expressions are in the analysis script.

Model	words Y/O	memory %	age ref. %	1st person %	excitement %	vulnerab. %
Gemma-3-27B	18.5/21.1***	7.2/88.0***	0.0/6.9***	65.8/93.9***	18.0/5.9***	19.8/23.3*
Qwen3-VL-235B	18.4/23.2***	7.3/72.0***	6.9/37.3***	7.0/50.4***	21.8/7.8***	17.7/27.4***
Qwen3-VL-32B	22.3/25.9***	3.7/63.8***	10.9/30.7***	0.0/14.1***	16.2/6.6***	17.2/31.9***
Claude-Sonnet-4.6	20.8/25.3***	5.9/53.9***	2.6/10.6***	1.2/18.3***	22.9/6.8***	14.7/28.3***
GPT-5-mini	11.8/14.5***	4.1/44.8***	0.3/11.0***	1.4/20.0***	20.8/8.6***	11.5/27.6***
Grok-4.6	10.1/11.4***	4.2/33.3***	0.7/1.9*	0.0/0.0	16.4/5.1***	5.1/20.6***
GLM-4.5V	16.0/19.3***	2.7/33.0***	0.2/2.9***	0.6/25.9***	13.7/4.0***	7.2/23.2***
Gemma-4-31B	12.1/13.7***	2.9/22.7***	0.4/2.2**	0.1/4.7***	12.2/3.8***	5.8/17.3***
Claude-Opus-5	17.5/21.0***	6.2/21.1***	0.9/1.3	0.4/4.2***	22.9/11.1***	12.6/20.4***
Qwen3-VL-8B-Think	7.1/7.6***	1.2/21.0***	0.7/0.9	0.0/0.1	10.5/5.0***	7.5/11.3**
Gemini-3-Flash	15.3/16.3***	5.1/20.8***	1.7/1.5	0.0/2.1***	14.7/5.3***	11.5/22.7***
GPT-5.2	13.2/15.6***	1.6/17.4***	0.3/2.2**	10.6/27.2***	14.3/7.0***	15.0/30.4***
Qwen3.8-27B	17.0/17.8***	2.0/15.3***	0.2/1.1*	0.1/0.0	11.7/5.5***	8.5/17.0***
Gemini-3.8-Flash	10.0/10.1	3.9/12.4***	0.3/1.1	0.0/0.0	12.2/3.1***	6.8/19.3***
Llama-4-Scout	5.8/6.5***	1.0/9.7***	0.2/0.3	1.4/4.6***	12.2/8.6***	5.6/9.1***
Qwen3-VL-8B	11.3/12.8***	2.6/7.4***	0.8/4.2***	0.7/0.2	12.6/7.7***	10.1/14.0***
GPT-5.6-Luna-Pro	11.6/12.2***	0.8/4.7***	0.3/0.1	0.1/0.0	13.2/7.5***	10.2/20.7***
Pooled	14.0/16.1	3.7/31.8	1.6/6.9	5.3/15.6	15.7/6.4	11.0/21.4

Stars: Wilcoxon (length) or exact binomial on discordant image pairs (flags), Benjamini–Hochberg across models within column; $^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$.

F.2.6 ANCHORED CALIBRATOR (S4, VERA-CAL)

Starting from the model's own no-age rating from the main study, outputs the three persona ratings, sizing the age differences to the evidence rather than copying the panels, which tend to exaggerate them.

F.2.7 AUDITOR (S4, VERA-CAL)

Flags caricature (age differences outside the range of the evidence) and stereotyped descriptions; flagged outputs are recalibrated once.

Table 32: Words most over-represented in older-persona descriptions (left) and in young-persona descriptions (right), pooled over models: weighted log-odds ratio with an informative Dirichlet prior (Monroe et al., 2008), prior counts taken from the pooled corpus of all four conditions; n_Y, n_O are the numbers of descriptions containing the word under each persona. The older persona is written in the register of reminiscence, the young one in the register of excitement.

older word	n_Y	n_O	z	young word	n_Y	n_O	z
nostalgia	154	1280	14.4	kinda	347	0	-11.0
reminds	35	880	12.9	cute	439	43	-9.2
memories	25	702	12.6	vibe	250	1	-9.1
reminiscent	44	763	12.5	excitement	1148	466	-8.7
reminding	19	653	12.3	really	313	38	-8.5
older	47	661	12.0	intense	893	317	-8.4
quiet	1128	2833	11.2	like	1086	545	-8.2
brings	8	516	10.9	adrenaline	261	14	-7.8
nostalgic	267	1040	10.1	romantic	414	73	-7.4
life	77	630	9.9	excited	245	33	-6.9
past	23	379	9.3	fun	228	33	-6.5
concern	343	1172	8.9	friends	289	76	-6.3
gentle	327	1098	8.7	relaxed	413	136	-6.0
safety	22	370	8.5	creating	178	9	-5.8
years	0	243	8.2	chill	99	2	-5.7
later	0	225	8.1	attraction	190	16	-5.6
grandchildren	0	194	7.6	super	86	0	-5.5
fond	6	218	7.5	curiosity	485	219	-5.3
times	21	281	7.4	rush	116	9	-5.2
worry	51	358	7.3	cozy	246	73	-5.1

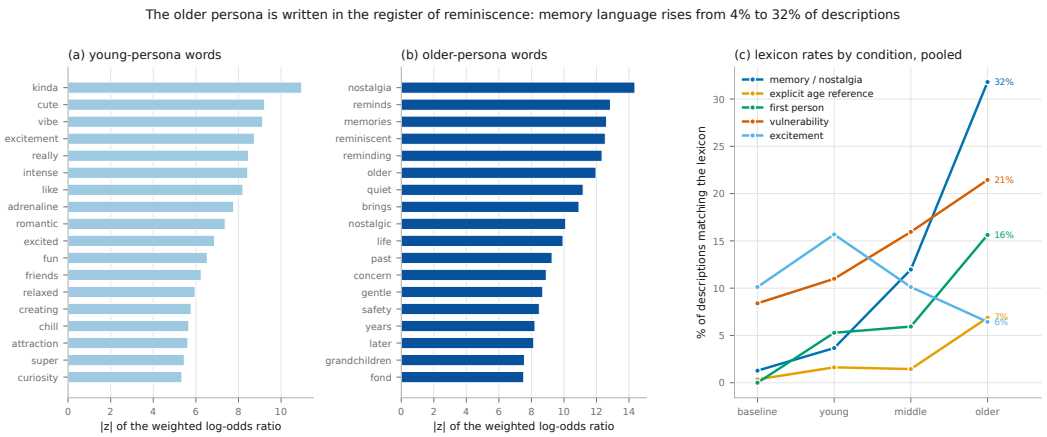


Figure 22: Words most diagnostic of the young (a) and older (b) personas by weighted log-odds, and lexicon rates by condition pooled over models (c).

F.2.8 Hypothesiser and code verifier (S5, VERA-Cal)

From three retrieved cases, the hypothesiser states one general, testable rule about how age changes responses to this kind of content and predicts the target's older-minus-young difference. In a text-only call the rule is applied to four further cases whose human answers are withheld from the prompt, and the predictions are scored *in code* against the human differences (same sign, or both below 0.05 and 0.1 in magnitude). A rule agreeing on fewer than three of four cases is revised once from the per-case errors and re-tested on four *fresh* cases the agent has never seen; if it fails again, the agent abstains and predicts no age difference. No case is ever used to grade a rule that was fitted to it.

