Theory of optimal experimental manipulation with generative models via reverse correlation of stimuli

Adam Sobieszek[1], Maciej Siemiątkowski[1]

Abstract

The article bridges experimental methodology and Artificial Intelligence (AI) generative models that construct realistic stimuli. The complex representations learned by neural networks can create stimuli with arbitrary features, which could enable researchers to vary a specific evoked psychological impression within an experimental manipulation. However, this complexity hinders the ability of existing reverse correlation approaches to accomplish this task. We develop the necessary mathematical foundations to address these limitations. We derive how linear model, morphing, and choice-based approaches systematically deviate from the optimal manipulation directions—the local gradient of the expected impression in feature space—and establish its psychometric properties as itself a valid target for empirical estimation. We provide state-of-the-art prediction models for 28 facial impressions for validation. We show that to construct robust manipulations, the problem has to be expressed in terms of feature distributions and propose a solution using the mathematics of flows and optimal transport, leading to our Psychometric Gradient Flow Matching algorithm. In four studies ($N_{total} = 1026$), we validate performance across methods and impressions of faces. Study 1 demonstrates the effectiveness of manipulating visual features (race typicality, happiness, age) with the gradient. Study 2 shows how this approach succeeds for impressions like dominance but fails for valence-laden traits like trustworthiness. Studies 3 and 4 show the superior performance of our flow-based algorithm across three valence-laden impressions (trustworthy, smart, attractive) and can uniquely work under controlled variables. The presented methods enable researchers to construct more complex orthogonal designs to disentangle the effects of highly interrelated psychological processes.

Keywords: generative models, experimental manipulation, reverse correlation, machine learning

[1]

[1]Faculty of Psychology, University of Warsaw, Poland

1 Introduction

One of the preeminent goals of experimental methodology is to enable controlled experimental interventions that can disentangle the effects of different causal factors. Progress in experimental manipulation methods can thus consist of finding new causal factors to intervene on, reducing general side-effects of the intervention, or finding a way to explicitly disentangle interventions on two known causal factors (enabling orthogonal designs). In psychology, where direct intervention on psychological phenomena in vivo is infeasible, the experimenter is particularly limited in experimental manipulation tools to create experimental conditions with valid, specific and orthogonal manipulations. The use of stimuli, assigned to different experimental conditions—facilitated by large, curated databases—has made significant steps towards rectifying this limitation.

Stimuli norm databases provide a key resource. Curated stimuli and the reported levels of emotional and socio-cognitive impressions they evoke have been collected e.g. for different emotions in word stimuli (Bradley & Lang, 1999), different first impressions, social categorizations or perceived emotional expressions in face stimuli (Ma et al., 2021; Peterson et al., 2022), emotions evoked by image stimuli (Marchewka et al., 2014), emotional prosody in voice stimuli (Lassalle et al., 2019) or affective dynamics in video stimuli (Samide et al., 2020). Additionally, whenever the evoked impression has a stable correspondence to quantitative stimuli features, databases enable modeling and extrapolating these subjective norms to new unrated word (Plisiecki & Sobieszek, 2024), voice (Wagner et al., 2023), or face stimuli (Peterson et al., 2022).

A key observation is that these developments do not solely concern the study of stimuli themselves, instead facilitating investigation in topics of interest across cognitive, social, emotion and clinical psychology. From important social phenomena such (uncovering impacts of race, gender, age and specific face impressions on perceived emotional state, Summerell et al., 2025, hiring decisions (Derous et al., 2017), or political election outcomes, Todorov et al., 2005), to observing subtle differences between clinical and non-clinical populations (e.g. reaction differences in tasks using emotional word stimuli, Zhukova et al., 2021), which could lead to new diagnostic tools (e.g. suicide risk screening Chung and Jeglic, 2017).

Orthogonal manipulations, however, are often hard to achieve with stimuli selection if the desired controlled and manipulated variables naturally covary. A stronger internal validity can be achieved with methods to selectively vary the level of a cue within the same control condition stimuli. Computer-based methods make this feasible: digitally manipulating a stimulus feature such as image hue or vocal pitch to obtain altered stimuli versions, matched between experimental conditions, affords markedly stronger causal inferences than selecting different items for each condition.

Hence, experimental methods benefits from developments in data-driven digital representations that allow one to independently control more meaningful, complex features. An early example of such an effort was the FaceGen approach of Blanz and Vetter (1999), where PCA is used to construct a *feature space representation* of face shapes that captures the major factors of variation of real laser-scanned 3D faces.

Noise-based reverse correlation has developed in parallel in the field of psychophysics (Eckstein & Ahumada, 2002), which conversely relies on hundreds of *unstructured* noise stimuli that observers classify, to construct an average *Classification Image* (CI), which reveals which stimulus-domain features were associated with their responses. Early success of the approach for audio (Ahumada & Lovell, 1971) and visual (Ringach et al., 1997) stimuli prompted its widespread adoption as a way to 'visualize internal mental representations' (Brinkman et al., 2017).

A significant step was made by Oosterhof and Todorov (2008), who applied a reverse correlation-inspired technique to the feature space modeled by FaceGen to identify *feature space directions* associated with different face impressions. The reduction afforded by a feature space allowed Oosterhof and Todorov show that data-driven approaches can not only discover features to *visualize* but also to *induce* a studied impression by using the discovered feature space direction to manipulate newly sampled faces.

Although a promising experimental manipulation tool, 3D models have been criticized in terms of their generalizability, due to the lack of ecological validity and smaller effect sizes they produce (Crookes et al., 2015). A different unique strand of manipulation methods comes from morphing approaches (Sutherland et al., 2017). Morphing techniques can uniquely maintain both a continuous change in impressions *and* ecological validity, but are limited in the type of stimuli they can be applied to.

The rapid progress in deep learning has significantly improved the realism of computer-generated stimuli. Models like Diffusion (Rombach et al., 2022) and Generative Adversarial Networks (GANs; Goodfellow et al. 2014) have enabled generation of images with nearly arbitrary combination of features. However, the complexity of their feature representations can limit applicability of existing reverse correlation methods to these models. Recently Peterson et al. (2022) introduced the One Million Impressions (OMI) dataset of norms of over 30 psycholog-

ical impressions for faces generated by the highly realistic StyleGAN2 model (Karras et al., 2020). Following the model-based approach of Oosterhof and Todorov (2008), they demonstrated that manipulating the GAN representations of faces along the direction given by linear model coefficients resulted in changes in perceived impressions.

More recently, Sobieszek et al. (2024) provided unexpected evidence that new methodological grounding is needed to truly take advantage of generative models in experimental manipulation. Within the setting of Peterson et al. (2022) they identify how maximizing impression model accuracy does not actually correspond to maximizing manipulation quality (shrinkage, for example, causing regression accuracy to increase and manipulation quality to decrease). Using much fewer impressions, they were able to find manipulation directions that, while being not the best linear models, resulted in much larger changes to the target impression at a smaller overall change in appearance, demonstrating a gap in manipulation-specific theoretical foundations. Given how similar generative models have already shown promise in controlling psychologically relevant cues in video (Yan et al., 2023), and voice stimuli (Cho et al., 2024), filling this research gap could unlock major methodological advancement in experimental psychology.

1.1 Aims and contributions

We pursued three primary aims, corresponding to theoretical, methodological, and empirical contributions:

1. **Theoretical.** Formalize Optimal Experimental Manipulation of Stimuli (OEMS):
   (a) Defining formal optimality criteria for manipulation directions 2.2 , paths 2.8 and a new distributional constraint necessary for orthogonal manipulation 2.7; as well as show how mathematics of optimal transport and flows enable their satisfaction 2.5.
   (b) Identify the class of generative models that make manipulation direction identifiable A.1, demonstrate that the gradient of the mean impression can be established within such spaces B.1 is locally optimal for manipulation 2.3 and psychometrically estimable B.2.
   (c) Derive (as function of this optimal direction) the estimands of choice- and noise-based reverse correlation C.2, linear model C.1, and morphing A.2 approaches to manipulation.
2. **Methodological.** Develop tools for experimental manipulation of stimuli that satisfy the theoretical requirements:
   (a) First open source, state-of-the-art, neural network prediction models for 28 face impressions 3.2 of the OMI dataset.
   (b) The *Psychometric Gradient Flow Matching* (PGFM) algorithm, which learns optimal manipulation paths that satisfy controlled variable constraints 3.1.
   (c) The *Model Classification Image* as a way to visualize stereotypes encoded in impression models for arbitrary categories e.g. of trustworthiness for faces categorized as old, white and male D.2.
   (d) Releasing trained models, stimuli, and code to facilitate future research 3.6.
3. **Empirical.** In four studies $N_{\text{total}} = 1026$ (4):
   (a) We test the effectiveness of linear, gradient, and PGFM approaches in their ability to effectively influence between-subject impressions, control confounding factors and preserve ecological validity
   (b) Test whether the PGFM approach can reliably increase valence-laden impressions (trustworthiness, attractiveness) challenging for prior methods, while constrained to control apparent age and gender.

