scientific reports

Check for updates

OPEN

Individualized models of social judgments and context-dependent representations

Daniel N. Albohn✉, Stefan Uddenberg & Alexander Todorov

How individuals view the world is critical to understanding human behavior. Yet, almost all research within perception and judgment has drawn inferences from group-level behavior, with little work focused on understanding how *the individual* perceives their world. However, for complex judgments (e.g., trustworthiness), most of the meaningful variance is due to factors specific to the individual. Here we showcase a data-driven reverse correlation method for visualizing any perceptually-derived stereotype at the individual level. We show that our method (1) produces photorealistic and reliable results related to a broad range of judgments, (2) produces valid, psychologically-aligned representations of what individuals are imagining "in their mind's eye", and (3) is capable of capturing visual representations sensitive enough to examine context-dependent categories (e.g., a trustworthy individual to babysit your children vs. to fix your car). Across all studies, we highlight the theoretical implications and utility of developing idiosyncratic models of visual perception.

Keywords Idiosyncratic models, Reverse correlation, Data-driven methods, Generative modeling

> "If the doors of perception were cleansed every thing would appear to man as it is, infinite."
>
> —William Blake, The Marriage of Heaven and Hell.

William Blake's[1] insight is that human perception of reality is limited by one's own experiences and biases: what one perceives is marred by one's past history and subjective interpretations. Yet, much of what is empirically known about preference, judgment, and decision making has been studied and interpreted through the lens of group-level behavior. That is, our understanding of how individuals perceive others[2,3], objects[4,5], scenes[6], and many other facets of our world[7] are based on data aggregated across many participants and observations.

However, group-level estimates (i.e., "averages") can be misleading. When data are averaged together, any "noncompliant" observations are treated as noise and are, essentially, discarded (e.g[8]). In other words, important idiosyncratic differences can be masked when averaging data. In fact, studies that partition the reliable variance of judgments into shared (e.g., due to stimulus features) and idiosyncratic variance (e.g., due to the individual and individual-by-stimulus interactions) show that the latter trumps the former in many cases. For example, studies on attractiveness judgments[9–14] find that idiosyncratic variance exceeds shared variance. The results are even more dramatic for complex judgments such as perceived trustworthiness from facial appearance. These judgments are highly individualized with over 50% of the variance due to idiosyncratic components and less than 5% due to shared components[10]. Likewise, social category judgments such as perceived sexual orientation, political ideology, and religiosity also appear to be largely explained by individual-level contributions[15]. These findings are not limited to studies of facial judgments and extend to diverse domains, including aesthetics[16–18], architecture[19], dancing[20], personality traits[21,22], voices[23,24], and even the evaluation of academic work through peer review[25]. Despite these findings, modeling and interpreting averages continues to be the norm across many subdisciplines of behavioral science[7].

To be clear, the insights gleaned from work that aggregates across participants and observations are fundamental, especially in the domain of modeling or visualizing the configurations of perceptual features that drive particular judgments. One way that group-level averages have been examined within perception science is through computational data-driven methods. These methods are capable of measuring and visualizing how people view stimuli in their "mind's eye" (for reviews see[26–28]), and have proven crucial to understanding the mental representations of a variety of stimuli, including visual stereotypes of social groups, facial emotions, specific facial attributes, identities, and objects[2,29–41].

Booth School of Business, University of Chicago, Chicago, IL, USA. ✉email: Daniel.Albohn@chicagobooth.edu

Computational data-driven models of facial judgments have been particularly successful in making the "ineffable" explicit[38,42]. In essence, these models capture the shared variance in judgments or consistent stimulus features that influence the specific judgment on average. In the case of judgments of trustworthiness, for example, multiple studies have shown that global attributes such as positive expressions and femininity increase these judgments[2,35,43–45], but even for these attributes there is large variation in how individuals weigh them in their individual judgments. The variation is even larger for specific features: while one individual may perceive large eyes and round faces as strong cues for trustworthiness, another individual may perceive smaller eyes and angular faces as such cues[10]. As a result, models that aggregate across participants provide only a limited understanding of perception and judgment.

Here we propose and validate a data-driven method that leverages generative artificial intelligence to capture, visualize, and quantify idiosyncratic representations of any social judgment that individuals hold in their "mind's eye." The method allows for the construction of *psychologically aligned models*: reliable, robust, photorealistic, and representing valid constructs at the level of the individual participant.

All computational data-driven models of judgments are a version of reverse correlation, which identifies a quantitative relationship between high-dimensional variables (e.g., visual stimuli) and judgments[28]. The key to these methods is that the input stimuli are randomly varied – either randomly generating faces from a multi-dimensional face space (e.g[2]), or superimposing randomly generated visual noise on facial stimuli (e.g[38]). These methods quantify and visualize the stimulus variation that is predictive of judgments. However, existing methods have been almost exclusively applied to aggregated judgments or group-level behaviors[2,3,32,34,35,46,47], because individual models cannot be reliably estimated or they are noisy and hard to interpret[48]. In this paper, capitalizing on recent advances in deep learning with respect to modeling face evaluation[3], we combine procedures from the existing data-driven methods to generate rich individual representations of a variety of social judgments. As in face-space-based reverse correlation methods, we randomly generate facial stimuli from a multidimensional space by either projecting real faces into the latent space using various encoding methods (Studies 1 & 2) or sampling directly from the latent space (Studies 3 & 4; Fig. 1). As in psychophysical reverse correlation methods, participants make categorical judgments about each of these stimuli and we use their judgments to build a model of the individual participant's judgment.

In four studies, we show that our method is capable of producing photorealistic and reliable visual models across a broad range of social judgments at both the group and individual levels. Building on prior research[10], in the first three studies we use two types of judgments: those with a clear mapping to physical cues that are consistent across observers (e.g., femininity/masculinity, age) and those with a mapping that is less consistent across observers (e.g., trustworthiness, familiarity). Correspondingly, whereas most of the variance accounting for the former is shared, most of the variance accounting for the latter is idiosyncratic. Irrespective of the type of judgments, we predicted that faces manipulated by models of individual participants would be rated by other participants according to the model's predictions (e.g., faces manipulated to appear more trustworthy would be rated as more trustworthy), but visual models of highly shared judgments would be more similar to each other than visual models of highly idiosyncratic judgments. More importantly, we predicted that for highly idiosyncratic judgments, participants' own visual models would be more predictive of their own ratings compared to visual models of other participants (Studies 2 and 3). Finally, we show that our method of generative reverse correlation is sensitive enough to visualize idiosyncratic representations across context-dependent social judgments (e.g., a trustworthy individual to babysit your children vs. a trustworthy individual to fix your car; Study 4).

Study 1

Generative reverse correlation

The primary objective of Study 1 was to introduce the generative reverse correlation procedure and to demonstrate that the models of individual participants' judgments are interpretable. We have discussed the procedure as a proof-of-concept, creating models of a few participants[9], but we have not tested it in a larger sample and we have not validated these models in a separate group of participants.

We used two judgments–femininity/masculinity and trustworthiness–as these two are clearly different with respect to the relative proportion of shared and idiosyncratic variance. Whereas most of the meaningful variance in femininity/masculinity is shared variance, most of the meaningful variance in trustworthiness judgments is idiosyncratic[9,10]. In the first stage of the experiment (Phase I), participants categorized randomly generated synthetic–but realistic-appearing–faces on the respective judgment dimensions. We used their categorizations to create idiosyncratic visual models of their judgments (Fig. 2A). Consistent with variance partitioning studies[9–11,15,24], we expected that the models of femininity/masculinity judgments would be more similar to each other than the models of trustworthiness judgments.

In the second stage of the experiment (Phase II), a new sample of participants rated faces generated by the models of individual participants from the first stage. The objective was to validate these individual models. We expected that participants' ratings would track with the models' predicted values (e.g., faces manipulated to appear trustworthy would be rated as more trustworthy). Further, given that judgments of trustworthiness and masculinity are negatively correlated[2], we expected that the judgments would be differentially sensitive to the two judgment models (e.g., the slope of trustworthiness judgments would be positive for faces generated from trustworthiness models and negative for faces generated from masculinity models).