F.3 Execution and validation

F.3.1 Leakage controls

Evidence never includes the target image, any image from its OASIS list, or an image of the same theme; we verified this for every image (0 violations). The verifier's test prompts contain no human

Table 33: Does the description predict the numeric shift on the same image (valence)? Per model, the older-minus-young rating is regressed on the image covariates of Table 26 (valence, valence^2, sensitive flag) plus the older-minus-young change in each text feature. Coefficients are rating points per unit change of the flag (or per SD for length). R^2_{img} uses image covariates only; $R^2_{+\text{text}}$ adds the text features.

Model	memory	age ref.	vulnerab.	excitement	Δlength	R^2_{img}	$R^2_{+\text{text}}$
Claude-Opus-5	+0.14***	-0.18	-0.06	+0.22***	-0.01	0.16	0.21
Claude-Sonnet-4.6	+0.14***	-0.13	-0.20***	+0.40***	-0.05*	0.11	0.23
Gemini-3-Flash	+0.10*	-0.12	-0.26***	+0.55***	+0.02	0.12	0.24
Gemini-3.8-Flash	+0.07	+0.08	-0.14**	+0.43***	+0.03	0.07	0.16
Gemma-3-27B	+0.41***	-0.21	-0.26***	+0.39***	-0.15***	0.09	0.25
Gemma-4-31B	+0.14**	+0.10	-0.12	+0.54***	-0.05*	0.10	0.19
GLM-4.5V	+0.36***	+0.11	-0.45***	+1.00***	-0.04	0.03	0.27
GPT-5-mini	+0.24***	-0.14	-0.20**	+0.66***	-0.01	0.07	0.20
GPT-5.2	+0.33***	-0.31	-0.15*	+0.68***	-0.03	0.10	0.23
GPT-5.6-Luna-Pro	+0.15*	+0.19	-0.17**	+0.40***	+0.01	0.05	0.12
Grok-4.6	+0.32***	+0.09	-0.23***	+0.76***	-0.07**	0.10	0.30
Llama-4-Scout	+0.08	-0.33	+0.04	+0.30**	+0.01	0.04	0.05
Qwen3-VL-235B	+0.13**	-0.05	-0.22**	+0.30***	-0.03	0.13	0.18
Qwen3-VL-32B	+0.29***	-0.00	-0.33***	+0.39***	-0.09**	0.15	0.31
Qwen3-VL-8B	-0.01	-0.05	+0.07*	+0.07*	-0.01	0.05	0.06
Qwen3-VL-8B-Think	+0.23***	-0.37	-0.02	+0.44***	-0.03	0.07	0.11
Qwen3.8-27B	+0.07	+0.06	-0.12*	+0.34***	+0.00	0.02	0.08

HC1 standard errors; stars Benjamini–Hochberg across models within column.

Table 34: Does the description predict the numeric shift on the same image (arousal)? Per model, the older-minus-young rating is regressed on the image covariates of Table 26 (valence, valence^2, sensitive flag) plus the older-minus-young change in each text feature. Coefficients are rating points per unit change of the flag (or per SD for length). R^2_{img} uses image covariates only; $R^2_{+\text{text}}$ adds the text features.

Model	memory	age ref.	vulnerab.	excitement	Δlength	R^2_{img}	$R^2_{+\text{text}}$
Claude-Opus-5	+0.03	-0.12	+0.01	+0.19***	-0.01	0.20	0.25
Claude-Sonnet-4.6	+0.01	-0.08	-0.05	+0.21***	+0.02	0.14	0.18
Gemini-3-Flash	-0.06	-0.13	+0.22**	+0.26**	-0.02	0.19	0.23
Gemini-3.8-Flash	-0.03	+0.09	+0.04	+0.20***	-0.04	0.06	0.09
Gemma-3-27B	+0.06	-0.23	-0.17**	+0.24***	-0.06	0.13	0.19
Gemma-4-31B	+0.00	-0.07	-0.03	+0.49***	+0.01	0.06	0.12
GLM-4.5V	+0.03	-0.09	+0.16	+0.51***	+0.01	0.07	0.11
GPT-5-mini	-0.06	-0.04	+0.02	+0.33***	-0.02	0.07	0.11
GPT-5.2	-0.00	-0.00	-0.00	+0.18***	+0.01	0.05	0.07
GPT-5.6-Luna-Pro	+0.08	+0.38	+0.06	+0.26***	-0.01	0.05	0.08
Grok-4.6	-0.02	+0.12	+0.10	+0.30***	-0.00	0.13	0.17
Llama-4-Scout	-0.15	-0.20	+0.04	+0.16	+0.04	0.01	0.02
Qwen3-VL-235B	-0.06	+0.00	+0.04	+0.22***	+0.01	0.06	0.09
Qwen3-VL-32B	+0.03	+0.01	-0.06	+0.31***	-0.01	0.05	0.09
Qwen3-VL-8B	-0.10	-0.02	+0.02	+0.17**	+0.01	0.04	0.05
Qwen3-VL-8B-Think	-0.09	-0.56	+0.15	+0.33**	-0.03	0.01	0.02
Qwen3.8-27B	-0.07	-0.02	-0.01	+0.39***	-0.01	0.02	0.06

HC1 standard errors; stars Benjamini–Hochberg across models within column.

answers, refinement uses test cases only, and re-verification uses unseen cases. The anchor is the model's own no-age rating, not a human value.

F.3.2 Execution, caching, and cost

All calls use `openai/gpt-5.6-luna-pro` through OpenRouter with temperature 0, one completion and a 1,500-token budget, matching the main protocol. Every call is cached by model, prompt text and image, and every intermediate output (perception, rationale, critique, neighbours, panels, rules and verification scores) is stored with the final rating. The six arms produced 99.7–100% valid ratings on the 300 images, for a total cost of $24.21.

Table 35: Across the 17 models, Spearman correlations between properties of a model and its response to the age persona ($n = 17$ models; exploratory, Benjamini–Hochberg within column). The size of the persona shift is unrelated to baseline competence and to over-dispersion, but tracks how much the model writes about memory and about its own age under the older persona.

Model-level predictor	Δ_{O-Y} valence	Δ_{O-Y} arousal	aversive net	r_{part} valence
over-dispersion (SD ratio, valence)	-0.09	-0.51	-0.08	-0.06
baseline r (valence)	-0.07	0.09	0.41	0.48
baseline MAE (valence)	-0.26	-0.40	-0.04	-0.26
description length (young)	-0.50	0.07	0.50	0.35
Δ description length (O$-$Y)	-0.73**	-0.12	0.61*	0.30
memory lexicon rate (older)	-0.76**	-0.44	0.67*	0.33
explicit age reference (older)	-0.65*	-0.39	0.53	0.20
first-person rate (older)	-0.60*	-0.27	0.37	0.16
run cost (USD, price proxy)	0.11	0.17	0.19	0.22

$^{*}q < .05$, $^{**}q < .01$, $^{***}q < .001$. Two-sided.