2 Theoretical Framework

The problem of optimal experimental manipulation can be approached from two perspectives:

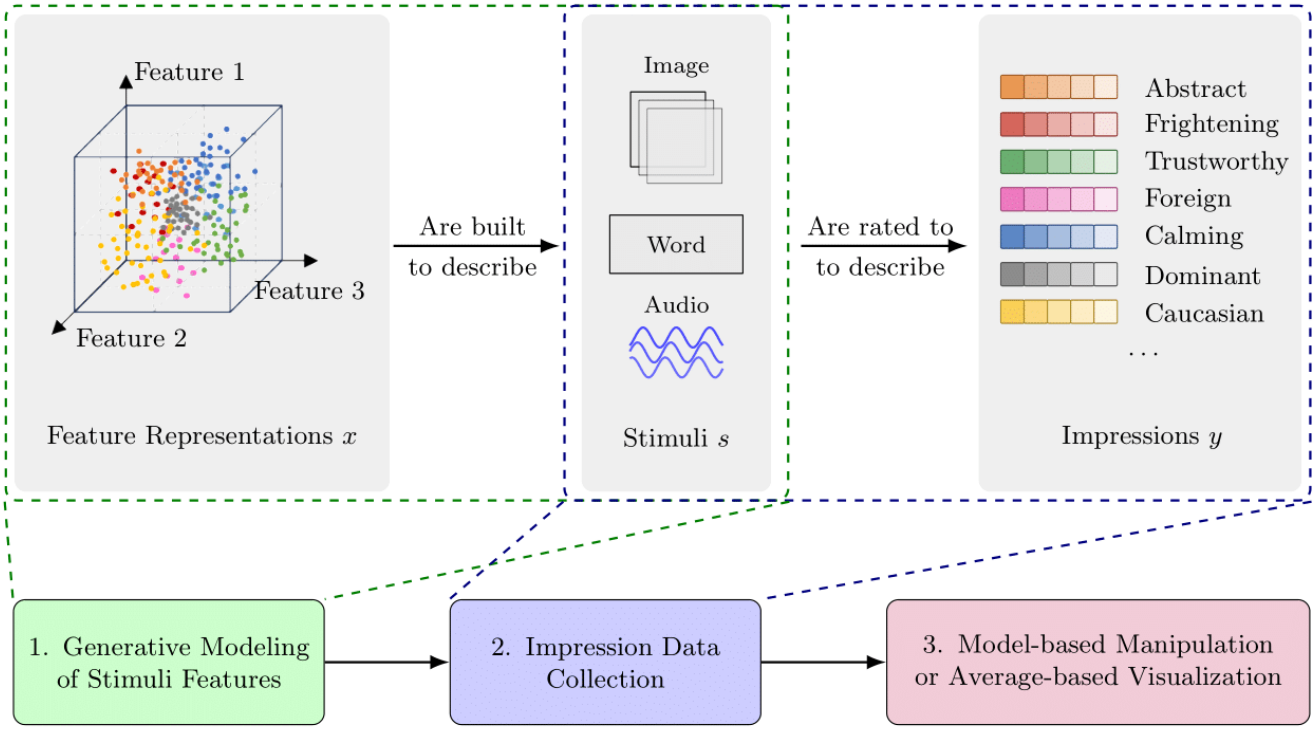
- **The *High Road*** is the top-down formalization, that seeks a general solution for optimally changing expected impressions of stimuli (represented as vector spaces) within a generative modeling framework.
- **The *Low Road*** is the bottom-up analysis of existing methods in psychophysics and face impressions, highlighting the open problems and limitations encountered along the way that we expect to tackle within a more principled approach.

The primary formal contribution of this work will be to bridge these two roads, showing how disparate existing methods can be understood in relation to a geometric notion of a *psychometric gradient* (the direction in features of the fastest change in mean impressions) and a probabilistic notion of the *distribution of stimuli* (capturing their natural feature covariation). Figure 1 illustrates the general steps involved in reverse correlation, both as a data-driven method to visualize and to manipulate stimuli.

2.1 The High Road

The limiting factor in the use of computer-based manipulation methods is that the feature spaces like those used

Figure 1: The reverse correlation steps involved in impression visualization and data-driven stimuli manipulation.



Note. Reverse correlation requires a quantitative description of both the stimuli and their associated impressions. (1) To describe the stimuli, generative modeling is used, where a feature space representation (of e.g., faces or speech) is constructed to capture the perceptual differences between them. (2) To describe the impressions they evoke, individuals are asked to report their subjective impressions (e.g., trustworthiness, abstractness, or dominance); (3) The data can be used to visualize the impression by constructing a prototypical stimuli or an impression model may be trained for use in experimental manipulation.

by Oosterhof and Todorov (2008) to find face impression directions had to be constructed by explicitly providing the factors of variation in the stimuli. Conversely, while all digital stimuli can be *described* by their raw pixel or amplitude values, this high-dimensional representation $\mathbb{R}^{d_{\text{pix}}}$ (for images) is poorly suited for manipulation. As the valid stimuli of a given type lay on a contoured, narrow surface within the overall space, geometric information like distances and directions correspond poorly to relevant perceptual factors of variation. This limiting factor can be overcome with the rising capacity of generative models.

Unsupervised learning in AI refer to methods that enable *representation learning* from unlabeled examples. With suitable training objectives neural network learned representations—like word embeddings—can encode high-level features, such that nearby points in this feature space encode meaningfully similar stimuli. In particular interest to us, such models can encode learned features as directions (see Mikolov et al., 2013 for words, and Karras et al., 2020 for images), which even for word representation (in a result reminiscent of Oosterhof and Todorov, 2008) can result in distinct directions capturing distinct emotional dimensions (Plisiecki & Sobieszek, 2024). How neural network achieve this can be understood from two complimentary views of the goal of generative modeling, as capturing the distribution (natural covariation of features, likelihood of examples), and the geometry (distances/differences between samples, their district feature directions) of a target domain.

The probabilistic view of the goal is to model the natural variation or *distribution* of valid stimuli, denoted $p_{\text{valid}}(s)$, from a large set of examples *sampled* from this class of stimuli. The probabilistic model p_G is trained not in order to predict uncertainty, but to allow sampling of new, unseen examples. To this end, the output domain is conditioned on a simpler *latent* space z, for which we know the sampling distribution $z \sim p(z)$. We aim for

$$\underbrace{p_{\text{valid}}(s)}_{\substack{\text{natural}\\\text{variation}}} \approx \int_{\mathcal{Z}} \underbrace{p_G(s \mid z)}_{\substack{\text{stimuli}\\\text{generation}}} \underbrace{p(z)}_{\text{sampling}}$$

The geometric view of the goal is to learn a deterministic mapping G from a lower-dimensional vector space to the space of stimuli, mapping the valid vector representations $\mathcal{X} \subset \mathbb{R}^d$ to the surface of valid stimuli $\mathcal{S}$:

$$\mathcal{X} \overset{G}{\mapsto} \mathcal{S} \subset \mathbb{R}^{d_{\text{pix}}}$$

In this formulation, the goal is *representation learning*, with the feature space x serving as a learned coordinate system for valid positions on the stimuli surface. The point $\mathbf{x} \in \mathcal{X}$ is then a vector of learned features describing the stimulus $\mathbf{s} = G(\mathbf{x})$ it maps to.

The geometric goal of *simple representations* and probabilistic goal of fully *capturing the natural variation* can clash. Fully probabilistic models like diffusion (Rombach et al., 2022) need a simple prior distribution (e.g. Gaussian Noise), while a model that represents distinct features of stimuli will require a complicated distribution $p_{\text{feature}}(x)$ that matches the covariation of these features in reality. The Generative Adversarial Network (GAN) class of models provide a powerful computational solution to this problem. The StyleGAN2 model (Karras et al., 2020) achieves both objectives by mapping the input sampling space z to an intermediate **feature space** x which is conditioned to learn distinct linear features of the output stimuli ($z \to x \to s$). Such a feature space, we call *well-conditioned*, satisfies formal requirements for capturing meaningful, independent feature directions. Appendix A establishes those axioms and shows a key result that lead to well-conditioned feature spaces containing identifiable feature directions: they make distances between points x_1 and x_2 correspond to perceptual differences between their associated stimuli and encourage straight paths in features $x_1 \to x_2$ to map to ideal morphs between them $s_1 \xrightarrow{\text{morph}} s_2$.