Similarity between idiosyncratic visual models

We used cosine similarity to assess the similarity between latent vectors for each of the participants' visual models constructed in Phase I. Because each model is a vector of real numbers in the generative model's latent

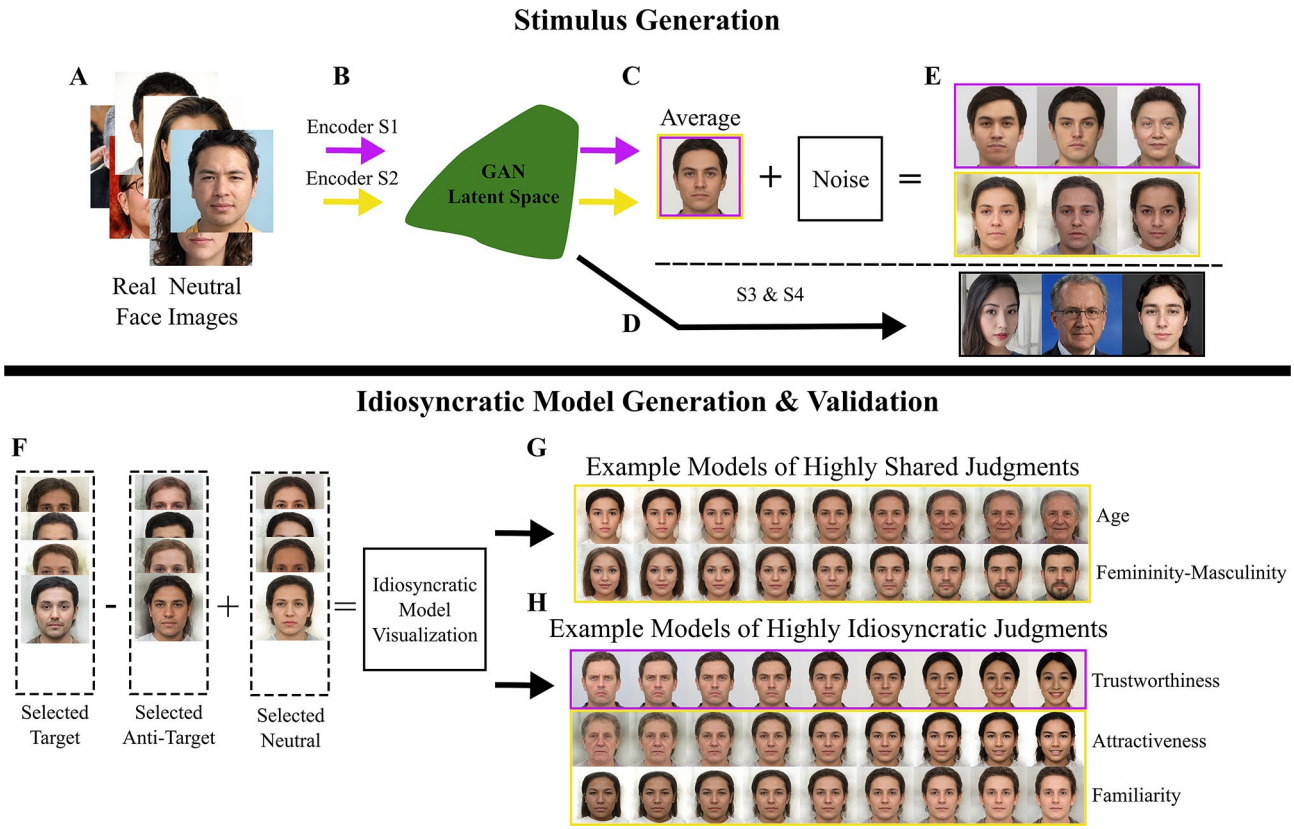

Fig. 1. Generative Reverse Correlation and Validation Methodology. Stimuli for generative reverse correlation (top half of figure) are obtained by either randomly generating them from real neutral faces projected into the generative model's latent space (A and B), averaging together a sample of projected faces, and adding noise (C), as in Studies 1 and 2 (S1, S2), or sampling directly from the generative model's latent space (D), as in Studies 3 and 4 (S3, S4). Examples of images generated through each method are shown in (E). The face images presented in (A) are synthetic faces for illustrative purposes only. Participants classify each of the generated stimuli as either the target category judgment (e.g., "trustworthy"), the conceptually opposite target category judgment (e.g., "untrustworthy"), or a "neutral" category. After participants classify the stimuli, idiosyncratic visual models are created by first averaging the latent values associated with each of the three categories. Next, the average latent vector of the conceptually opposite target judgment is subtracted from the average latent vector of the target judgment. We call this latent vector the "idiosyncratic visual model". To visualize images from each participant's model, the average latent vector of the faces selected as the neutral category is added to the idiosyncratic visual model (F). Finally, by multiplying the idiosyncratic visual model by a constant (e.g., -/+2), different values of the target judgment can be visualized by interpolating through the participant's own latent space. (G) displays examples of images generated from individual participants' models of judgments that are highly shared (i.e., femininity-masculinity and age); and (H) displays examples of images from individual models of judgments that are highly idiosyncratic (i.e., trustworthiness, attractiveness, and familiarity). The center image in each row represents the average of all the latent vectors that the participant selected as the "neutral" category. Images to the left and right of the center image represent the model interpolation values at +/- 2, 4, 6, and 8, respectively.

space, we calculated the average cosine similarity for each participants' individualized model vector and every other participants' model vector.

Consistent with previous variance partitioning studies that show more consensus for judgments of femininity and masculinity compared to judgements of trustworthiness[9–11], the average similarity of feminine-masculine idiosyncratic visual models was significantly higher than that of trustworthy-untrustworthy idiosyncratic visual models, $t(46.98) = 21.98$, $p < .001$, $d = 5.09$ (Fig. 3A).

Validation of idiosyncratic models

Across all studies, validation data were analyzed using linear mixed-effects regressions with fixed effects for type of visual model (e.g., the intended target judgment) and model value (i.e., linear interpolation value), along with random intercepts for participant and image (full linear mixed-effects regression tables are provided in the Supplemental Material). All validation studies included repeat trials used to assess participants' test-retest reliability. Participant scores were averaged over repeat observations for each validation regression.

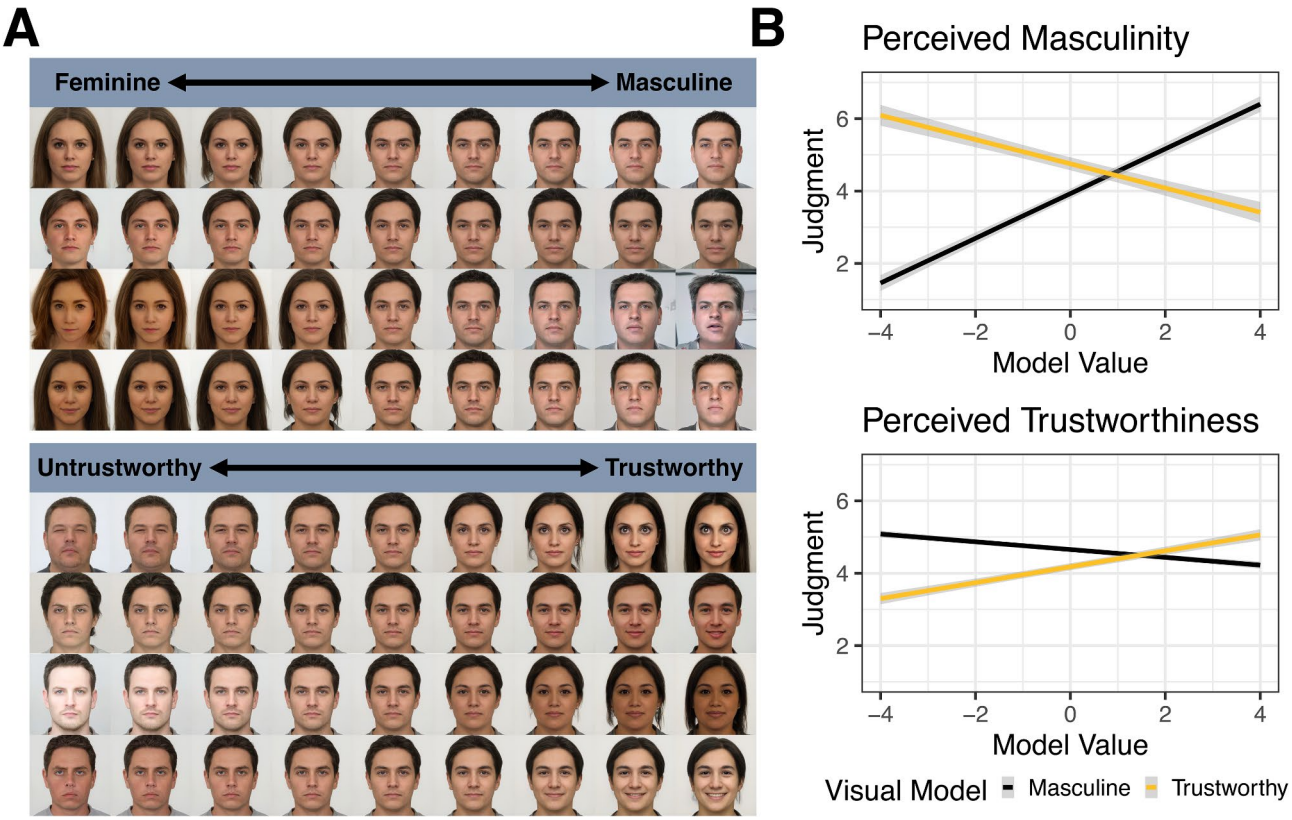

Fig. 2. Validation results of Study 1. Each row in (A) displays example images generated from Phase I participants' idiosyncratic visual models for each condition. The center image in each row represents the average of all of the latents each participant selected as the "neutral" category. Images to the left and right of the center image represent the linear interpolation at +/- 2, 4, 6, and 8, respectively. All included participants' models can be viewed in the Visualization Supplemental Materials. (B) displays the Phase II validation results whereby a second group of participants judged the +/-4 and +/-2 idiosyncratic visual model images generated in Phase I on how "masculine" or "trustworthy" each appeared. The x-axis represents the model interpolation values and the y-axis represents participants' responses. Shaded areas around each line display 95% confidence intervals.