Table 36: Controlled comparisons within a developer. Each group holds the developer fixed and varies one thing: parameter count (the Qwen3-VL ladder), inference-time reasoning at a fixed 8B (Thinking), model generation, or product tier. Columns: baseline r and SD ratio (valence); the older-minus-young shift on valence and arousal; the partial correlation with the human age effect; the share of images the persona does not move; labels changed and net aversive labels gained; and the share of older-persona descriptions containing memory language or an explicit age reference. The step that most increases the persona shift (Qwen3-VL 8B to 32B) is also the step that most increases memory and age-reference language; newer generations and higher tiers (Gemma-4, GPT-5.2 and 5.6, Claude-Opus-5) write less in that register and, in most cases, shift less. No step moves the human alignment out of the noise.

Group	Model	r_{base}	SD ratio	Δ^{V}_{O-Y}	Δ^{A}_{O-Y}	r_{part}	flat	changed	avers.	memory %	age ref. %
Qwen3-VL ladder	Qwen3-VL-8B	0.885	1.34	-0.020	-0.296	0.001	73%	105	-9	7.4	4.2
	Qwen3-VL-32B	0.866	1.43	-0.219	-0.295	0.002	0%	202	+57	63.8	30.7
	Qwen3-VL-235B	0.902	1.31	-0.226	-0.245	0.072	0%	208	+53	72.0	37.3
Reasoning at 8B	Qwen3-VL-8B	0.885	1.34	-0.020	-0.296	0.001	73%	105	-9	7.4	4.2
	Qwen3-VL-8B-Think	0.837	1.66	+0.010	-0.441	0.044	14%	175	+4	21.0	0.9
Qwen generation	Qwen3-VL-32B	0.866	1.43	-0.219	-0.295	0.002	0%	202	+57	63.8	30.7
	Qwen3.8-27B	0.855	1.22	-0.006	-0.157	0.024	30%	156	-9	15.3	1.1
Gemma generation	Gemma-3-27B	0.863	1.40	-0.157	-0.402	0.009	1%	195	+33	88.0	6.9
	Gemma-4-31B	0.908	1.35	-0.057	-0.429	0.042	16%	197	+39	22.7	2.2
Gemini generation	Gemini-3-Flash	0.912	1.36	-0.026	-0.395	0.065	21%	150	+31	20.8	1.5
	Gemini-3.8-Flash	0.919	1.20	-0.045	-0.275	0.059	9%	150	+30	12.4	1.1
OpenAI tier	GPT-5-mini	0.877	1.50	-0.141	-0.518	0.037	4%	155	+28	44.8	11.0
	GPT-5.2	0.891	1.38	-0.091	-0.231	-0.030	7%	175	+43	17.4	2.2
	GPT-5.6-Luna-Pro	0.890	1.51	-0.076	-0.193	0.006	0%	122	+17	4.7	0.1
Anthropic tier	Claude-Sonnet-4.6	0.900	1.36	-0.126	-0.272	0.086	0%	206	+52	53.9	10.6
	Claude-Opus-5	0.923	1.28	-0.116	-0.208	0.058	11%	140	+43	21.1	1.3

F.3.3 Pipeline diagrams

F.4 Evaluation and results

F.4.1 Attainable ceiling

The human age effects ΔH_i are posterior means, so their reliability follows from the shrinkage identity, $\rho = \mathrm{Var}(\Delta H)/(\mathrm{Var}(\Delta H) + \overline{\tau^2})$, where τ_i is the posterior SD. For r_{part} the relevant quantity is the age effect after the quadratic in pooled valence is partialled out; with residual e_i, $\rho_{\text{part}} = \mathrm{Var}(e)/(\mathrm{Var}(e) + \overline{\tau^2})$ and the attainable correlation is $\sqrt{\rho_{\text{part}}} = 0.60$ (0.86 for the raw age effect). The attainable direction agreement is $\frac{1}{n}\sum_i \Phi(|\Delta H_i|/\tau_i) = 85\%$, the rate at which the true effect matches the sign of its estimate.

F.4.2 Results and paired statistical tests

We report every arm against the direct prompt below, with paired tests on the same images: a paired bootstrap for r_{part}, an exact McNemar test on per-image direction correctness and a Wilcoxon signed-

Table 37: **Qwen3-VL scale ladder, the only comparison in which developer, generation and instruction-tuning recipe are held fixed.** Cross-developer differences elsewhere confound parameter count with data and alignment procedure and are not read as scale effects.

Model	Scale (Model Parameters)	Dimension	r_{part}	Flat	Δ_{O-Y}	Labels changed
Qwen3-VL-8B	8B	arousal	-0.108	0%	-0.296	105
Qwen3-VL-32B	32B	arousal	-0.190	0%	-0.295	202
Qwen3-VL-235B	235B	arousal	-0.067	2%	-0.245	208
Qwen3-VL-8B	8B	valence	0.001	73%	-0.020	105
Qwen3-VL-32B	32B	valence	0.002	0%	-0.219	202
Qwen3-VL-235B	235B	valence	0.072	0%	-0.226	208

Table 38: Mitigation arms and the ingredient each isolates.

Arm	Type	Instruction	Critic loop	Human evidence	Panels + anchored calibration	Verified rule
S0 direct prompt	baseline					
S1 debias instruction	reasoning	✓				
S2 deliberation agent	reasoning		✓			
S3 grounded agent	grounded		✓	✓		
S4 contrastive panel	grounded		audit	✓	✓	
S5 VERA	grounded			✓	anchored only	✓
VERA-Cal	grounded		audit	✓	✓	✓

rank test on per-image absolute error, Holm-corrected across the six arms (Figure 26). Grounded arms fix the direction of many more images than they break (e.g. 211 vs. 15 for S4) and halve the error against the human change (0.48 to 0.20; $p < .001$), whereas S1 and S2 change neither significantly. The advantage of S4 in r_{part} (+0.19) is significant before but not after correction (Holm $p = .14$).

F.4.3 Comparison with the 17 VLMs

Figure 27 places each arm among the 17 VLMs on the same images. Reasoning arms stay within the range of existing models. On sensitive-flagged pleasant images, where older OASIS raters are more positive than young raters ($+0.24$), all 17 VLMs move in the opposite direction (-0.29 to -1.65); the debiasing instruction attenuates this (-0.37), deliberation neutralises it (-0.01), and only the grounded arms move with humans ($+0.12$ to $+0.17$).

F.4.4 Sources of the gains

Direction agreement alone cannot separate genuine age knowledge from pleasantness: a rule that predicts the human age effect from pooled valence alone, fitted on the other lists, agrees with humans on 88% of images. Grounded arms reach the same level, which is why we rely on r_{part}. Their gains in agreement appear in all three valence classes, not only on pleasant images (Figure 28b). The panel's simulated raters are less varied than real raters (panel SD 0.49 of the human SD), and the grounded arms' age differences are smaller than the human estimates on average (caricature ratio 0.75–0.97), so the gain is not driven by exaggeration.

F.4.5 Verified rules

Across the 300 images, VERA-Cal's first rule passed the hidden-case test for 222 images (74%); 78 rules were revised, of which 55 passed on fresh cases, and the agent abstained on 23 images (8%). We list verbatim rule examples below. Explicit rules make the agent's assumptions inspectable, but compressing the evidence into one rule costs fidelity: VERA and VERA-Cal match the panel in direction agreement (87%) yet reach about half its r_{part} (0.14–0.15 vs. 0.27).