Definition 2.1 (Problem Setting)**.** Assume an impression y covaries with the factors of variation in a class of stimuli, such that the expected impression can be modeled as a function of the stimuli features. Let:

- $X \subset \mathbb{R}^d$ be a continuous feature representation of stimuli s, describing their perceptual appearance.
- $f_y : X \to \mathbb{R}$ be a function modeling the expected impression y evoked by each stimuli based on their perceptual features x.
- $p_{\text{feature}}(x)$ be the probability of co-occurrence of different stimuli features describing the distribution of valid stimuli.

The objective is to identify for each stimuli the best manipulation direction. The solution:

- $\vec{v}^*(x) : \mathbb{R}^d \to \mathbb{R}^d$ assigns to each stimuli what features to change to induce the fastest change in expected y.

We start by framing the problem locally, in terms of a manipulation direction for a given stimuli s_1 whose feature representation x_1 falls within the support of $p_{\text{feature}}(x)$ ($x \in \text{supp}(P_X) \implies s$ stimuli with this combination of features aren't so unlikely as to lack ecological validity and destroy the accuracy of f_y):

Definition 2.2 (Locally Optimal Manipulation Direction)**.** Restricting ourselves to a local neighborhood around a stimuli s_1, the objective is to find a direction in feature space $\vec{v}$, that when manipulating s_1 along this direction to obtain s_2:

1. Validity: Produces the largest change in the target psychological trait:

$$\max_{\vec{v}} \Delta y = y(s_2) - y(s_1)$$

2. Specificity: Minimizes perceptual change of the stimuli, which corresponds to minimizing the distance in a well-conditioned feature space (see Appendix A):

$$\min_{\vec{v}} \text{diff}(s_1, s_2) \approx \text{dist}(x_1, x_2)$$

3. Explicit control: Avoids influencing important confounding variables, which for K features $\{\omega_k\}$ that should be held constant (e.g., age, gender) can be expressed as a constraint:

$$\vec{v} : \quad |\omega_k(s_1) - \omega_k(s_2)| \leq \epsilon_k \quad \forall_{k \in \{1,\ldots,K\}}$$

where ϵ_k is a small tolerance for each controlled factor.

4. Ecological Validity: Maintains ecological validity of the associated stimuli.

The first two requirements present a tradeoff between the amount of manipulation obtained Δy and amount of change in appearance $d(s_1, s_2)$ put in. To assess which direction optimizes them jointly we can define **manipulation efficiency** as the impression change over stimulus difference $\left(\frac{\text{validity}}{\text{specicity}}\right)$, i.e. the impression change per unit distance. For local manipulation, the task becomes well-defined when we restrict ourselves to directions $\mathbf{v}$ in feature space among unit-length vectors $||\mathbf{v}|| = 1$. Finding the direction which maximizes the combined objective:

$$\max_{\mathbf{v}} \frac{\Delta y}{d(x_1,\, x_1 + \epsilon\mathbf{v})} = \frac{f_y(x_1 + \epsilon\mathbf{v}) - f_y(x_1)}{||\epsilon\mathbf{v}||}$$

Which under small perturbations ϵ is the directional derivative of f_y in the direction $\mathbf{v}$:

$$\lim_{\epsilon \to 0} \frac{f_y(x_1 + \epsilon\mathbf{v}) - f_y(x_1)}{\epsilon} = D_{\mathbf{v}} f_y$$

This formulation directly demonstrates the optimal local direction for manipulation is the direction maximizing the directional derivative — the gradient.

Proposition 2.3 (Local Optimality Criterion)**.** *For any stimuli s whose feature representation x falls within the support of $p_{feature}$, the direction of change $\vec{v}$ that locally maximizes the increase in impression y while minimizing the perceptual change is the direction of the gradient of the the impression model f_y with respect to x, i.e., $\nabla f_y(x)$.*

$$\boxed{x \in \text{supp}(p_{feature}) \implies \vec{v}^*(x) \doteq \nabla f_y(x)}$$

Where $\doteq$ indicates vectors are equal up to a constant and thus point in the same direction.

This special role of the local gradient direction comes from its coincidence, within this setting, of this geometric notion with a causal notion of the differential effects of features on ratings. Around a point x_0, in a local enough area within the support, the relative weight $\nabla f_y(x_0)$ assigns to each dimension x^i of x space, can be understood variously as the:

- **Structural coefficient** β in structural equations, where $\boxed{Y := \beta^\top (X - x_0) + y_0 + \varepsilon}$.
- **Marginal causal effect** in econometrics and causal inference, where $\boxed{\frac{\partial}{\partial x^i}\mathbb{E}[Y \mid do(X{=}x)\ ]}$.
- **Maximal dose–response** of a continuous "treatment" in a potential outcomes/heterogeneous treatment effects framework.

Generative adversarial networks in particular can uniquely facilitate the estimation of these quantities (Athey et al., 2024), which—as we derive in Appendix B—makes the gradient ∇y in such a feature space a (surprisingly measurement invariant) estimand, that we posit is the target of reverse correlation methods in psychology. We will indeed find, that manipulation and visualization methods—like the manipulation directions found with linear models (Oosterhof & Todorov, 2008; Peterson et al., 2022)—can be understood in relation to this target and its interaction with the natural distribution of stimuli features, which will be formalized in the following sections.

2.2 The Low Road

Reverse correlation methods initially focused on recovering from purely noise stimuli the pattern that visualizes the central tendency of the source of activation e.g. of a neuron (Ahumada and Lovell, 1971; Ringach et al., 1997). The pure-noise setting, while appealing, required many trials and necessarily limited the approach to low-resolution stimuli. In a modern Two-Alternative Forced-Choice (2AFC) task, to increase the statistical power and study more complex subjective impressions, a base stimulus $\mathbf{s}_0$ (e.g. an average face) is used and superimposed with independent noise patterns (Dotsch et al., 2008). Participants are asked to make judgments (e.g., "Which face is more trustworthy?"). Taking the pixel-mean of the stimuli that were chosen $\bar{\mathbf{s}}_{\text{high}}$ and subtracting the rejected $\bar{\mathbf{s}}_{\text{low}}$ removes the noise patterns unrelated to the judgment and uncovers a visual proxy for the observer's internal stereotype for that trait (Brinkman et al., 2017). However, as noted, the pixel space isn't suitable for finding directions within valid stimuli, that could enable the pixel-space average CI to be used for manipulation.

This possibility only appeared with the work of Oosterhof and Todorov (2008), who used the FaceGen feature space rather than the pixel space. By working in a parametric face model's feature space, their approach allowed not only for visualization but active manipulation. Their key observation was that in such well-behaved feature spaces directions can encode particular stimuli transformations which could in turn systematically alter impressions. The approach they propose is to identify this manipulation direction as the ordinary least squares coefficients $\hat{\beta}_{OLS}$ of linear regression. As shown in detail in Appendix C.1, this aligns with the local optimality criterion (maximizes manipulation efficiency) under the assumption, that there is no interaction between the feature values x and the effect of features on impressions (neither in size nor direction), as in such a case the structural equation we have written for the local effect could be generalized to the whole feature space using the standardized regression coefficients $\hat{\beta}_{OLS} \approx \nabla y(x)$.