Perceived masculinity

Phase II participants judged the faces produced from each Phase I visual model as intended (Fig. 2B). Faces manipulated to appear more masculine were rated as more masculine. Faces manipulated to appear more trustworthy were rated as less masculine. The latter finding reflects the fact that masculine faces are perceived as less trustworthy[2]. These findings were reflected in a significant interaction between visual model type (feminine-masculine vs. trustworthiness) and model value (-4 to + 4), $b = 0.96$, $t(256.14) = 27.14$, $p < .001$ (the main effect of visual model was also significant, $b = 0.83$, $t(181.69) = 7.39$, $p < .001$). Simple slopes analyses for each visual model showed a positive and significant effect for feminine-masculine visual models, $b = 0.57$, $t(258) = 25.95$, $p < .001$, and a negative and significant effect for trustworthy visual models, $b = -0.34$, $t(254) = 12.94$, $p < .001$.

Perceived trustworthiness

The pattern of findings was similar to judgments of trustworthiness. Faces manipulated to appear more trustworthy were rated as more trustworthy, whereas faces manipulated to appear more masculine were rated as less trustworthy. The interaction between model type (feminine-masculine vs. trustworthiness) and model value (-4 to + 4) was significant, $b = 0.32$, $t(260) = 19.94$, $p < .001$ (the main effect of model was also significant $b = -0.46$, $t(260.56) = 9.08$, $p < .001$). Simple slopes analyses for each visual model showed a positive and significant effect for trustworthy visual models, $b = 0.21$, $t(253) = 17.98$, $p < .001$, and a negative and significant effect for feminine-masculine visual models, $b = -0.11$, $t(269) = 9.98$, $p < .001$.

Taken together, these results show that images produced from the idiosyncratic models in Phase I were judged as intended across each model value. In other words, images generated through participants' visual models of feminine-masculine or trustworthy-untrustworthy images both qualitatively (i.e., through visual inspection) and empirically represented those categories, replicating our previous research[9].

While our generative reverse correlation methodology shows promise in terms of both the construction of idiosyncratic visual models and their validation by other observers, a stronger validation would be to show that images produced from a participant's own visual model are judged *by that same participant* as more

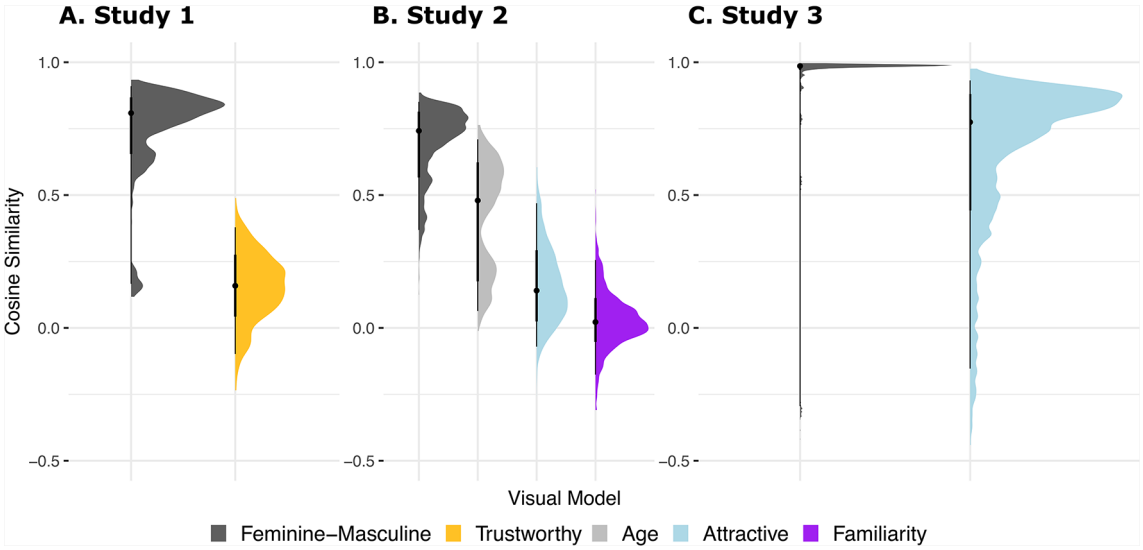

Fig. 3. Distribution of the cosine similarities for Studies 1–3 (Panels A - C, respectively). The cosine similarity (y axis) was calculated by taking the average similarity of each participant's idiosyncratic visual model and every other participant's visual model within a particular judgment category (x-axis; colored distributions). Across all studies, we predicted that the similarity for highly shared judgments (e.g., feminine-masculine, age) would be larger than highly idiosyncratic judgments (e.g., trustworthy, attractive). Error bars represent 95% confidence intervals.

representative of their judgments than images from other participants' visual models. Thus, in Study 2, we have the same participants who created the visual models return and judge both their own and a random sample of images generated from other participants' visual models.

Study 2

Idiosyncratic validation of Generative Reverse correlation

The objective of Study 2 was to evaluate the validity of generative reverse correlation at the individual participant level and across a more diverse set of judgments. Instead of having an additional set of participants judge each resultant visual model image on the target categories, we had the same participants return to evaluate their own visual models. We selected femininity/masculinity and age to represent judgments high in shared variance and attractiveness and familiarity to represent judgments high in idiosyncratic variance. We expected that images created from participants' own visual models (as opposed to models from other participants) would be more predictive of their subsequent judgments of images generated by models of highly idiosyncratic judgments (i.e., attractiveness and familiarity), but not necessarily of images generated by models of shared judgments (i.e., masculinity and age). In addition, consistent with Study 1, we expected that visual models of highly shared judgments to be more similar with one another than visual models of highly idiosyncratic judgments. For this study, we changed the method used to project neutral faces into the latent space to retain more detail and increase diversity of the generated stimuli (Fig. 4A).

Similarity between idiosyncratic visual models

Like Study 1, we used cosine similarity to assess the similarity between latent vectors for each of the participants' visual models constructed in Phase I. However, we grouped the four judgments used in Study 2 into two groups for analysis: judgments that are highly shared (feminine-masculine and age) and judgments that are highly idiosyncratic (attractiveness and familiarity). We replicated the finding from Study 1 and previous variance partitioning studies: The average similarity of visual models of highly shared judgments was significantly higher than the similarity of visual models of highly idiosyncratic judgments, $t(176.51) = 26.42$, $p < .001$, $d = 3.34$ (Fig. 3B).

Idiosyncratic visual model validation

Highly Shared judgments: Age and feminine-masculine visual models

Participants rated the images created by the models to appear "older" as older, evidenced by a main effect of model value, $b = 0.48$, $t(50.17) = 16.53$, $p < .001$. There was no main effect of visual model type (a participant's own model vs. other participants' models), $b = 0.01$, $t(50.07) = 0.07$, $p = .948$, or an interaction, b = -0.05, $t(50.14) = 1.59$, $p = .117$, suggesting that participants' ratings were not more sensitive for their own than for other participants' models (Fig. 4B).

Likewise, participants rated the images manipulated by the models to appear more "masculine" as more masculine, $b = 0.55$, $t(50) = 8.47$, $p < .001$. There was no main effect of visual model type (own vs. other), $b = 0.03$,

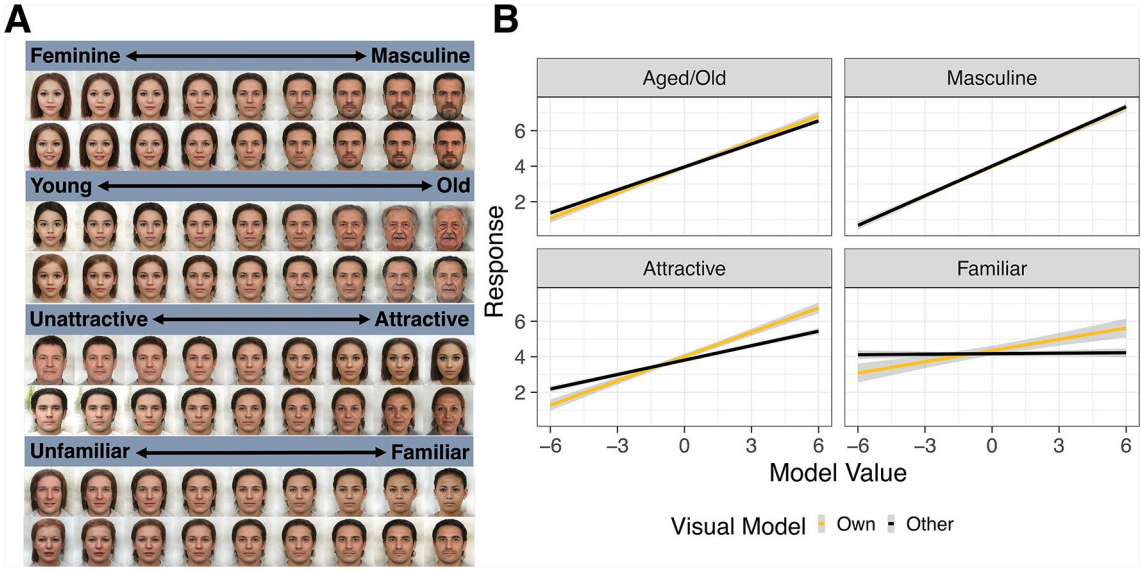

Fig. 4. Validation results of Study 2. (A) displays examples of images generated from idiosyncratic visual models of two participants across each judgment condition. The center image in each row represents the average of all of the latents each participant selected as the "neutral" category. Images to the left and right of the center image represent the linear interpolation at +/- 2, 4, 6, and 8, respectively (see Visualization Supplemental Materials for all individual and group visual models). (B) displays the validation results. Each participant judged images generated by their own visual model (yellow lines) and those generated from a random subsample of other participants' visual models (black lines). The x-axis represents the linear interpolation model values and the y-axis represents participants' raw judgment responses. Shaded areas around each line display 95% confidence intervals.