F.5 Limitations

We use one model, one persona wording and 300 images. At this size the r_{part} intervals are about ± 0.11, so the advantage of the best arm over the direct prompt and over the best VLM is suggestive but not significant after correction. The human evidence is itself noisy (only 41 raters are aged 60 or older), which bounds any grounded method, and the agents receive aggregate human data rather than individual raters. A 60-image pilot run before the fixes described in Appendix F.2 is not reported.

Table 39: The balanced subset against the full stimulus set.

	Subset (300)	All (900)	KS p
Person / Scene / Object / Animal	38 / 25 / 22 / 15%	38 / 24 / 22 / 15%	–
Pleasant / neutral / unpleasant	51 / 24 / 25%	51 / 24 / 25%	–
Sensitive-flagged	17%	19%	–
Human valence, mean (SD)	4.33 (1.21)	4.33 (1.23)	0.95
Human arousal, mean (SD)	3.64 (0.84)	3.67 (0.84)	0.93
Human age effect (valence), SD	0.107	0.106	0.97
Human age effect (arousal), SD	0.057	0.056	0.97

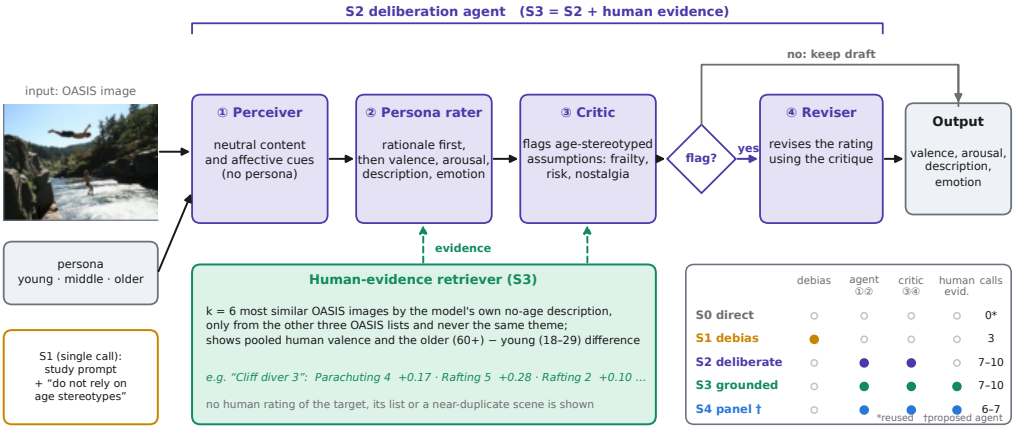


Figure 23: Reasoning and grounded agents (S1–S3) and the components of every arm.

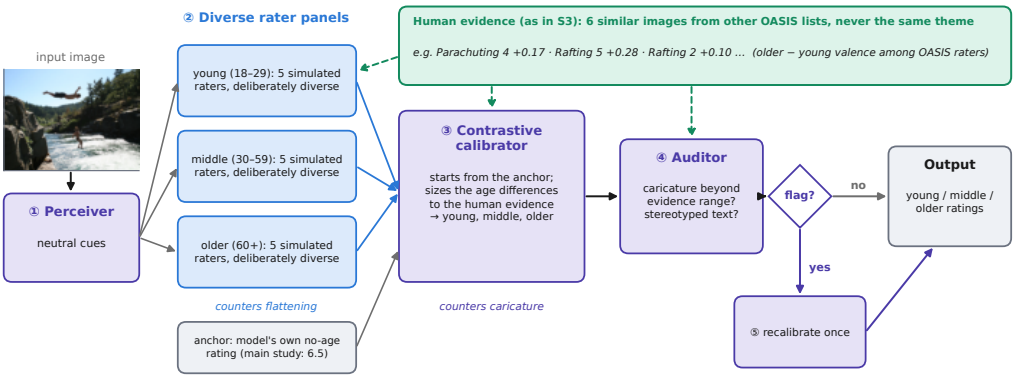


Figure 24: The grounded contrastive panel (S4): diverse rater panels per age band, a calibrator anchored on the model's own no-age rating, and an auditor for caricature and stereotyped text, all with human evidence from the other OASIS lists.

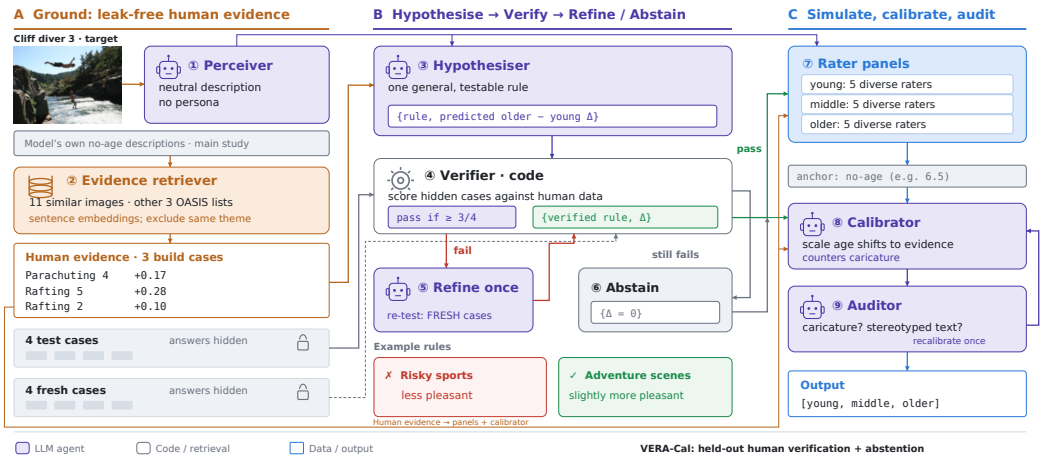


Figure 25: **VERA-Cal.** (A) Human evidence is retrieved from similar images in the other three OASIS lists, never of the same theme, using sentence embeddings of the model's own no-age descriptions, and split into build, test and fresh cases. (B) A hypothesiser states one testable rule about age differences from the build cases; a verifier implemented in code scores it on test cases whose human answers are hidden (pass: $\geq 3/4$); a failing rule is refined once and re-tested on fresh cases, otherwise the agent abstains ($\Delta = 0$). (C) Diverse rater panels, a calibrator anchored on the model's own no-age rating and an auditor produce the three persona ratings; the human evidence and the verified rule are passed to both. Example rules and values are illustrative.

Table 40: Alignment with human age differences on the same 300 images for all systems. r_{part} measures image level fidelity after controlling for pleasantness. Values for system groups are means with 95% joint image bootstrap confidence intervals. The attainable ceiling is estimated from the reliability of the human age effects. Direction agreement can approach or exceed this ceiling because it is strongly driven by pleasantness and should not be read as human level image specific fidelity.