Keeping this assumption for a moment, the linear model literature also provides an established solution to the third criterion in our definition—explicit control. Stating it in terms of gradients, to find the direction for y controlling for c we subtract the components of ∇y that are collinear with ∇c:

Definition 2.4 (Control by orthogonalization)**.** Assume ∇ and ∇c vectors are normalized, and $\langle \cdot | \cdot \rangle$ indicates the dot product which for normalized vectors describes their co-linearity. The manipulation direction for y controlling for c is:

$$\vec{v} = \nabla y - \langle \nabla y \mid \nabla c \rangle \cdot \nabla c$$

Morphing literature provides an interesting alternate solution, that will help explain whether the original reverse correlation estimates a similar feature direction as the model-based approach of Oosterhof and Todorov (2008). In their seminal work Sutherland et al. (2013) observe that for impressions to be even approximately linear in features, such a space has to discard a significant amount of information. Without requiring a feature space, their solution instead starts similarly to the CI methods—by determining which stimuli were rated low or high among a set of *valid* stimuli. Using morphing instead of pixel-averaging, a prototypical low and high rated faces can be created. The *method-of-constant-stimuli* (see

Sutherland et al., 2017) builds these prototypes to average out the features unrelated to the target impression and construct a sequence of valid stimuli reliably increasing in the target impression. By morphing $\bar{\mathbf{s}}_{\text{low}} \xrightarrow{\text{morph}} \bar{\mathbf{s}}_{\text{high}}$ the obtained sequence can be used to study the impression in a design in which the independent variable is the morph proportion $t \in [0, 1]$.

In appendix A we show that this morphing approach actually corresponds to the original reverse correlation formulation as choice-based averaging, but conducted within a well-conditioned feature space (c.f. Albohn et al., 2022), which preserves ecological validity and enables the comparison of directions found with choice-based averaging and model-based regression. In Appendix C, we show that the CI can be described as an estimate of the covariance-aligned gradient

$$\boxed{CI \approx \Sigma_{\mathbf{x}} \cdot \nabla y}$$

where $\Sigma_{\mathbf{x}}$ is the feature covariance matrix. As a result, the original reverse correlation direction systematically deviates from the optimal manipulation direction (the gradient of the impression), being influenced by $\Sigma_{\mathbf{x}}$ to align more with the major factors of variation of the feature distribution.

Finally, we come back to the simplifying assumption of linear models—that the same feature direction captures the impression across the whole space. This assumption is surely not correct for the actual impression being modeled, as complex impression differ in how they are influenced by features both in the direction (features that make male- and female-presenting faces attractive differ) and magnitude (the impact on trustworthiness of increasing the smile from a completely neutral position has a larger effect than increasing it more and more, which gives diminishing returns), and even simpler judgments are non-linear due to ceiling effects, imperfect feature representations or perceptual biases. Thus the directions estimated with linear models are also systematically biased by the natural feature covariance, which we show in Appendix C.1 makes the estimand of the linear model direction a spatial average of gradient directions under the sampling feature distribution p_{feature}. To model these interactions—changing the gradient direction—a non-linear, differentiable model is needed. The most prominent and successful method in this class of models is the neural network.

Stimuli (gradient) Descent is a recent computational extension of finding manipulation directions with neural network predictions that emerged in the emotional word methods literature (Plisiecki & Sobieszek, 2024). Because neural network models are differentiable, we can calculate the gradient $\nabla f_y(x)$ of a trained model for any input stimuli representation to find a manipulation direction. Two key differences that emerge when using the gradient of a neural network prediction, instead of a averaged global direction is that (a) the predicted direction changes as we move along the direction, making the path traced with this approach a curve, instead of a straight line, (b) we additionally have to consider the distribution of the stimuli in feature space to make sure we don't move outside of the region where model predictions remain valid. In the general setting, the second requirement also requires the path to end on a valid stimuli, a problem not tackled in its generality by any method, as Stimuli Descent 'snaps to' a nearest valid word representation and morphing manipulates a face explicitly in the direction of another valid stimuli. A unified presentation of this and other properties and estimands of the discussed methods are presented in Table 1.

2.3 The Low Road: Open problems

Impression model accuracy (understood as the correlation of the prediction $f_y(x)$ with human judgments) is necessarily limited in linear models—first by how linearly the feature space represents stimuli features and secondly by the existence of interactions between the type of stimuli and which features impact the impression most. While the first limitation diminishes rapidly with newer generative models, the second limitation is not inherent and requires the use of nonlinear impression models. In the present study we take the impression dataset collected by Peterson et al. (2022) and show how the use of neural network regression models can drastically increase the prediction accuracy. The accuracy of impression models is however also affected by what Plisiecki and Sobieszek (2024) call the *missing not at random problem* (MNAR), where impressions whose values are correlated with the likelihood of stimuli being picked for norm databases experience a systematic loss in accuracy for out-of-distribution stimuli (c.f. Snefjella and Blank, 2020).

Definition 2.5 (MNAR problem)**.** If an impression y is not independent of the likelihood of being included in the norm database, the apparent accuracy of f_y does not extrapolate to all stimuli:

$$\nabla y(x) \not\perp p(x) \mid x \sim p(x) \implies \begin{array}{c} f_y(x) \text{ fails} \\ \text{out-of-distribution} \end{array}$$

The above shorthand is a statement about conditional-independence between random variables ("A is not conditionally independent of B given C=c"), and more formally it should be read as "conditional on x (in the sampling distribution) the estimator (fisher information) of the true direction ∇y in x-space is not conditionally-independent of (the slope direction of) the likelihood of being included in the training set".

Although the experimentally observed loss of accuracy of neural network impression models for out-of-distribution words was negligible, in our continuous setting, MNAR

Table 1: A Unified Framework of Reverse Correlation and Manipulation Methods

Paradigm	Input Data	Method	Estimand (CI/Gradient)	Estimator	Respects $p_{\text{valid}}(s)$?	Manipulation Path Type	Relevant works
Impression Visualization	Noise $\mathbf{v}_i$ Choice c_i	Linear Models (STA/GLM)	$\mathbb{E}_s[\nabla p_{\mathbf{c}}]$	$\hat{\beta}_{GLM}$ for $p_{\mathbf{c}}(+1\|\cdot)$	No	(Yields feature importance)	Ahumada and Lovell (1971) and Chichilnisky (2001)
		Choice-based Average	$CI \approx \Sigma_{\mathbf{v}} \cdot \nabla y$	$\sum_i \frac{1}{N}[c_i \mathbf{v}_i]$	Approx.	(Yields an image)	Abbey and Eckstein (2002) and Dotsch et al. (2008)
+ **Manipulation**	Stimuli $\{\mathbf{s}\}_{\text{low}}$, $\{\mathbf{s}\}_{\text{high}}$	Morphing	$CI_{\text{low}} \rightarrow CI_{\text{high}}$	$\bar{\mathbf{s}}_{\text{low}} \xrightarrow{\text{morph}} \bar{\mathbf{s}}_{\text{high}}$	Yes	S-space Interpolation	Sutherland et al. (2017, 2013)
	N Stimuli, Model f_y	**Model CI**	CI implied by f_y	$\sum_i \frac{1}{N} \Sigma_{\mathbf{x}} \cdot \nabla f_y(x_i)$	Yes	X-space Interpolation	**This Work**
Manipulation Direction Estimation	Features x, Ratings y	Linear	$\mathbb{E}_x[\nabla y]$	$\hat{\beta}_{OLS}$	No	Linear	Oosterhof and Todorov (2008)
		LASSO / Ridge	$\mathbb{E}_x[\nabla y]$ Shrunk towards CI_{high}	$\hat{\beta}_{OLS}$; $\|\|\beta\|\|_k$ penalty	Approx.	Linear	Peterson et al. (2022)
		Logistic	$\mathbb{E}_s[\nabla p_{\text{high}}]$	$\hat{\beta}_{GLM}$ for $p(y_{\text{high}}\|\cdot)$	Approx.	Linear	Sobieszek et al. (2024)
Manipulation Path Estimation	Features x, Ratings y	Gradient Flow	Local $\nabla y(x)$	$\nabla f_y(x)$	No	Greedy Gradient Ascent	This work; Plisiecki and Sobieszek (2024)
	Model + OT	**PGFM**	Optimal Transport $\vec{v}^*$ solution for y	$\vec{v}(x)$ minimizing the PGFM loss	**Yes**	**Flow Matching**	**This Work**

Note. Column 'Respects $p_{\text{valid}}(s)$' indicates whether the resulting stimuli are constrained to remain within the ecologically valid stimuli distribution. Abbreviations: Classification Image (CI); Spike-Triggered Average (STA); Generalized Linear Model (GLM); Ordinary Least Squares (OLS); Optimal Transport (OT); Psychometric Gradient Flow Matching (PGFM); Feature Covariance ($\Sigma_{\mathbf{x}}$).

provides an important reason for why manipulation directions may easily fail to reflect human judgments once a manipulation violates the fourth condition of ecological validity—i.e., moves outside the support of f_y. We will see, however, that this property can also be used to understand how to overcome much of the current limitations of stimuli manipulation.