$t(50) = 0.10$, $p = .923$, or an interaction, $b = 0.004$, $t(50) = 0.05$, $p = .958$, again suggesting no participant-level discrimination between their own model images and others' model images.

These two results suggest that idiosyncratic visual models of social judgments that have high agreement, or high amounts of shared variance, are not any more predictive than random participants' visual models. However, this does not mean these models are not representative of the target judgment. On the contrary, they appear able to visually capture the shared agreement inherent in these judgments across observers.

Highly idiosyncratic judgments: attractiveness and familiarity visual models

Participants rated the images manipulated by the model to appear more "attractive" as more attractive, $b = 0.45$, $t(49.31) = 15.91$, $p < .001$. There was no main effect of visual model type (own vs. other), $b = -0.19$, $t(49.83) = 1.71$, $p = .093$. However, there was a significant interaction, $b = -0.18$, $t(49.37) = 5.76$, $p < .001$. Participants' judgments were more sensitive to their own visual models ($b = 0.45$) than to other participants' visual models ($b = 0.27$). In particular, they judged faces at high values of their model as more attractive than faces at high values of other participants' models, and vice versa for faces at low values of the models.

Similarly, images manipulated by the model to appear more "familiar" were rated as more familiar, $b = 0.21$, $t(50) = 5.57$, $p < .001$. There was no main effect of visual model type (own vs. other), $b = -0.18$, $t(50) = 1.20$, $p = .236$, but there was a significant interaction, $b = -0.20$, $t(50) = 4.86$, $p < .001$. Participants judged images manipulated by their own models to appear more familiar ($b = 0.21$) as more familiar, but not faces manipulated by other participants' models ($b = 0.01$).

Together, these results further show that our method produces psychologically aligned visual representations across a variety of highly shared and highly idiosyncratic judgments. Specifically, participants judged images generated from their own visual models as more representative of the target judgment compared to images generated from others' visual models. However, this only appears to be the case when the judgment being evaluated is high in idiosyncratic variability, such as attractiveness and familiarity. On the other hand, judgments that are known to have high agreement (i.e., high amounts of shared variance), such as age or femininity/masculinity, do not show individualized preferences. Participants judged older and masculine visual representations similarly across model types (i.e., their own model vs. other participants' models).

Studies 1 and 2 provide strong evidence that the generative reverse correlation procedure is capable of visually capturing a diverse set of social judgments that not only appear like the judgment being examined but are also more predictive of the individual participant's own preferences. However, it is still unknown whether generative reverse correlation is dependent on the underlying model's latent space used to generate the images. Similarly, it is also unknown whether this methodology is capable of visualizing judgments beyond broad social judgment categories. In the next two studies we address both of these issues.

Study 3

Invariance to Model Latent Space

The stimuli for Studies 1 and 2 were created by projecting real faces into the latent space of a pretrained face model. This was done to ensure that each study had a quality stimulus set that was neutral in expressivity, as the StyleGAN-2 FFHQ latent space used in the previous studies has an overrepresentation of smiling faces. However, projecting neutral faces into the model's latent space artificially constrains it to the subsection bounded by the faces projected into it. This results in less variability in stimulus appearance. The objective of Study 3 was to test generative reverse correlation with stimuli sampled from a larger, more heterogeneous latent space trained exclusively on neutral and minimally expressive face images (see Supplemental Materials for full training details). Importantly, if generative reverse correlation is robust, valid, and generalizable, the results should be invariant to the underlying model and latent space used to create the stimuli.

The methods for Study 3 were nearly identical to those of Study 2. We had participants categorize faces randomly generated from a latent space that was trained on faces that were only neutral in appearance. As before, we selected two types of judgments to focus on: "feminine-masculine" as the highly shared judgment and "unattractive-attractive" as the highly idiosyncratic judgment. After categorization, we constructed idiosyncratic visual models for each participant and had them return to rate images manipulated by their own model and images manipulated from a random selection of other participants' models (Fig. 5A).

Similarity between idiosyncratic visual models

As before, we replicated Studies 1 and 2 showing that the average similarity of feminine-masculine visual models was significantly higher than the average similarity of attractiveness visual models, $t(116.3) = 8.23$, $p < .001$, $d = 1.50$ (Fig. 3C).

Idiosyncratic visual model validation

Ratings of masculinity

As in Study 2, participants rated the images manipulated by the models to appear more "masculine" as more masculine, $b = 1.47$, $t(26) = 6.22$, $p < .001$ (Fig. 5B). There was no main effect of visual model type (participant's own model vs. other participants' models), $b = 0.05$, $t(26) = 0.15$, $p = .879$, or an interaction, $b = 0.06$, $t(26) = 0.22$, $p = .831$.

Ratings of attractiveness

Participants rated the images manipulated by the models to appear more "attractive" as more attractive, $b = 1.09$, $t(26) = 11.15$, $p < .001$ (Fig. 5B). There was no main effect of visual model type (own vs. other), $b = -0.02$, $t(26) = 0.11$, $p = .911$, but there was a significant interaction, $b = 0.28$, $t(26) = 2.59$, $p = .016$. Participants' judgments were more sensitive to their own visual models ($b = 1.09$) than to other participants' visual models ($b = 0.82$).

Together, these results replicate those in Study 2 but with one important insight: generative reverse correlation appears to be invariant to the underlying latent space. Our previous method required projecting real neutral faces into a pretrained model's latent space in order to have a robust stimulus set that was not oversaturated with smiling faces. However, this approach limited the diversity of the stimuli generated. To address this, we trained

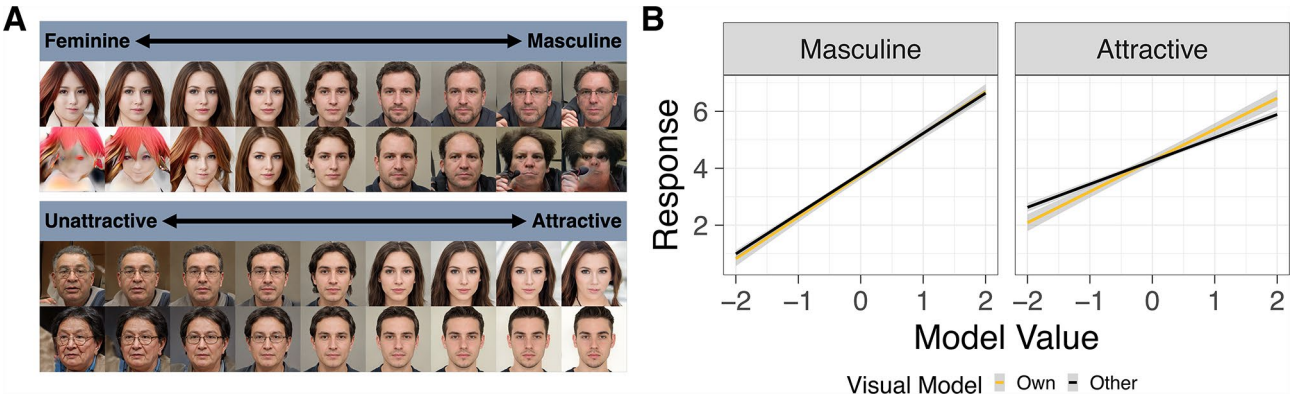

Fig. 5. Validation results of Study 3. (A) displays examples of images generated from each idiosyncratic visual model across conditions. The middle image represents the average of all faces participants categorized as "unsure". Each image to the right and left of the middle represents a +/-1 model interpolation step value, respectively. The second example participant's visual model in the feminine-masculine condition (second row from top) shows an example of a model going out-of-bounds at the extremes (> +/-3). (B) displays the validation results. Each participant judged images generated from their own visual model (yellow lines) and those generated from a random subsample of other participants' visual models (black lines). The x-axis represents the model interpolation values and the y-axis represents participants' responses. Shaded areas around each line display 95% confidence intervals.