Level	r_{part} [95% CI]	% of ceiling	Direction agreement
17 VLMs, direct prompt	0.06 [−0.04, 0.15]	9%	51%
Reasoning agents (S1, S2)	−0.01 [−0.09, 0.07]	−1%	42%
Human grounded agents (S3 to S5, VERA Cal)	0.16 [0.07, 0.25]	27%	83%
Best: contrastive panel (S4)†	0.27 [0.16, 0.37]	44%	88%
Attainable ceiling	0.60	100%	85%
Bootstrap tests of group mean differences			
Grounded − VLMs	+0.11 [−0.03, 0.25], $p = .13$		+33 pts, $p < .001$
Grounded − reasoning	+0.17 [0.04, 0.30], $p = .011$		+41 pts, $p < .001$
Reasoning − VLMs	−0.06 [−0.16, 0.03], $p = .20$		−8 pts, $p = .006$
Ceiling − every level	all $p < .001$		

†S4 was selected post hoc. Compared with the best VLM, Qwen3 VL 235B, $\Delta r_{\text{part}} = +0.13$ [−0.02, 0.29], $p = .09$.

Table 41: Human alignment of the mitigation arms (GPT-5.6-Luna-Pro, 300 images). r_{part}: fidelity to the human age effect with pleasantness partialled out (95% image bootstrap CI). *Agree*: direction agreement on images the model moves (binomial test vs. 50%). Δr_{part}: paired bootstrap difference from the direct prompt. McNemar: images whose direction becomes correct / incorrect relative to the direct prompt. MAE: mean absolute error of the older-minus-young change. p-values Holm-corrected across the six arms; $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

System	r_{part} [95% CI]	Pearson	Agree	Δr_{part} [95% CI]	McNemar	MAE	Wilcoxon p
S0 direct	0.08 [-0.06, 0.22]	-0.11	39%**	–	–	0.48	–
S1 debias instruction	0.00 [-0.11, 0.11]	-0.03	44%	-0.08 [-0.21, +0.05]	45/27	0.47	.604
S2 deliberation agent	-0.02 [-0.13, 0.10]	-0.22	40%*	-0.10 [-0.27, +0.07]	51/36	0.54	.248
S3 grounded agent	0.09 [-0.04, 0.22]	+0.30	71%***	+0.01 [-0.20, +0.21]	146/27***	0.37	$< .001$
S4 grounded contrastive panel	0.27 [0.16, 0.37]	+0.74	88%***	+0.19 [+0.02, +0.35]	211/15***	0.20	$< .001$
S5 VERA	0.14 [0.03, 0.26]	+0.75	87%***	+0.06 [-0.12, +0.24]	193/14***	0.20	$< .001$
VERA-Cal	0.15 [0.04, 0.27]	+0.75	87%***	+0.07 [-0.11, +0.25]	200/14***	0.20	$< .001$
17 VLMs (range)	-0.07 to 0.14	-0.13 to +0.19	34–64%	–	–	0.38–0.62	–

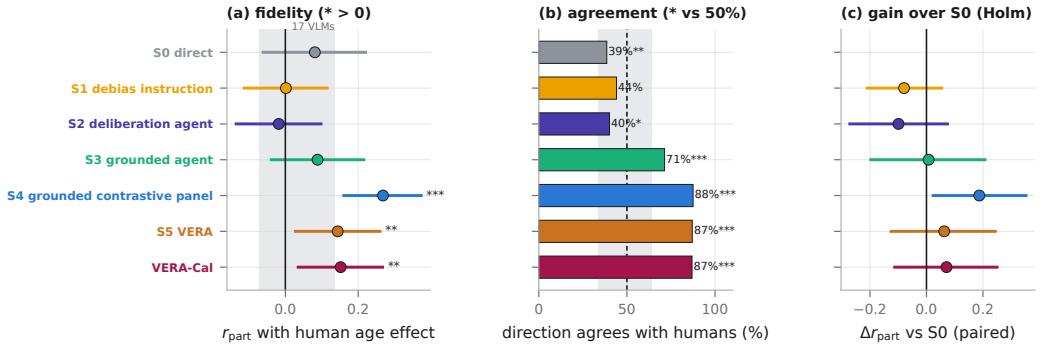


Figure 26: Human alignment of each arm (300 images). (a) r_{part} with 95% CI; stars mark $r_{\text{part}} > 0$; the shaded band is the range of the 17 VLMs. (b) Direction agreement; stars mark a binomial difference from 50% (S0 and S2 are significantly *below* chance). (c) Paired gain in r_{part} over S0; none survives Holm correction.

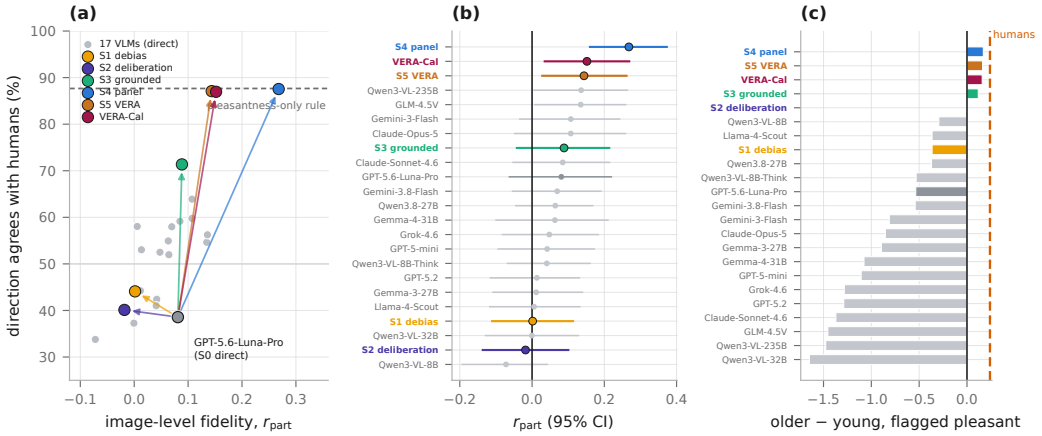


Figure 27: Mitigation arms against the 17 VLMs on the same 300 images. (a) Direction agreement against r_{part}; arrows start from GPT-5.6-Luna-Pro's direct prompt; the dashed line is a pleasantness-only rule fitted on the other lists. (b) r_{part} for all systems. (c) Shift on sensitive-flagged pleasant images.

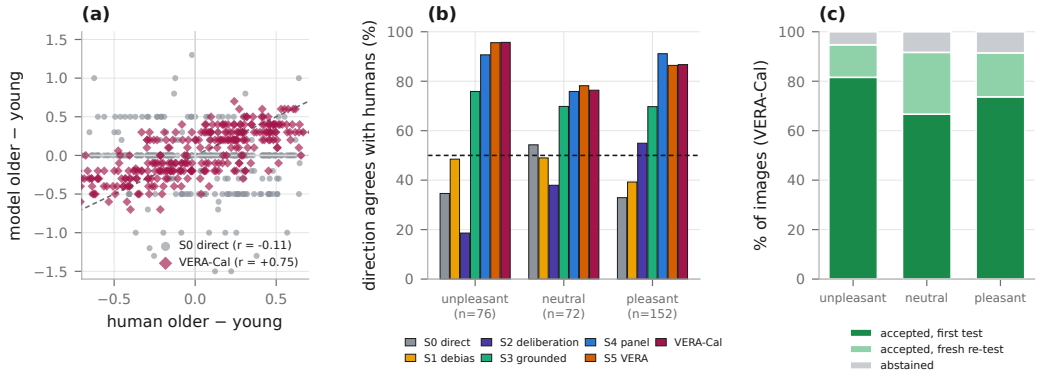


Figure 28: (a) Per image, the model's older-minus-young change against the human estimate for the direct prompt and VERA-Cal. (b) Direction agreement by valence class for every arm. (c) Outcome of VERA-Cal's rules: accepted on the first test, accepted after a fresh re-test, or abstained.