Existence of goodness-of-manipulation measures has recently been put into question. Sobieszek et al. (2024)point to a gap in the theoretical foundations, where the *goodness-of-fit* measures used to find $\hat{\beta}$ regression coefficients with the best **prediction accuracy** (held as the goal since Oosterhof and Todorov, 2008) do not overlap with measures of the best manipulation direction. They use as an example the method of Peterson et al. (2022) of using LASSO regression to obtain higher accuracy. LASSO or Ridge regression adds a penalty to shrink the OLS parameters (decrease $||\beta||$), yet when used for manipulation in the obtained $\hat{\beta}$ direction, the amount of distance needed to traverse to increase ratings by 1 grows inversely as $||\beta||$ decreases (Sobieszek et al., 2024). While correct for the L1 penalty of LASSO (which imposes sparsity in the nominal values of a direction vector), we prove in appendix D that the *L2 penalty of Ridge* has a useful reason for decreasing alignment with the gradient—as the obtained direction rotates precisely towards the direction of the Classification Image, providing a continuous connection between these two estimation approaches.

Idiosyncrasy in perceptions of trustworthiness and attractiveness has been given a significant amount of attention in recent literature (Albohn et al., 2024; Jaeger, 2021; Lavan & Sutherland, 2024; Martinez et al., 2020) and has been posed as an explanation for the difficulties these impressions pose for the discussed methods (Todorov & Oh, 2021). This could be corroborated by results of Peterson et al. (2022), where even with thousands of trustworthiness impression ratings, the linear model manipulation fails to obtain a significant increase in ratings—as noted by Sobieszek et al. (2024). They, however, pose alternative explanations: (a) valence-laden impressions are sensitive to loss of ecological validity under manipulation (MNAR), showing similar changes to trustworthiness ratings with either manipulation to change dominance or trustworthiness; they also highlight (b) the existence of fungible solutions in high-dimensional multiregression—where multiple directions can be associated with an increase in ratings, and thus the manipulation direction found with a strictly linear model may not be the one that leads to the best manipulation.

Indeed appendix B.6 we show that the gradient of one-

rating-per-participant mean models actually captures the average of gradients of (inestimable on their own) idiosyncratic individual impression models.

Within our method and validation studies we will tackle prediction model accuracy, Having trained more accurate impression models on the same data as Peterson et al. (2022), the results of our studies will test whether (a) valence-laden dimensions cannot be reliably increased due to their high idiosyncrasy, (b) past difficulties were caused by low impression model accuracy and are rectified by better impression models, or (c) past difficulties were caused by loss of ecological validity (MNAR) and are rectified with more theoretically-grounded selection of manipulation paths, which we formalize in the next section.

2.4 The High Road: From Local Directions to Manipulation Paths

The local analysis provides us with an optimal *direction* of manipulation at a valid point in feature space. To perform a substantial change in a psychological trait, however, we must move beyond an infinitesimal step and trace a continuous path through the space. We formalize this as a **manipulation path**. This path describes the evolution of a stimulus

$$x = \gamma(\ell), \quad \ell \geq 0$$

where ℓ denotes the distance traveled in feature space, with $\gamma(0) = x_0$ the initial stimulus.

A simplest manipulation path is the one used in linear model–based methods (Oosterhof & Todorov, 2008; Peterson et al., 2022). Using a linear impression model $f_y(x) = \hat{\beta}_{\text{OLS}} \cdot x + C$, the linear manipulation path is given by the constant shift by the vector of normalized regression coefficients (the direction of the gradient of f_y w.r.t. x):

$$\gamma(\ell) = x_0 + \ell\, \vec{v}, \quad \text{where } \vec{v} = \pm \frac{\hat{\beta}_{\text{OLS}}}{\|\hat{\beta}_{\text{OLS}}\|^2}$$

In general, we say a path is *generated* by a **manipulation vector field** $\vec{v}(x)$ (assigning a manipulation direction to every point x) if the path $\gamma(\ell)$ starting at a point $\gamma(0) = x_0$ satisfies the ordinary differential equation:

$$\frac{dx}{d\ell} = \vec{v}(x)$$

Which simply means the path follows, at each step, the direction of the manipulation vector field. Formally the collection of all such paths constitutes a **flow**, describing how all points in feature space move to increase their y ratings.

As our derivation of the locally optimal direction did not require linearity of the entire function f_y. The definition of the manipulation direction as the gradient allows the manipulation direction to change with x, accounting for interactions between the type of stimuli and influential features. This more general class of manipulation paths that linear model paths are an example of is given by the **gradient flow** generated by

$$\vec{v}(x) = \frac{\nabla f_y(x)}{\|\nabla f_y(x)\|^2}$$

the normalized impression gradient. Figure 2 illustrates how the manipulation paths differ in shape and how they relate to the vector field they follow. Both the linear and gradient flow paths capture only the **geometric** aspects of how the impression changes in feature space, but do not consider the **distribution** of valid feature combinations (which we aim to solve with our PGFM algorithm). The gradient-based paths follow, at each step, the locally optimal direction, which doesn't logically guarantee they maximize for a given ℓ the overall manipulation efficiency. This necessitates more comprehensive path-wise optimality criteria.

To understand why we note how constructing a manipulation path $\gamma(\ell)$ can be described as a (sequential) decision of which direction to follow at each step. To this end we use the impression model f_y, whose predictions have a given average MSE. As we integrate these steps, the MSE can accumulate or eventually explode once we leave the support of the feature distribution. Following statistical decision theory we can formulate this error as the **risk along the manipulation path**, $\mathcal{R}_\gamma(\ell)$:

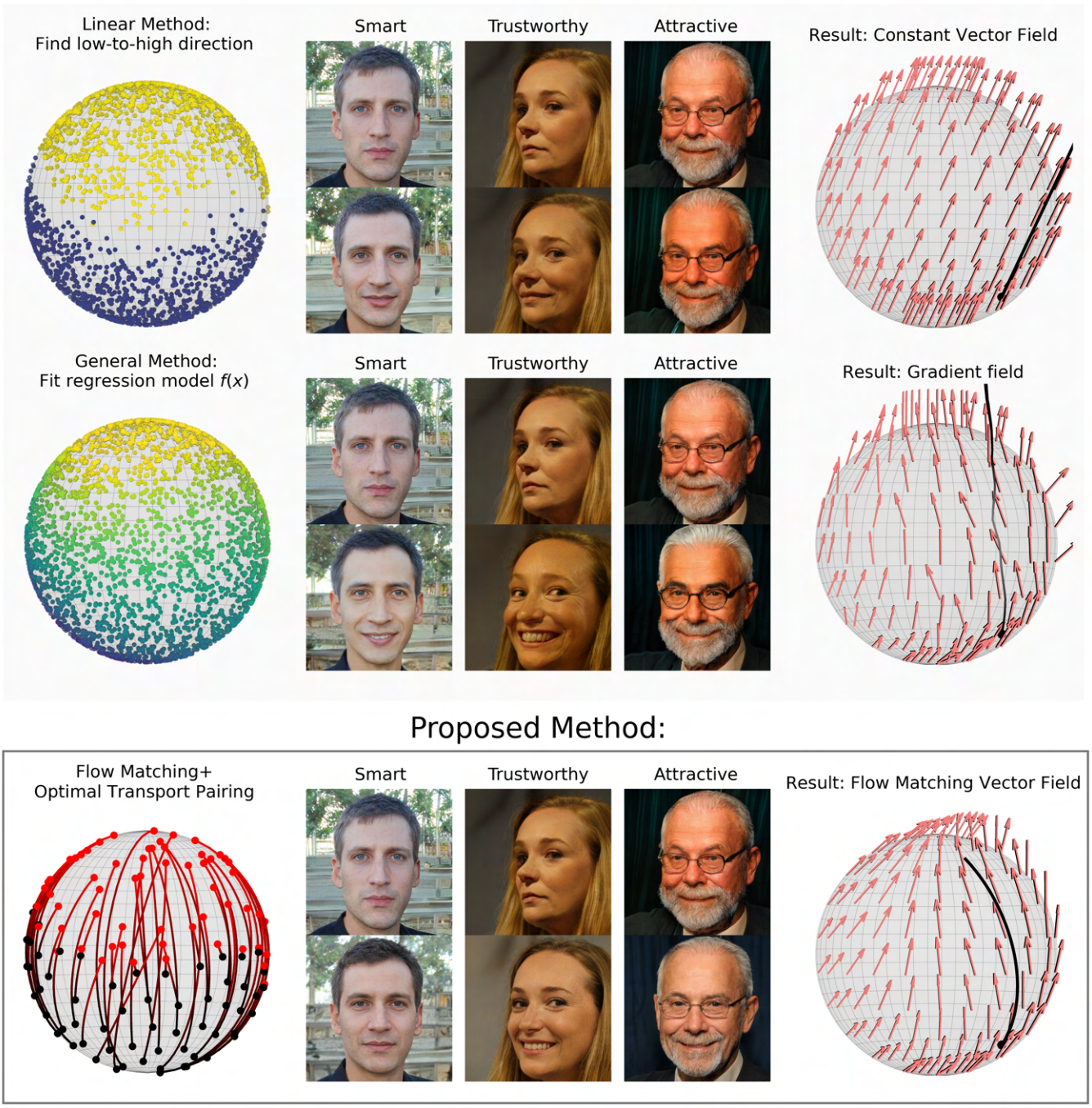
Definition 2.6 (Path risk)**.** Denote by y^* the true population mean that f_y is trying to estimate. We define the *risk along manipulation path* γ as the expected Mean-Squared Error at time t:

$$\mathcal{R}_\gamma(t) \triangleq \mathbb{E}_{x_0 \sim p_{\text{valid}}}\left[(y^*(x_t) - f_y(x_t))^2\right] \text{ with } x_t = \gamma(x_0, t)$$

We know that for $t = 0$, the statistical risk $\mathcal{R}_\gamma(0)$ recovers the setting of the well-established estimator MSE theory for i.i.d. data, capturing the bias/variance tradeoff in f_y model accuracy. However for $t > 0$, the expectation is taken over a different marginal distribution, causing all non-infinitesimal manipulation path estimation methods to have different, much more complicated accuracy criteria.

Considering the *risk along the manipulation path* explains the lack of statistical criteria for general manipulation methods, as the risk calculation depends on intractable integrals involving estimation errors of f_y combined with representation space and true $y^*(x)$ curvature. This insight however can let us identify a subset of methods for constructing such paths with superior a priori accuracy estimates. We consider the expected path risk over a family of manipulation paths constructed with a given method $\mathbb{E}_{\gamma \sim \text{method}}\left[\mathcal{R}_\gamma(t)\right]$, and observe that:

Figure 2: Comparison of controlled stimuli manipulation methods



Note. The spheres represent the support of the distribution of valid stimuli within a larger feature space. **Column 1** illustrates the feature representations and how they are used to train each method—a point represents a rated stimulus used to find a direction, fit a neural network, or construct optimal transport paths, **Column 2** shows the corresponding face manipulations using each method to increase 3 valence-laden impressions, while controlling for age and gender, **Column 3** illustrates the kind of vector field obtained with each method and an example manipulation path following its direction.

(MNAR $\implies$) In any high dimensional feature space, due to error-accumulation a generic method of constructing paths γ causes statistical risk to grow quadratically with distance traveled out-of-distribution. Conversely the expected path risk stays *approximately bounded* on an appropriate interval $t \in [0, T_{\max}]$ if one can construct paths γ that stay, on average, within high density regions of the initial data distribution support.

We can now understand that a greedy gradient minimizing path, while locally efficient, may rapidly accumulate errors in direction estimation and move to regions where f_y no longer accurately reflects human judgments. Such paths correspond to not only losing ecological validity, but also manipulation efficiency. We can now posit an explanation to the difficulties in manipulating valence-laden impressions. While most judgments are barely affected by how typical or valid a stimulus is (e.g. we can judge the color of an object whether it is realistic or not), emotional and valence-laden impressions are uniquely attuned to perceived unnaturalness. As a consequence, the unnaturalness of any feature itself induces a strong negative impression, and regardless of e.g. the strong positive correlation between a smile and trustworthiness within average faces, ignoring the feature distribution confounds the observed effects of the intended manipulation. Figure 3 shows empirical estimates of the risk along manipulation path for trustworthiness manipulations in our studies. The result lays credence to our claim that a successful manipulation method for the study of these important psychological phenomena requires the additional consideration of the stimuli distribution.

We now turn to our last consideration needed to satisfy the optimality conditions for paths. We note that while valence-laden dimensions are among the most adversely influenced by leaving the feature distribution support, all impression models, including those for controlled variables, experience a compounding increase in MSE along manipulation paths with no support preservation. This error turns into a systematic bias if the controlled variable correlates with the manipulated one. If we hope to de-confound and gain the ability to orthogonally manipulate different factors, we need to reconsider how we ought to introduce control.

Proposition 2.7 (Conditional Support Preservation (CSP))**.** *Let $p_{feature}(x \mid c)$ be the distribution of valid feature space points with controlled variable values c.*

Because for both manipulated and controlled variables:

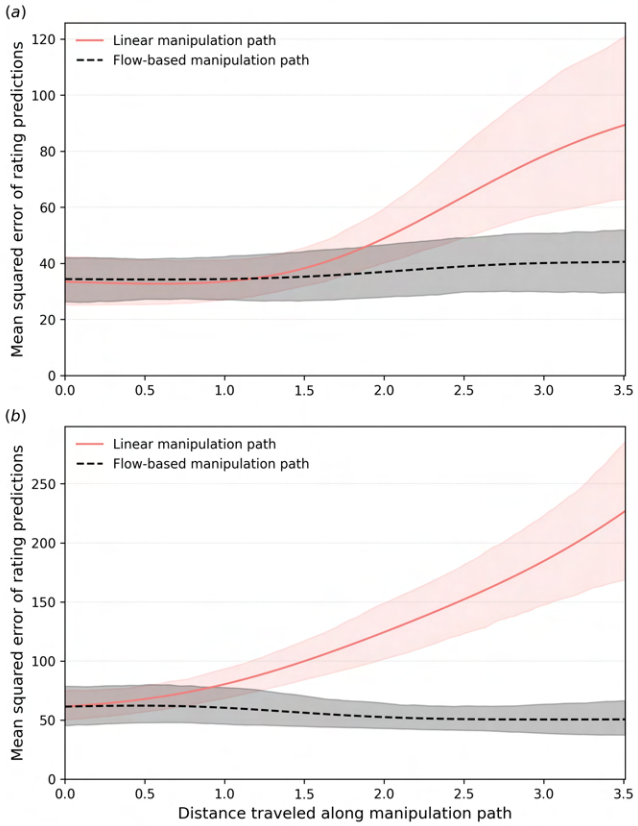
$$MNAR \implies$$
$$\mathbb{E}[\mathcal{R}_\gamma(\ell)] \text{ is bounded} \iff \gamma(\ell) \in \text{supp}(p_{feature}(x))$$

The orthogonalization procedure in Definition 2.4 is not enough to preserve controlled variable ratings. The combined criterion, we call Conditional Support Preservation, requires that a manipulation path stays within:

$$\gamma(\ell) \in \text{supp}\left(p_{feature}(x \mid c)\right) \text{ for all } \ell \in [0, \ell_{max}].$$

The remarkable result is that given a manipulation method that can successfully achieve this restriction, the stimuli sets constructed for different experimental conditions will not only have the additional variables included as a statistically controlled confound, but allow for fully orthogonal investigation of the manipulated and all explicitly controlled variables.

Figure 3: Empirical risk along manipulation paths of trustworthiness manipulation from the present studies.



Note. Panels use ratings from (a) study 3, and (b) study 4, of face stimuli randomly sampled (0 on the x-axis) and at 4 increasing levels of manipulation. The lines indicate the local average of mean squared errors of the impression model in predicting the actual mean rating observed. The x-axis represents the distance in feature space from the original point. Manipulations in study 4 were additionally controlled for age and gender, leading to the faster increase in MSE for the linear manipulation path. The shaded regions show a bootstrapped 95% confidence interval.

Definition 2.8 (Optimal Manipulation Path)**.** An optimal manipulation path $\gamma(\ell)$ with $\gamma(0) = x_0$ is one that:

1. Manipulation Efficiency: Maximizes the change in the impression per unit distance:

$$\gamma: \quad \max_{\gamma} \frac{\Delta y(\ell)}{\text{dist}(x_0, \gamma(\ell))} \qquad \left(\frac{\text{validity}}{\text{specicity}}\right)$$

2. Conditional Support Preservation (CSP): remains within support of $p_{\text{feature}}(x \mid c_i)$ so that all intermediate manipulations remain valid and do not affect the levels of explicitly controlled factors c_i (risk along

manipulation path does not explode along).

$$\Big(\text{ecological validity} \ \wedge \ \text{explicit control}\Big)$$

3. Dose–Response Relationship: The amount of manipulation ℓ has a stable correspondence with the predicted change in the impression, increasing $\Delta y(\ell)$.