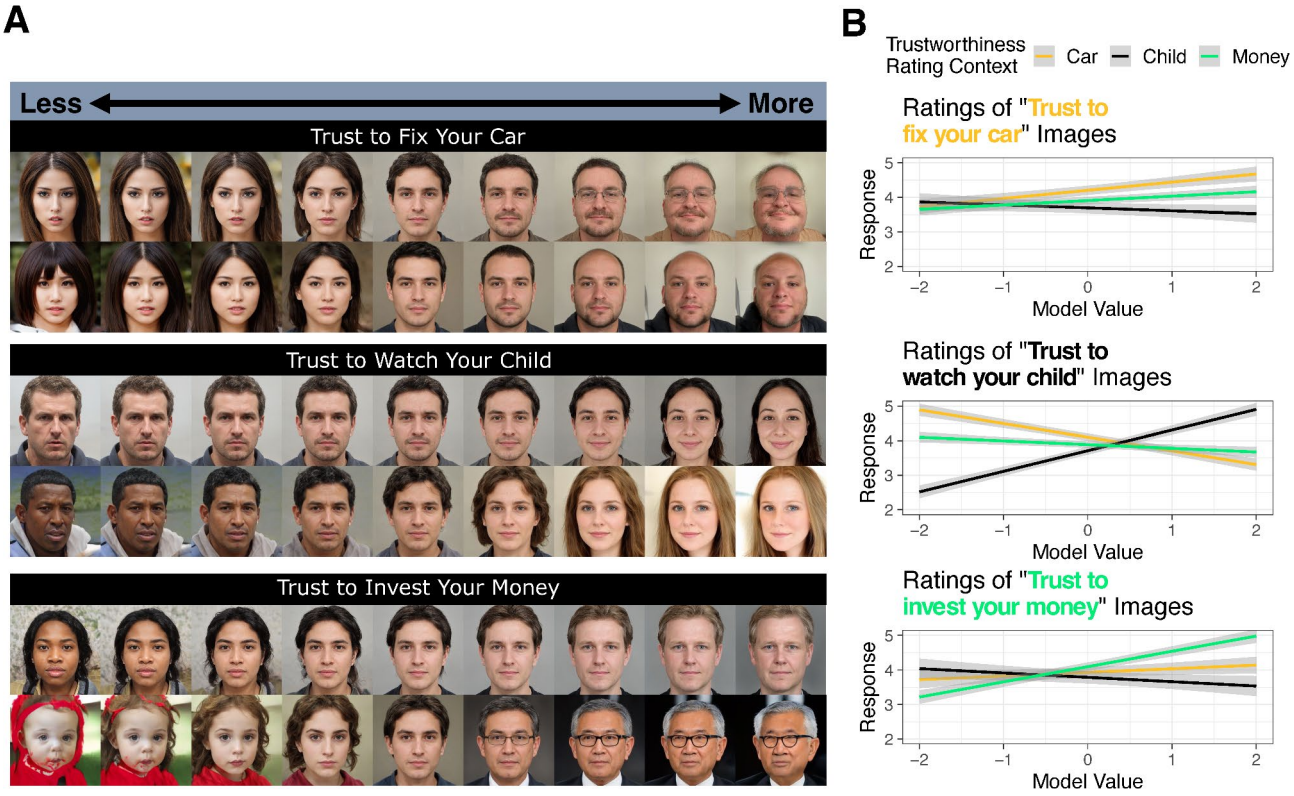

Fig. 6. Validation results of Study 4. (**A**) displays examples of idiosyncratic visual models generated from participants' context-dependent responses. The middle image in each row represents the average of all faces that the example participants categorized as the "neutral" category. Each image to the right and left of the middle represents a +/-1 model interpolation value step, up to +/-4, respectively. All included participants' models can be viewed in Visualization Supplemental Materials. (**B**) shows the validation results for across all three context-dependent judgment conditions (from top to bottom: "trust to fix your car" "trust to watch your child", and "trust to invest your money"). Line colors represent the condition in which the Phase II participants judged each of the images, irrespective of the Phase I image context (i.e. all images were judged on all "trustworthiness" contexts). The x-axis represents the model interpolation values and the y-axis represents participants' responses. Shaded areas around each line display 95% confidence intervals.

a new model using nearly 48,000 faces with neutral appearance. The resulting model is able to create a diverse range of novel faces with neutral expressions while maintaining a broad range of other desirable characteristics, such as perceived age and race (see Supplemental Material). Importantly, the results from this study replicated the results from Study 2. Participants' own visual models of attractiveness (but not femininity-masculinity) were more predictive of their subsequent judgments.

Study 4

Context-dependent representations

Studies 1 through 3 provide strong evidence that the generative reverse correlation is capable of visually capturing a diverse set of social judgments that not only appear like the judgment being examined, but are also more predictive of the individual participant's own preferences. In Study 4, we test whether generative reverse correlation is capable of capturing representations beyond broad social judgments by examining the sensitivity of idiosyncratic visual models to context-dependent evaluations of trustworthiness. For example, the mental prototype of who to trust to watch your child is likely different from the prototype of who to trust to fix your car, despite both being derived from evaluations of "trustworthiness." In Phase I, we randomly assigned participants to make context-dependent trustworthiness judgments–who they would trust to "fix their car", "watch their child", or "invest their money"–and used their judgments to create idiosyncratic visual models. In Phase II, a separate group of participants judged the images generated from each Phase I participant on how much they trusted each individual depicted to "fix their car", "watch their child", or "invest their money".

"Trust to Fix Your Car" Images.

Across all three Phase I conditions, participant ratings were in line with the visual model from which the images were created (Fig. 6B). Faces manipulated to appear more trustworthy in the car context were perceived as more trustworthy within that context compared to faces manipulated in both the "trust to watch your child" images and "trust to invest your money" contexts, as evidenced by significant two-way interactions between visual model type and model value ("trust to fix your car" vs. "trust to watch your child": $b = -0.32$, $t(4620.02) = 11.57$,

$p < .001$; "trust to fix your car" vs. "trust to invest your money": $b = -0.11$, $t(4617.50) = 3.97$, $p < .001$). There was also a significant main effect for "trust to fix your car" images vs. "trust to watch your child" images, $b = -0.45$, $t(217.68) = 2.84$, $p = .005$, but not "trust to fix your car" images vs. "trust to watch your money" images, b = -0.24, $t(218.37) = 1.54$, $p = .126$.

Simple slopes analyses for each visual model showed positive and significant slopes for "trust to fix your car" images, $b = 0.23$, $t(479) = 8.53$, $p < .001$, and "trust to invest your money" images, $b = 0.12$, $t(427) = 4.45$, $p < .001$, and a significant negative slope for "trust to watch your child" images, b = -0.10, $t(397) = 3.72$, $p < .001$.

"Trust to Watch Your Child" Images.

A similar pattern emerged for images manipulated to appear more trustworthy within a "watch your child" context. There was a significant difference in slopes between "trust to watch your child" images and both "trust to fix your car" (b = -0.97, $t(4546.11) = 37.41$, $p < .001$) and "trust to invest your money" images (b = -0.71, $t(4540.58) = 28.54$, $p < .001$). There was a significant main effect for "trust to watch your child" images vs. "trust to fix your car" images, $b = 0.43$, $t(217.84) = 2.78$, $p = .006$, but not "trust to watch your child" images vs. "trust to invest your money" images, $b = 0.22$, $t(217.47) = 1.46$, $p = .145$.

Simple slopes analyses for each visual model showed a positive and significant slope for "trust to watch your child" images, $b = 0.58$, $t(392) = 24.18$, $p < .001$, and significant negative slopes for "trust to fix your car" images, b = -0.39, $t(462) = 4.81$, $p < .001$, and "trust to invest your money" images, b = -0.12, $t(382) = 5.16$, $p < .001$.

"Trust to Invest Your Money" Images.

The pattern was also similar for images manipulated to appear more trustworthy within a "trust to invest your money" context, as indicated in significant differences between the slopes ("trust to invest your money" vs. "trust to fix your car": $b = -0.55$, $t(3256.21) = 17.04$, $p < .001$; "trust to invest your money" vs. "trust to watch your child": $b = -0.29$, $t(3241) = 8.88$, $p < .001$). There was also a significant main effect for "trust to invest your money" images vs. "trust to watch your child" images, $b = -0.328$, $t(217.45) = 2.14$, $p = .034$, but not "trust to fix your car" images, b = -0.12, $t(216.60) = 0.74$, $p = .463$.

Simple slopes analyses for each visual model showed positive and significant slopes for "trust to invest your money" images, $b = 0.43$, $t(282) = 14.46$, $p < .001$, and "trust to fix your car" images, $b = 0.14$, $t(343) = 4.47$, $p < .001$, but a significant negative slope for "trust to watch your child" images, b = -0.12, $t(319) = 3.85$, $p < .001$.

Previous research on modeling judgments has focused almost exclusively on gestalt social judgments (e.g., judging "trustworthiness" or "attractiveness" without further context), rather than on situation- or context-dependent social judgments. We show that generative reverse correlation provides a way to accurately capture visual representations that are highly sensitive and context-dependent.

Discussion

If two randomly selected individuals were asked to evaluate the attractiveness, trustworthiness, or familiarity of a face, the probability that they would agree with each other is low. Individuals bring their own biases, learned experiences, history, and desires when making complex judgments. Correspondingly, variance partitioning studies show that the variance of such judgments due to idiosyncratic factors exceeds the variance due to stimulus features[10,11,13,15,24]. Yet decades of research within social, cognitive, and perception science have largely ignored idiosyncratic contributions to judgments. Rather, research has focused on drawing inferences from group-level averages and treated idiosyncratic differences as noise.

Despite the documented importance of investigating idiosyncratic contributions to judgments, little work has attempted to explain what causes these differences and the research that has attempted to explain it has shown that predicting these idiosyncratic differences is difficult[10,49,50]. To address these limitations, we developed and validated a data-driven, reverse correlation method that leverages generative artificial intelligence to accurately and reliably visualize photo-realistic representations of social judgments from faces. Across four studies, we show that our method is capable of producing psychologically aligned visual representations of social judgments. Highly idiosyncratic judgments are important to study, yet difficult to predict. Our work allows for visualizing these idiosyncratic contributions, which can then be used to systematically investigate differences in perception and judgment beyond descriptive or predictive modeling.

Generative artificial intelligence provides an opportunity to leverage advances in computer science to answer important questions, particularly with respect to individualized models of behavior. Generative modeling, in conjunction with reverse correlation, allows for all of the benefits of data-driven reverse correlation, such as less potential experimenter bias, but with the addition of realistic appearing stimuli–both for the participant to categorize and the resulting models that are computed, studied, and visualized. This is in contrast to many of the previous studies and methods that require averaging over the entire set of images generated. Producing and visualizing psychologically aligned constructs at the individual level represents a significant step forward for accurately revealing individualized preferences across a host of judgments and contexts.