Table 42: Examples of rules stated by the agent, verbatim, with their agreement on hidden human cases (first test; fresh re-test after refinement) and the outcome.

Image	First rule	Pass	Refined rule (fresh pass)	Outcome
Alcohol 7	Compared with younger adults, older adults tend to rate alcohol-related scenes more pleasantly when they convey warm social conviviality, but less pleasantly when they are primarily impersonal product displays.	75%	–	accepted
Camping 2	For calm, positive scenes of children engaged together, adults aged 60 and older tend to rate pleasantness modestly higher than adults aged 18 to 29.	75%	–	accepted
Camping 4	For warm, affectionate, and playful interpersonal scenes, older adults tend to rate pleasantness modestly higher than younger adults, by roughly three-tenths of a scale point.	100%	–	accepted
Ambulance 1	For subdued or peaceful scenes, older adults' pleasantness ratings differ only modestly from younger adults' ratings and show no consistent directional shift across content.	50%	Older adults' pleasantness ratings shift upward relative to younger adults for benign or comforting scenes but downward for scenes suggesting vulnerability, threat, or wartime distress, with larger differences when those cues are salient. (75%)	refined
Cat 13	For calm, gentle, and affectionate animal images, older adults tend to give higher pleasantness ratings than younger adults.	50%	Older adults tend to rate emotionally positive and affiliative images, especially animals and babies, as more pleasant than younger adults, while neutral images show little or no age difference. (100%)	refined
Cold 7	The older-minus-younger pleasantness difference is context-dependent, tending to be positive for clearly peaceful, contentment-evoking landscapes but negative for stark snowy scenes emphasizing solitude or melancholy.	50%	Older adults generally rate scenes as more pleasant than younger adults, including quiet or solitary snowy scenes, although the size of this positive difference varies substantially by image. (100%)	refined
Cat 6	For calm, neutral-to-gently endearing faces or animals, older adults tend to give slightly higher pleasantness ratings than young adults.	25%	Older adults tend to rate clearly appealing and positive animals as more pleasant than young adults, but show little or lower pleasantness for neutral, ambiguous, sad, bored, or mildly intimidating stimuli. (25%)	abstained
Cat 9	Compared with younger adults, older adults tend to rate calm, solitary animals with gentle, attentive expressions as more pleasant, while this age advantage is smaller for interaction-focused scenes.	50%	Older adults tend to rate clearly warm, affiliative companion-animal scenes as more pleasant than younger adults, but show little or no advantage for merely attentive, ambiguous, or neutral expressions. (50%)	abstained
Cliff diver 1	For images depicting physically risky or exhilarating acts, older adults tend to report slightly lower pleasantness than younger adults on average, with the difference varying by scene.	25%	For physically risky or exhilarating scenes, older adults tend to report slightly higher pleasantness than younger adults, especially for aerial or open-water acts, while responses to climbing scenes may be near zero or slightly lower. (50%)	abstained

G QUALITATIVE EXAMPLES

We use these examples to show, image by image, the patterns summarised in Appendix C. The examples illustrate the main-study findings; they are not additional evidence. We base the main-study quantitative claims on the full 900-image analyses. The separate mitigation study uses the 300-image subset described in Appendix F.1.1.

G.1 IMAGE AND MODEL SELECTION

We draw 15 OASIS images that are not sensitive-flagged and that carry a supported human valence age effect (Appendix B.1). To cover the stimulus set, we stratify by OASIS category (Person, Scene, Animal, Object) and by pooled human valence, take in each stratum the image with the largest 17-model older-minus-young shift, and allow one image per theme. Each page shows four models. GPT-5.6-Luna-Pro, the strongest GPT model we evaluate, appears on every page as a fixed reference. The other three come from three other developers, chosen as those with the largest absolute shift on that image, and none of these additional models appears on more than four pages. All values and quotes are copied verbatim from the released model responses.

G.2 PANEL DEFINITIONS AND COLOUR CODING

The *ground truth* panel reports pooled OASIS valence and arousal (1–7), the number of raters, and the image-specific human age effect ΔH_i (per SD of rater age), estimated from all 822 participants. The age effect is labelled green when older raters rate the image more pleasant and red when they rate it less pleasant. Each *model* panel reports:

- baseline (no-age) valence and arousal, green when within one scale point of the human mean;
- valence and emotion label under the young, middle and older personas;
- the older-minus-young shift, green when it points in the direction of ΔH_i, red when it points the opposite way, and grey when it is smaller than 0.1;
- the older persona's description, with vulnerability, memory and excitement words coloured by the lexica in Table 31.

The comparison is directional only: ΔH_i is a per-SD slope and the model shift is a difference between two personas, so their magnitudes are not on a common scale.

G.3 PATTERNS ILLUSTRATED BY THE EXAMPLES

Four patterns recur across developers. **(i) Direction is close to chance.** The 17-model mean shift agrees in sign with the human age effect on 7 of the 15 images, points the opposite way on 7, and does not move on one (Picnic 2). This is consistent with the near-zero image-level alignment in Table 21. Agreement, where it occurs, also appears on images where human and model shifts share a sign for unrelated reasons (Fire 6, Boxing 2).

(ii) Leisure and risk scenes are scripted as danger. On Skydiving 3, Rollercoaster 3 and Cliff diver 3, older OASIS raters rate the image slightly *more* pleasant, yet models switch from *joy* to *fear*. On Cliff diver 3, GPT-5.6-Luna-Pro moves from 6.0 to 3.0 and describes the scene in terms of risk, worry and safety.

(iii) The older persona reminisces. Descriptions add nostalgia, memories and invented grandchildren (Picnic 2, Pigeon 6, Car race 4), sometimes while the rating is unchanged, so the text ages independently of the judgment.

(iv) The middle persona usually copies one end. Across the 60 model panels, the middle rating tends to equal either the young or the older rating rather than lie between them, matching the step pattern in Figure 12.