Unlike for local directions, the path criteria do not simplify to a simple objective, instead representing a highly complex variational optimization problem over paths with additional distributional constraints. This both makes the required formal machinery more complicated, but also opens the space for constructing better manipulation estimation methods, owing to very recent advancements in machine learning algorithms for learning *probability distributions flows*.

2.5 Global Optimality: Flow Matching and Optimal Transport

Previously, to choose a step, the local analysis advised us to maximize the alignment with the gradient. Meanwhile the global path criteria advise us to better align with the data distribution if that step would've caused us to move too much outside of the distribution. What we need now is to discover where in feature space to take which advice. To show how to find complex manipulation paths that stay ecologically valid, respect controlled variables and maximize manipulation efficiency, in this final section we will bring together:

(1) The theory of Optimal Transport, for formally determining **shortest paths within valid** $p_{\textbf{feature}}(x)$. (2) Flow Matching, a ground-breaking algorithm for learning **complex global paths** from many simpler interpolation paths. (3) A transport cost that **maximizes manipulation efficiency** when used within an Optimal Transport Flow Matching algorithm. (4) A full-factorial stratification of controlled variables that will satisfy CSP and constrain the learned paths to those that satisfy **explicit control**.

Optimal Transport (OT) theory provides a mathematical framework for finding the most efficient way to move mass between one probability distribution and another (see Villani, 2008), which is optimal in the sense of minimizing the total distance that transport requires (a transport cost, measured e.g. by $\|x_1 - x_2\|$). To define source and target distributions for our task, we may consider drawing $2n$ samples from the overall data distribution $x \sim p_{\text{feature}}(x)$ and separating them based on the predicted impressions into two subsets, low and high in y ratings.

Given discrete samples from two distributions:

$$X_{\text{low}} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(n)} \end{bmatrix}, \qquad X_{\text{high}} = \begin{bmatrix} x^{(n+1)} \\ \vdots \\ x^{(2n)} \end{bmatrix}.$$

OT seeks a solution that minimizes total distance traveled when moving from points X_{low} to X_{high}, finding the optimal coupling (transport plan π) that pairs each sample from X_{low} with one unique sample from X_{high}. Pairing π can be thought of as a joint probability, where the non-zero entries $\pi(x^i, x^j) \neq 0$ indicate which example from X_{low} moves to which in X_{high}.

In this formulation, we can see OT as a *shortest path* problem, providing the set of unique linear interpolation paths that minimizes the total distance traversed. To apply OT theory to our optimal manipulation path problem we need to: (a) find a way to efficiently learn a general solution that would continuously transport any new starting point from X_{low} towards X_{high} (b) modify the learning algorithm to work across the whole range of values of y.

The Flow Matching algorithm is a recent groundbreaking algorithm for learning complex paths or *flows between probability distributions*, by training a model $u_\theta(x)$ to predict the underlying (manipulation) vector field.

We have discussed how a manipulation path from x_1, benefits from reaching an intended x_2 in as short (straight) path as possible (linear interpolation to the target being the shortest possible path). In their seminal paper Lipman et al. (2023), start from a similar insight. While we do not know what our true intended target x_2 is, they propose to solve the task of learning complex paths by constructing a larger training dataset of simple paths based on linear interpolations between any pair of points:

$$\gamma(t) = x_1 + t\vec{v}, \quad \text{where } \vec{v} = (x_2 - x_1), \quad t \in [0, 1]$$

Where such an interpolation path $x_1 \to x_2$ for all pairs (x_1, x_2) suggest their own plausible manipulation direction $\vec{v} = (x_2 - x_1)$, that we may consider when at a position x between the two points.

Lipman et al. (2023) show how to reconcile these different suggested directions, and find one, global vector field, by training a neural network $u_\theta(x)$ to minimize the direction prediction error across all such linear interpolation paths:

$$\text{FM Loss}(\theta) = MSE\left(u_\theta(x),\ \vec{v}(x \mid x_1, x_2)\right)$$

Importantly, they show that a model trained to minimize this objective learns a complex vector field $u_\theta(x)$ that can successfully transport any **new sample** x **from the first distribution** $x \sim p(x|y_{\textbf{low}})$ **to a valid sample** $x' \sim p(x|y_{\textbf{high}})$ **in the second**, preserving ecological validity.

Optimal Transport and Flow Matching can be combined (OTFM, see Tong et al., 2024) to additionally encode our preference about which kinds of manipulation paths are to be preferred during training. While the original Flow Matching algorithm uses all possible linear interpolations between X_{low} and X_{high}, OTFM additionally selects the pairs from which to learn according to an optimal transport plan to learn **the shortest paths** between them. In the above loss this corresponds to evaluating the MSE not on randomly sampled pairs but on pairs sampled from the transport plan (Tong et al., 2024 thus extend the result that FM preserves validity, thus doing this with minimal distance). The problem for our general optimal manipulation path problem is that both FM and OTFM are are used to transport between *just two* distinct distributions (e.g. from noise to images), thus requiring us to throw away all information about the impression y other than the binary "low or high".

Psychometric Gradient Flow Matching starts by observing that we may instead encode the the information about low-to-high transport directions in the choice of what we mean by "shortest" path. Formally, this requires constructing a distance metric associated with manipulating a particular impression, which we could assign to the space of stimuli together with the impressions they cause. manipulation efficiency Because we have defined it terms the ration of a distance (specificity) and a difference (validity), distance metric—that we posit enables the use of optimal transport theory for optimal manipulation—is:

Definition 2.9 (Transport Cost)**.** For any pair of stimuli (s_1, s_2), each associated with a the Manipulation Transport Cost (MTC) is the inverse of manipulation efficiency:

$$\text{MTC}(s_1, s_2) = \begin{cases} \dfrac{\|x_1 - x_2\|}{y_2 - y_1} & \text{if } \; y_2 > y_1 \\ +\infty & \text{if } \; y_2 \leq y_1 \end{cases}$$

Note that this definition of manipulation efficiency corresponds to our path-level criterion, not the gradient-based local criterion. In this way, MTC can be viewed as a better (lower variance) estimator of the actual manipulation efficiency along the interpolation path $x_1 \to x_2$.

The proposed cost penalizes any "inefficient" motion w.r.t. increasing the impression. Choosing pairs that differ more in features unrelated to y increases the numerator $\|x_2 - x_1\|$, increasing the cost, while pairs differing more in the target impression increase the denominator Δy, thus decreasing the cost. By the argument for proposition 2.3, the optimal parings that minimize MTC are the ones that align most with the gradient of f_y. We can thus use the flow matching algorithm to learn the most efficient manipulation paths for the whole space. Training on many simpler, maximally efficient interpolations paths that stay inside the ecologically valid stimuli. Notably, this approach enables learning manipulation paths based only on the initial and target point ratings, overcoming the limitation of OLS/GLM parameter estimation approaches requiring explicit assumptions about the shape (linear, logistic) of the relationship between features and impressions.

This way PGFM satisfies the manipulation efficiency objective under the restriction to direction that lay within the support. The last requirement to solve is the explicit control of confounding variables. Directly from our conditional support preservation condition we can deduce that given how restricting the pairings to those that lay within the support of $p_{\text{feature}}(x)$, we can moreover satisfy the requirement to stay within the conditional support of $p_{\text{feature}}(x|c_i)$ by further requiring the pairings to be made between stimuli of the matching controlled variable values. This is achieved by fully stratifying the feature space points into bins, in our case, of (age, gender) values and calculating optimal pairings with these strata.

Remark 2.10 (Transport Cost with controlled variables). Computing many optimal transport plans, one within each strata can be equivalently formalized as one global transport plan computation, by augmenting our *MTC* definition to have an infinite transport cost between samples with different values of controlled variables. Let vectors ω_1, ω_2 indicate the values of controlled variables:

$$\text{MTC}(s_1, s_2) = \begin{cases} \dfrac{\|x_1 - x_2\|}{y_2 - y_1} & \text{if } \; y_2 > y_1 \\ +\infty & \text{if } \; y_2 \leq y_1 \\ +\infty & \text{if } \; \omega_1 \neq \omega_2 \end{cases}$$

Which interestingly suggests that optimal experimental manipulation in orthogonal experimental designs has the topology of a Fibered Optimal Transport (see Peszek and Poyato, 2023).