Using generative reverse correlation is a promising avenue for visualizing and examining both shared and idiosyncratic models of perception. Future work could extend the study of idiosyncratic contributions to a diverse set of judgments, including, but not limited to, health, occupation, or any type of social identity. Additionally, this process could be adapted to create individualized clinical training tools, for example, to help those with body dysmorphia or emotion recognition deficits.

Similarly, extending this work beyond faces would allow for systematic investigations of highly shared and idiosyncratic judgments beyond person perception. With a model trained on consumer products, for example, one could create bespoke representations of the ideal product for individuals to engage with or purchase. Such models can also be trained on novel shapes[16] and used to facilitate the creation of aesthetic designs tailored to particular groups of people. The possibilities for studying idiosyncratic representations across a myriad of domains is limited only by acquiring a high quality set of stimuli for training new generative models.

Lastly, there are a number of opportunities for future research to explore methodological improvements for generative reverse correlation. For instance, while we explored the number of trials needed for accurately visualizing results (see Supplemental Materials), additional research could implement an adaptive staircase procedure to determine optimal stopping on a per-participant basis. Likewise, face images generated from both StyleGAN-2 latent spaces do not control for low-level features like image background, which likely varies idiosyncratically and might influence evaluations. In principle, it would be possible to train a model without any backgrounds, remove all background features before participants make judgments, or more systematically incorporate these lower-level features into the model before validation.

How individuals perceive and interpret the world around them is critical for understanding human behavior. Individuals have their own biases, past history, wishes, desires, motives, and experiences–all of which Blake would argue obscure the doors of perception. Recent research has documented the importance of studying idiosyncrasies in perception and judgment. Yet, much of what is known about perception, judgment, and decision making is inferred from group-level averages of human behavior, which mask important idiosyncratic differences. Part of the issue is that idiosyncratic contributions to judgments are relatively easy to identify, but hard to formally model and predict. The methods described here provide the tools for visualizing idiosyncratic representations in a data-driven manner. These idiosyncratic visual models are robust, photorealistic, and psychologically aligned representations of how individuals perceive their world. Furthermore, these methods are highly extendible, opening up exciting potential to research how different sources of variance influence preference, judgments, and decisions beyond person perception.

Methods

Study 1

Participants

Seventy-six participants (M_{age} = 39.61, SD_{age} = 10.30) were recruited through CloudResearch for the first stage of the experiment. Participants in Phase I self-identified as follows: 26 women, 47 men, 1 non-binary, 2 not reported; 6 Asian, 12 Black, 2 Latinx, 49 White, 3 more than one race, and 4 other/not reported. One-hundred and ten participants (M_{age} = 39.63, SD_{age} = 10.26) were also recruited through CloudResearch to judge the stimuli generated by the visual models of participants from the first stage. Participants in Phase II self-identified as follows: 43 women, 62 men, 3 non-binary, 2 other/prefer not to answer; 4 Asian, 19 Black, 3 Latinx, 76 White, 6 more than one race, and 2 other/not reported. All studies using CloudResearch participants (Studies 1, 3, and 4) were prescreened to be located within the United States and were "CloudResearch approved"[51].

We randomly selected 10% of the trials in Phase I and all of the trials in Phase II to be repeated in order to assess test-retest reliability for each participant. Participants were excluded if they had negative or near zero ($r < .05$) test-retest correlation. Based on this criterion, 11 participants' visual models from Phase I and 16 participants from Phase II were removed from analyses.

Phase I: Model Construction

Constructing idiosyncratic visual models

The method for constructing idiosyncratic visual models shares some similarities to typical psychophysical reverse correlation procedures[2,35]. However, there are several key differences that take advantage of generative models. Here we briefly describe the process for constructing idiosyncratic (and group-level) visual models and refer the reader to previous work[9] and the Supplemental Materials for more details. Model construction used Python version 3.6.13 (https://www.python.org/).

The method begins by generating a set of random stimuli from a pretrained generative model either by projecting real faces into its latent space and adding Gaussian noise (Studies 1 and 2) or sampling directly from the latent space (Studies 3 and 4). In the current set of studies, we used a version of StyleGAN-2 to produce photorealistic stimuli of faces[52,53]. Participants categorize each of the sampled stimuli into one of three categories: (1) the target judgment (e.g., masculine), (2) the conceptually opposite target judgment (e.g., feminine), or (3) a neutral, neither, or unsure category. The neutral or unsure category was included as prior work has shown that what individuals consider "neutral" also varies idiosyncratically[31,32]. Because each stimulus is represented as a matrix of numeric values in the latent space, idiosyncratic models for each participant are computed through matrix arithmetic: the average of all the stimulus latents selected as the anti-target is subtracted from the average of all the stimulus latents selected as the target attribute, and the average of all the stimulus latents selected as "neutral/neither/unsure" is added back. This results in a directional vector in the model's latent space that can be traversed using linear interpolation (i.e., the model values) to visualize each participant's idiosyncratic model at varying intensities.

Stimuli

We generated 300 neutral face stimuli by projecting real neutral faces into the latent space, averaging 10 faces together, and adding noise to further differentiate each face from the real faces (see Supplemental Materials for more details). We sampled noise from a Gaussian distribution with parameters $\mu = 0$ and $= 0.4$ to be added to each averaged image. Next, we computed individualized images for each of the participants using linear interpolation (i.e., model values) with values ranging from −8 to +8. If participants did not categorize any stimuli as "neutral/neither", a random sample of 25 stimuli were averaged together to act as a starting point in the latent space and added to their idiosyncratic visual models. Example visualizations from individualized models are presented in Fig. 2A.

The removal of 11 poor quality participants from Phase I resulted in a total of 260 images generated by the idiosyncratic visual models to be judged by participants in Phase II (65 participants × 4 images each at model values −4, -2, +2, and +4).

Participant Procedure

Participants read brief instructions that stated the nature of the task and asked to take a moment and briefly imagine an individual that represented the category and the conceptually opposite category to which they were randomly assigned. Participants were then presented with each face one at a time and asked to categorize each stimulus into one of the three categories assigned to them using the "E", "I", or "Space" keys on their keyboard. Every participant saw the same 300 faces, though the presentation order was randomized between participants. Thirty faces were selected at random for each participant to judge twice as a measure of quality assurance (test-retest reliability). There was a 100 ms fixation cross between trials.

After the main portion of the study, participants filled out demographic information (age, gender, race), answered two debrief questions ("What did you think this study was about?" and "Did you notice anything odd or off about the images?"), and were told the purpose of the study.

Phase II: validation of individual models

The linearly interpolated images at values of -4, -2, +2, +4 were selected from each Phase I participant's model to act as stimuli for Phase II ($N = 304$; 76 Phase I participants x 4 images each [Due to a coding error, participants in Phase II judged all images from Phase I, though the poor quality participants' images were removed from analysis.]). We decided to only include model images at values up to +/- 4 to prevent participants in Phase II from judging images with potential artifacts (i.e., the visual model going out of sample when interpolating). All reported statistics were analyzed using R version 4.4.0 (https://cran.r-project.org/).

Participant Procedure

Like in Phase I, participants first were given brief instructions detailing the condition to which they were assigned (i.e., judgments of masculinity or trustworthiness) and instructed that they would be making judgments for a number of face stimuli. Participants were randomly assigned to judge the "masculinity" or "trustworthiness" of a random sample of 60 images from all of those generated in Phase I (regardless of the condition that the Phase I participant was assigned). Phase II participants saw each image presented one at a time and asked, "How [masculine/trustworthy] does this individual appear?" Participants responded using a 7 point Likert-type scale by either pressing a number on their keyboard or by selecting the corresponding number displayed along a scale under each image. The scales had anchors 1 = "Not at all [judgment]" and 7 = "Very [judgment]". Participants rated each face twice (across two randomized blocks) as a measure of quality assurance. Each face used in the reported analyses was judged by an average of 21.73 participants (median = 21, $SD = 4.06$, range = [11, 35]). There was a 100 ms fixation cross between trials. Lastly, participants filled out the same demographic information and debriefing questions as in Phase I.

Study 2

Participants

This study was conducted both in person and online through the authors' university participant pool. Due to the asynchronous and time-consuming nature of this study, some participants started but did not finish the study [One participant failed to complete the study multiple times.] ($N_{incomplete} = 16$) or took the survey multiple times ($N_{restart} = 36$), causing multiple responses across different conditions for the same participants. Because of this, we only included in the analyses participants who completed the study in its entirety and did not restart. This left us with a final sample of 211 participants across each of the four conditions: young/old = 49; feminine/masculine = 64; un/attractive = 56; un/familiar = 42). One hundred and sixteen out of the 211 participants completed both parts of the study.

As in Study 1, we computed test-retest reliabilities for each participant. In Phase I (image model generation), trials were grouped into subsets of 100 trials and randomly presented to participants within each subgroup. Within each subgroup, 10% of images were randomly selected to be shown twice to participants in order to calculate test-retest reliabilities. For Phase II (image ratings), participants judged all images twice across both blocks. Six participants had negative or near zero test-retest reliabilities ($r < .05$). The final sample of participants that completed both Phase I and Phase II and used for the validation analyses were: young/old = 34; feminine/masculine = 33, un/attractive = 24, and un/familiar = 19 [Excluding poor quality, repeat, and incomplete participants reduces the final sample for the attractive and familiar conditions below our target of 30 participants per condition. However, including all participants in the analyses does not change the significance, direction, or interpretation of the results presented.].