G.4 EXAMPLES

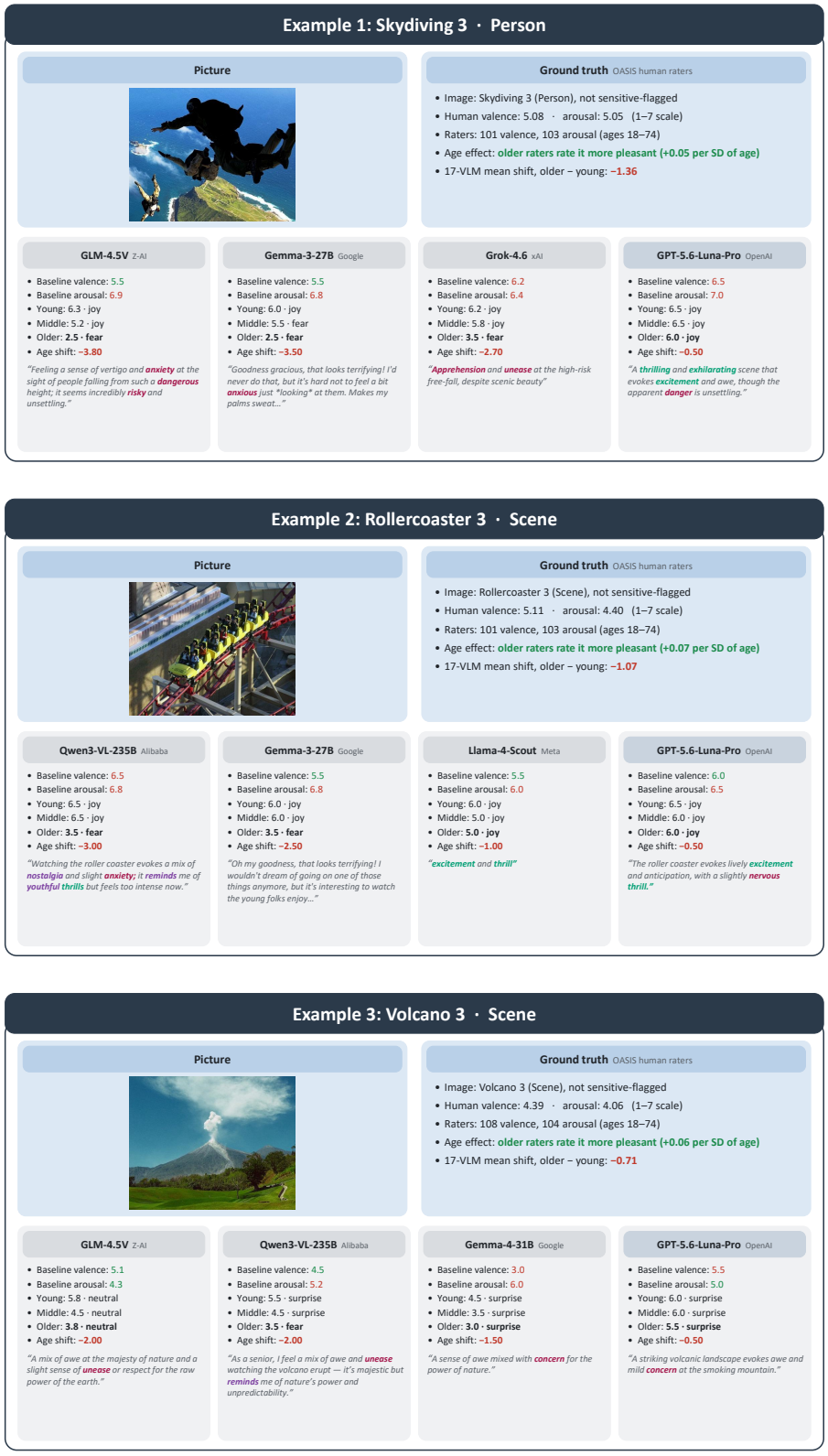
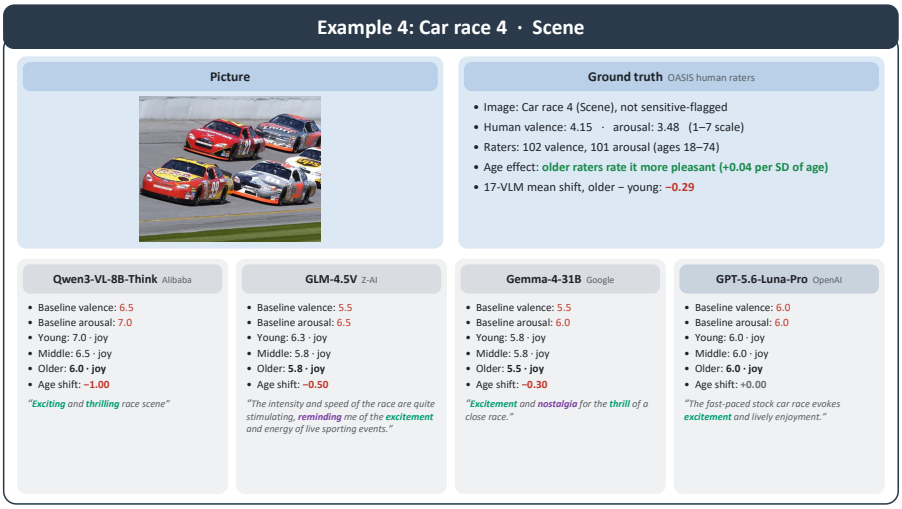


Figure 29: Qualitative examples 1–3. Panel definitions and colour coding are given in Appendix G.





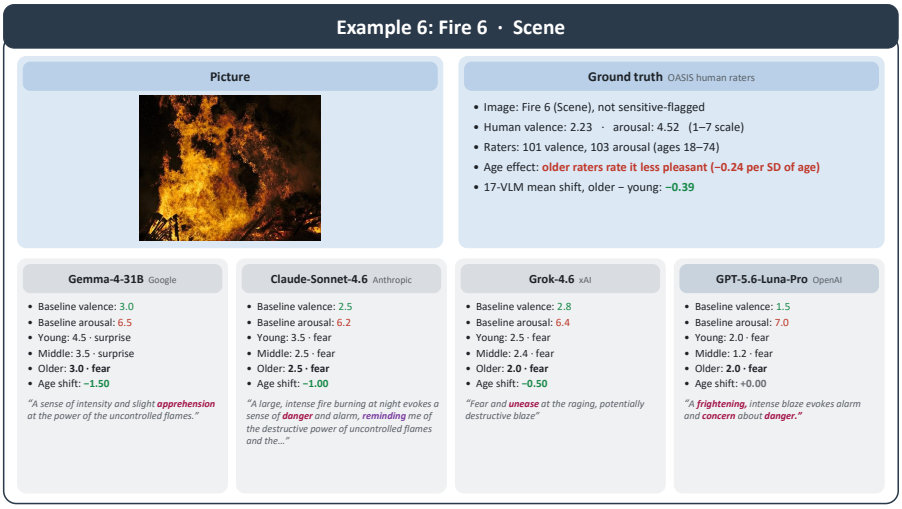


Figure 30: Qualitative examples 4–6. Panel definitions and colour coding are given in Appendix G.