Participants ($M_{age} = 29.22$, $SD_{age} = 10.64$) self-identified as: 130 women, 74 men, 4 non-binary, 3 not reported; 1 American Indian/Alaskan Native, 81 Asian, 14 Black, 20 Latinx, 8 Middle Eastern or North African, 67 White, 13 more than one race, and 7 other/not reported.

Stimuli

Stimuli for the image generation phase were created following the same procedure as Study 1 with two exceptions. First, we changed the image projection method (i.e., encoder) used to project the neutral faces into the StyleGAN-2 latent space. We used a minimally modified version of the *FeatureStyle*encoder[54], which in turn uses a modified version of the *Pixel2Style2Pixel*encoder[55]. We opted to change the image encoder in an effort to create stimuli with greater detail after the inversion and averaging process.

Second, we created 1,000 images (instead of 300) by averaging 10 randomly sampled real (but projected) neutral faces. We applied the same amount of random Gaussian noise as in Study 1 to each averaged latent to further differentiate the face.

Images for the judgment validation phase were similarly generated following the procedure outlined in Study 1 Phase I. Each participant judged their own visual model's images plus that of five randomly selected other participants' images within the same condition.

Participant Procedure

The participant procedure for creating the individualized image models was nearly identical to Study 1 with the exception of the new target judgments (age, feminine-masculine, attractiveness, and familiarity) and an increased number of trials (from 300 to 1000). We increased the number of trials to assess the optimal number of trials required for generative reverse correlation in a secondary analysis (see Supplemental Materials).

After taking part in the image generation phase, participants were invited back in a separate asynchronous online session to judge images on the target category they were assigned to. Stimuli for this portion of the experiment consisted of images created using their own visual model (at model values +/-6, +/-4, +/-2, +/-1, 0) as well as a random sample of five other participants' images (at model values +/-6, +/-4, +/-2, +/-1, 0) for a total of 54 images. Thus, a participant who made age categorizations, which were used to create their visual model of "age," in Phase I, always saw faces generated by "age" models of other participants in Phase II. Participants judged all images in Phase II twice across two blocks. Images were randomized across participants and blocks. All other aspects of Phase II were identical to that of Study 1 Phase II (including scale, demographic, and debrief questions).

Study 3

Participants

One hundred and twenty-four participants took this study online through CloudResearch. Like Study 2, we only included in the analyses participants who completed the study in its entirety and did not restart. This left us with a final sample of 115 participants across each condition: feminine/masculine = 58; un/attractive = 57.

Participants (M_{age} = 42.06, SD_{age} = 11.27) self-identified as follows: 44 women, 68 men, 1 non-binary, 1 trans (transgender, trans man, trans woman); 1 American Indian/Alaskan Native, 5 Asian, 11 Black, 4 Latinx, 84 White, 9 more than one race, and 1 other/not reported. Out of the 115 usable participants in Phase I (image model construction), 96 returned for the rating component (Phase II). One participant completed the survey twice and their data was removed from analyses. Additionally, four participants had negative or near zero test-retest reliabilities ($r < .05$) in Phase I. However, none of these participants completed Phase II, so their data did not affect subsequent analyses. This left a final sample of 96 participants across each condition for Phase II: 45 for feminine/masculine and 51 for un/attractiveness.

Stimuli

In order to generate high-quality, neutral-appearing face stimuli, we fine-tuned the StyleGAN-2 FFHQ model with a set of 47,724 high quality neutral faces (over 75,000 training images with augmentation). We trained the new generative adversarial network (GAN) for an additional 4,000 epochs reaching a final Fréchet inception distance score of 4.19, which is comparable to the original StyleGAN-2 FFHQ model trained on 70,000 face images. Additional training details and examples of random images sampled from the new model's latent space are presented in the Supplemental Materials. We have made this model free for researchers (https://github.com/PsychInsight/model-zoo).

In order to select the stimuli that participants saw, we first randomly generated a large set of face images from the new model (> 1000). Next, we manually inspected the images and removed any faces that contained artifacts or were clearly warped. Based on a secondary analysis performed on Study 2's data, we concluded that 300 image trials were enough to obtain stable visualizations from participant's idiosyncratic visual model vectors. Thus, we selected the first 300 images out of the remaining pool to act as stimuli in the experiment.

Participant Procedure

The participant procedure was nearly identical to Study 2 with three minor differences. First, instead of 1000 trials, participants completed 300 trials. Second, we elected to only examine and validate two judgments: femininity/masculinity and attractiveness. Finally, during the validation phase of the study (i.e., Phase II), participants judged visual representations from their own and other's idiosyncratic visual models at model values of +/-1, +/-2, and 0. We reduced the model interpolation value range because images constructed at higher values quickly degraded in quality for some attributes and participants (i.e., the idiosyncratic latent space went out of sample; see Supplemental Material for details and Fig. 5A, second row for an example).

Study 4

Participants.

One hundred and forty-six participants (M_{age} = 40.22, SD_{age} = 10.97) completed the image generation phase of this experiment online through CloudResearch. Participants self-identified as follows: 58 women, 85 men, 1 non-binary, 1 trans (transgender, trans woman, trans man), 1 not reported; 3 Asian, 18 Black, 1 Latinx, 111 White, 12 more than one race, 1 not reported.

Two hundred and sixty participants completed the image judgment phase of this experiment online through CloudResearch. Two participants completed the experiment twice and an additional 11 restarted the experiment; these participants were removed from additional analysis. The remaining 247 participants (M_{age} = 43.88, SD_{age}

= 12.58) self-identified as follows: 123 women, 116 men, 3 non-binary, 2 trans (transgender, trans woman, or trans man), 3 not reported; 1 American Indian/Alaskan Native, 17 Asian, 18 Black, 9 Latinx, 1 Middle Eastern, 1 Native Hawaiian or Other Pacific Islander, 179 White, 1 other, 1 not reported, and 16 more than one race.

Stimuli

The stimuli for the image generation phase of this experiment were the same stimuli as in Study 3. The images generated during this phase were then used in the second phase as stimuli (see Fig. 6A for examples of each context-dependent trustworthy visualization). Twenty-six participants had negative or near zero test-retest reliability ($r < .05$) and one participant did not use all three response options at least once. This left us with a final sample of 119 participants across the three conditions: 45 "trust to fix your car"; 43 "trust to watch your child"; and 31 "trust to invest your money".

Stimuli in the image judgment phase were images generated from each valid individual participant's model from the image generation phase. For each individual model, we constructed an image at +/- 1 and +/- 2 for a total of 476 unique images to be judged (4 images x 119 participants). After removing poor quality participants in Phase II (test-retest $r < .05$; $n = 26$), each of the 476 images was judged on average by 27.86 participants.

Participant Procedure

The participant procedure for the image generation phase of this study was largely the same as Studies 1–3. The only difference was that before categorizing each image, participants were presented with a brief context that situated the subsequent trustworthiness judgments. Participants were randomly assigned to one of three conditions: "trust to fix your car", "trust to watch your child", and "trust to invest your money". As an example, participants who were randomly assigned to the "trust to fix your car" condition read the following during the instructions, "Imagine your car recently broke down and you need to find a reputable mechanic. You will be shown images of potential mechanics in your area. Your task is to provide ratings on how much you trust each individual to fix your car in an honest and satisfactory manner." The other context-dependent scenarios can be viewed online in this study's preregistration.

The participant procedure for the image judgment phase consisted of participants first being randomly assigned to judge the trustworthiness of each face within one of three context conditions, similar to that in the image generation phase. Next, participants serially judged a random subset of 60 of the 476 images on a scale of 1 = "not at all" to 7 = "very". Importantly, all images were judged on all trustworthiness conditions, not just those assigned to similar context-dependent conditions. For example, images generated from the "trust to fix your car" condition in the first phase were judged in the second phase on the trustworthiness within all three categories, not just "trust to fix your car". Finally, participants in this phase of the study judged each image twice across two separate blocks to assess their test-retest reliability. The order in which stimuli appeared across blocks and participants was randomized.

Author Contributions.

DNA collected the data, analyzed the results, and wrote the initial manuscript; AT obtained funding and provided supervision; DNA, SU, and AT contributed to manuscript conceptualization, writing, and interpretation of the results.

Data availability

All data and analysis code are available online (https://osf.io/aqgfw/). Each study's design and analyses were pre-registered (https://aspredicted.org/4SW_4TZ; https://aspredicted.org/QW3_FCB; https://aspredicted.org/22C_K6R; https://aspredicted.org/35X_N3Y).