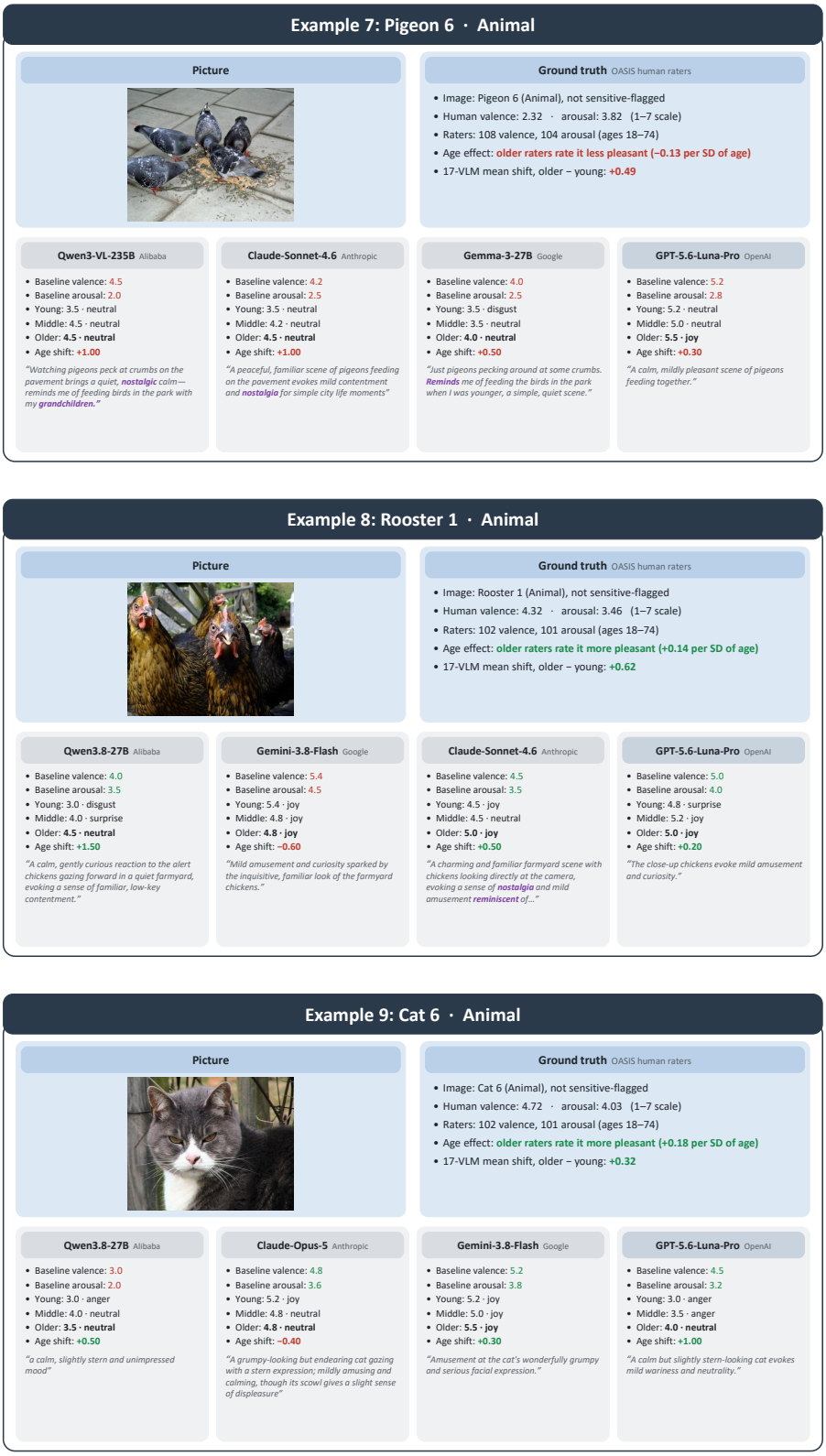


Figure 31: Qualitative examples 7–9. Panel definitions and colour coding are given in Appendix G.

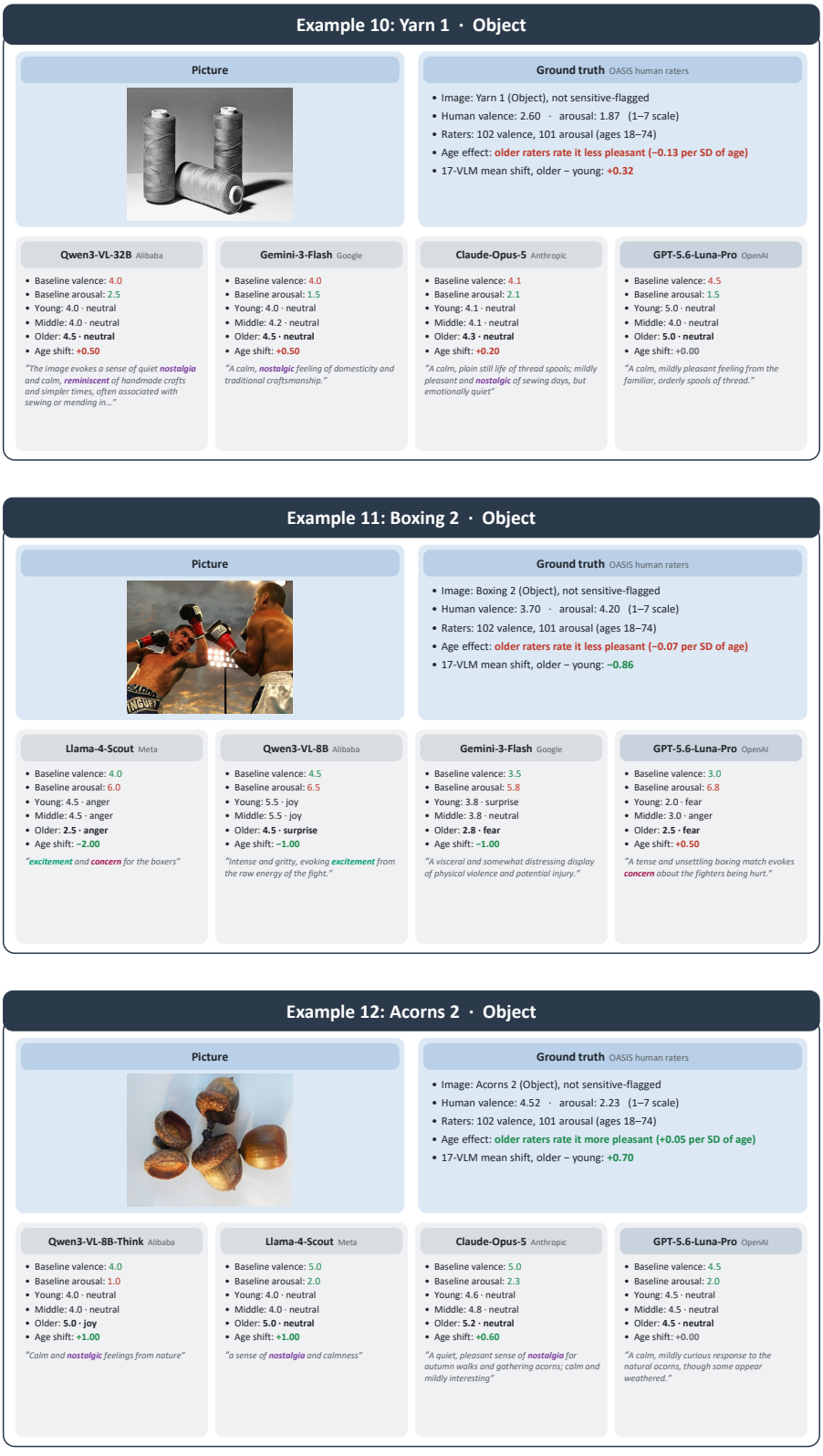


Figure 32: Qualitative examples 10–12. Panel definitions and colour coding are given in Appendix G.



Figure 33: Qualitative examples 13–15. Panel definitions and colour coding are given in Appendix G.