Received: 26 August 2024; Accepted: 7 January 2025
Published online: 04 February 2025

References

1. Blake, W. *The Marriage of Heaven and Hell* (The Dunyazad Digital Library, 2020).
2. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl. Acad. Sci.* **105**, 11087–11092 (2008).
3. Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A. & Suchow, J. W. Deep models of superficial face judgments. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2115228119 (2022).
4. Park, J., Shimojo, E. & Shimojo, S. Roles of familiarity and novelty in visual preference judgments are segregated across object categories. *Proc. Natl. Acad. Sci.* **107**, 14552–14555 (2010).
5. Richler, J. J., Wilmer, J. B. & Gauthier, I. General object recognition is specific: evidence from novel and familiar objects. *Cognition* **166**, 42–55 (2017).
6. Coburn, A. et al. Psychological responses to natural patterns in architecture. *J. Environ. Psychol.* **62**, 133–145 (2019).
7. McManus, R. M., Young, L. & Sweetman, J. Psychology is a property of persons, not averages or distributions: confronting the Group-to-person generalizability problem in experimental psychology. *Adv. Methods Practices Psychol. Sci.* **6**, 25152459231186615 (2023).
8. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise: A Flaw in Human Judgment* (William Collins, 2021).
9. Albohn, D. N., Uddenberg, S. & Todorov, A. A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Front. Psychol.* **13**, (2022).
10. Albohn, D. N., Martinez, J. E. & Todorov, A. Determinants of shared and idiosyncratic contributions to judgments of faces. *J. Exp. Psychol. Hum. Percept. Perform.* **50**, 11, 1117–1130 (2024).
11. Hehman, E., Sutherland, C. A. M., Flake, J. K. & Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *J. Personal. Soc. Psychol.* **113**, 513–529 (2017).
12. Hönekopp, J. Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 199–209 (2006).

13. Martinez, J. E., Funk, F. & Todorov, A. Quantifying idiosyncratic and shared contributions to judgment. *Behav. Res.* https://doi.org/10.3758/s13428-019-01323-0 (2020).
14. Thornhill, R. & Gangestad, S. W. Facial attractiveness. *Trends Cogn. Sci.* **3**, 452–460 (1999).
15. Bjornsdottir, R. T., Hehman, E. & Human, L. J. Consensus enables Accurate Social judgments. *Social Psychol. Personality Sci.* **194855062110470** https://doi.org/10.1177/19485506211047095 (2021).
16. Kurosu, A. & Todorov, A. The shape of novel objects contributes to shared impressions. *J. Vis.* **17**, 14 (2017).
17. Leder, H., Goller, J., Rigotti, T. & Forster, M. Private and Shared taste in art and face appreciation. *Front. Hum. Neurosci.* **10**, (2016).
18. Specker, E. et al. Warm, lively, rough? Assessing agreement on aesthetic effects of artworks. *PLOS ONE*. **15**, e0232083 (2020).
19. Vessel, E. A., Maurer, N., Denker, A. H. & Starr, G. G. Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition* **179**, 121–131 (2018).
20. Isik, A. I. & Vessel, E. A. Continuous ratings of movie watching reveal idiosyncratic dynamics of aesthetic enjoyment. *PLOS ONE*. **14**, e0223896 (2019).
21. Albright, L., Kenny, D. A. & MaUoy, T. E. Consensus in personality judgments at zero acquaintance. *Interpers. Relations Group. Processes*. **55**, 387–395 (1988).
22. Kenny, D. A. & La Voie, L. The Social Relations Model. in *Advances in Experimental Social Psychology* vol. 18 141–182Elsevier, (1984).
23. Lavan, N., Mileva, M., Burton, A. M., Young, A. W. & McGettigan, C. Trait evaluations of faces and voices: comparing within- and between-person variability. *J. Exp. Psychol. Gen.* **150**, 1854–1869 (2021).
24. Lavan, N. & Sutherland, C. A. M. Idiosyncratic and shared contributions shape impressions from voices and faces. *Cognition* **251**, 105881 (2024).
25. Jirschitzka, J., Oeberst, A., Göllner, R. & Cress, U. Inter-rater reliability and validity of peer reviews in an interdisciplinary field. *Scientometrics* **113**, 1059–1092 (2017).
26. Brinkman, L., Todorov, A. & Dotsch, R. Visualising mental representations: a primer on noise-based reverse correlation in social psychology. *Eur. Rev. Social Psychol.* **28**, 333–361 (2017).
27. Murray, R. F. Classification images: a review. *J. Vis.* **11**, 2–2 (2011).
28. Todorov, A., Dotsch, R., Wigboldus, D. H. J. & Said, C. P. Data-driven methods for modeling Social Perception: modeling Social Perception. *Soc. Pers. Psychol. Compass*. **5**, 775–791 (2011).
29. Abel, L. A. & Quick, R. F. Wiener analysis of grating contrast judgments. *Vision. Res.* **18**, 1031–1039 (1978).
30. Ahumada, A. J. Classification image weights and internal noise level estimation. *J. Vis.* **2**, 8 (2002).
31. Albohn, D. N., Brandenburg, J. C. & Adams, R. B. Perceiving emotion in the Neutral Face: a powerful mechanism of Person Perception. in The Social Nature of Emotion Expression (eds Hess, U. & Hareli, S.) 25–47 (Springer International Publishing, Cham, doi:https://doi.org/10.1007/978-3-030-32968-6_3. (2019).
32. Albohn, D. N. & Adams, R. B. Everyday beliefs about emotion perceptually derived from Neutral Facial Appearance. *Front. Psychol.* **11**, 264 (2020).
33. Beard, B. L. & Ahumada, A. J. A Technique to Extract Relevant Image Features for Visual Tasks. in *Proceedings of SPIE Human Vision and Electronic Imaging III* vol. 3299 79–85 (1998).
34. Dotsch, R., Wigboldus, D. H. J., Langner, O. & van Knippenberg, A. Ethnic out-Group faces are biased in the Prejudiced mind. *Psychol. Sci.* **19**, 978–980 (2008).
35. Dotsch, R. & Todorov, A. Reverse correlating Social Face Perception. *Social Psychol. Personality Sci.* **3**, 562–571 (2012).
36. Gosselin, F. & Schyns, P. G. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision. Res.* **41**, 2261–2271 (2001).
37. Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R. & Schyns, P. G. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 7241–7244 (2012).
38. Mangini, M. C. & Biederman, I. Making the ineffable explicit: estimating the information employed for face classifications. *Cogn. Sci.* **28**, 209–226 (2004).
39. Martin-Malivel, J., Mangini, M. C., Fagot, J. & Biederman, I. Do humans and baboons use the same information when Categorizing Human and Baboon faces? *Psychol. Sci.* **17**, 599–607 (2006).
40. Moon, K., Kim, S., Kim, J., Kim, H. & Ko, Y. The Mirror of mind: visualizing Mental representations of Self through Reverse correlation. *Front. Psychol.* **11**, 1149 (2020).
41. Yu, H., Garrod, O. G. B. & Schyns, P. G. Perception-driven facial expression synthesis. *Computers Graphics*. **36**, 152–162 (2012).
42. Todorov, A. & Oh, D. The structure and perceptual basis of social judgments from faces. in *Advances in Experimental Social Psychology* S0065260120300290Elsevier, (2021). https://doi.org/10.1016/bs.aesp.2020.11.004
43. Dzhelyova, M., Perrett, D. I. & Jentzsch, I. Temporal dynamics of trustworthiness perception. *Brain Res.* **1435**, 81–90 (2012).
44. Galinsky, D. F. et al. Do I trust you when you smile? Effects of sex and emotional expression on facial trustworthiness appraisal. *PLoS ONE*. **15**, e0243230 (2020).
45. Mehu, M., Little, A. C. & Dunbar, R. I. M. Sex differences in the effect of smiling on social judgments: an evolutionary approach. *J. Social Evolutionary Cult. Psychol.* **2**, 103–121 (2008).
46. Oliveira, M., Garcia-Marques, T., Dotsch, R. & Garcia-Marques, L. Dominance and competence face to face: dissociations obtained with a reverse correlation approach. *Euro. J. Social Psych*. **49**, 888–902 (2019).
47. Scarfe, P. & Hibbard, P. B. Reverse correlation reveals how observers sample visual information when estimating three-dimensional shape. *Vision. Res.* **86**, 115–127 (2013).
48. Cone, J., Brown-Iannuzzi, J. L., Lei, R. & Dotsch, R. Type I error is inflated in the two-phase reverse correlation Procedure. *Social Psychol. Personality Sci.* **194855062093861** https://doi.org/10.1177/1948550620938616 (2020).
49. Martinez, J. E. & Paluck, E. L. *Quantifying Shared and Idiosyncratic Judgments of Racism in Social Discourse*. https://osf.io/kfpjg (2020). https://doi.org/10.31234/osf.io/kfpjg
50. Martinez, J. E., Oh, D. & Todorov, A. *Mental Representations of Immigrants Encode Racialized Expectations of Socio-Structural Positions*. https://osf.io/cvhze (2021). https://doi.org/10.31234/osf.io/cvhze
51. Hauser, D. J. et al. Evaluating CloudResearch's approved Group as a solution for problematic data quality on MTurk. *Behav. Res.* https://doi.org/10.3758/s13428-022-01999-x (2022).
52. Karras, T., Laine, S. & Aila, T. A. Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]* (2018).
53. Karras, T. et al. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]* (2020).
54. Yao, X., Newson, A., Gousseau, Y. & Hellier, P. Feature-Style Encoder for Style-Based GAN Inversion. (2022). https://doi.org/10.48550/ARXIV.2202.02183
55. Richardson, E. et al. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. Preprint at (2021). http://arxiv.org/abs/2008.00951

Author contributions

DNA collected the data, analyzed the results, and wrote the initial manuscript; AT obtained funding and provided supervision; all authors contributed to conceptualization, writing, and interpretation of results.

Declarations

Competing interests
All authors are named inventors on a pending patent related to this work.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86056-1.

Correspondence and requests for materials should be addressed to D.N.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.