Finally, we need to observe that while the original "dynamic optimal transport" formulation used in the Conditional Flow Matching algorithm transports points over a fixed, unit time interval $t \in [0, 1]$. Meanwhile, the manipulation paths of reverse correlation can in principle continue for all $t \geq 0$ until the representation no longer corresponds to a valid output. To account for this we define the following parametrization of interpolation paths that changes the training objective for along any learned 'consensus path' to increase the attribute y at a constant rate:

$$\gamma(t) = x_1 + t\vec{v}, \quad \text{where } \vec{v} = \frac{(x_2 - x_1)}{|y_2 - y_1|^2}, \quad t \in [0, y_2 - y_1]$$

A more complete list of expressions for the Flow Matching Loss and Manipulation Transport Cost can be found in the online notebook tutorial: https://wiki.deixis.ai/Papers/Sobieszek%2C+Siemi%C4%85tkowski/Traversal-Flow+Matching+Appendix

3 Method

3.1 Psychometric Gradient FM Algorithm

1. **Train impression prediction models:** For the manipulated y and controlled c_i variables.
2. **Define source and target:** We sample a large number of feature space points $x \sim p_{\text{feature}}(x)$ and use the impression models to obtain ratings $s \to (x, y, c_1, c_2 \dots)$.
   Any point (apart from the max/min y values) could be used as either a start or end point of an interpolation path if it has a small enough MTC(s_1, s_2).
3. **Split into groups based on control strata**
4. **Construct Optimal parings:** Within each strata approximate the optimal transport plan w.r.t. to the manipulation transport cost 2.9.
   A useful heuristic here is to assume for each s_1 we find two stimuli s_2 and s_2' that minimize the MTC for either increasing or decreasing the impression.
5. **Learn the manipulation vector field:** Train a neural network $v_\theta(x)$ over batches of points x along interpolation paths $\gamma(t \mid x_1, x_2)$ between the OT pairings, learning to minimize the MSE to manipulation directions $\vec{v} = (x_2 - x_1)$
6. **Generate Manipulations:** To manipulate a new face x_1, we flow along the learned manipulation vector field $d\gamma/dt = \vec{v}_\theta(x)$. The obtained points correspond to increasingly manipulated stimuli.

The algorithm is implemented in full detail in the PGFM tutorial notebook shared in section 3.6, where additional theoretical guarantees are described (given by the Benamou-Brenier formulation, see Villani, 2008).

3.2 Predictive models of face impressions

In our approach, we utilized StyleGAN2 (Karras et al., 2020), a generative adversarial network (GAN) that maps points in a feature space onto high-resolution, photorealistic face images. StyleGAN2's architecture ensures the its 512-dimensional feature space is *well-conditioned*, meaning it does not encode discrete features as individual values but as directions within this space, which include facial features, expressions, hairstyles, positioning, clothing and background. We used the StyleGAN feature space to train impression models on the OMI dataset (Peterson et al., 2022), which contains ratings of subjective and visual impressions.

We trained neural network prediction models to estimate the mean human impression for each trait based on feature space representations. The resulting models are fully differentiable, allowing us to estimate not only ratings but also directions of fastest change at each point via the model gradient. To asses accuracy, we performed cross-validation using 10% of the OMI dataset as a held-out validation set, evaluating correlations with human ratings. Compared to results obtained by Peterson et al. (2022), our method significantly increased prediction accuracy (Figure 4).

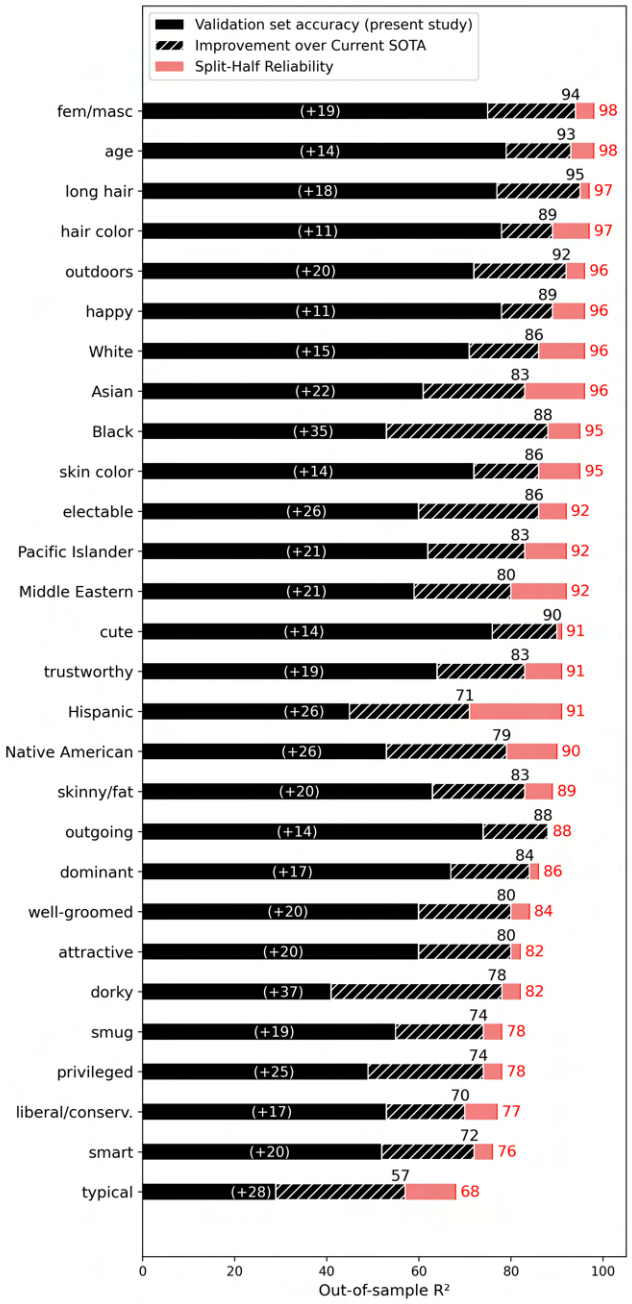
Prediction accuracy alone, however, is not sufficient for our purposes. To minimize error accumulation along manipulation paths, the model's gradients must be sufficiently smooth to prevent sudden directional changes that easily lead paths out of distribution. We therefore built an ensemble model, training eight smaller models (3 layers, batch normalization, ReLu nonlinearity) independently and average their predictions, which effectively averages their gradient directions when differentiated. Each model was heavily regularized to minimize overfitting. As shown in Appendix D, ridge (L2) regularization encourages alignment with the data distribution, converging toward the classification image—the prototypical direction most aligned with the feature distribution. We can connect this insight—via a known result that neural network models (with ReLU activations) learn a piecewise linear function of the input—to the regularization technique of high weight decay, which adds L2 regularization to network weights, and is crucial for low-variance estimation of gradient directions. The models were trained for 300 epochs with weight decay of 0.1, learning rate 1e-3 (cosine-schedule), and batch size of 128.

3.3 Flow models

For our models that learn a flow vector field (a model mapping any x to a direction $u_\theta : \mathbb{R}^d \to \mathbb{R}^d$), we constructed a vector-to-vector neural network with a conditional transformer architecture (inspired by Peebles and Xie, 2023). The architecture diagram is available in the Supplementary Materials. Transformer architectures are largely responsible for the recent explosion in AI methods, thanks to their high representational capacity, which enables the learning of highly complex transformations from large datasets. For flow models, the dataset consists of paired examples for interpolation. The consistent finding across the flow matching literature is that such flow models do not simply save the pairings between which to interpolate, but actually find remarkably sophisticated transformations that generalize parts of transformations across the training set (see e.g. Eskimez et al., 2024 for remarkable generalization in voice editing). To ensure this property, a large diverse dataset is needed, to ensure this, within each strata, we created our flow training dataset with 200 thousand valid face coordinates in each 50 age $\times$ 2 gender strata.

The computation needed to construct such datasets may be prohibitively slow on consumer hardware, so we share the combined 30GB of training data created for this work, consisting of the feature space points and ratings in each

Figure 4: Cross-validation prediction accuracy comparison.



Note: Black bars represent out-of-sample R^2 values for our neural network regression models across different psychological dimensions. Hatched parts indicate improvement over current state-of-the-art (SOTA) prediction models (Peterson et al., 2022), with improvement values shown in parentheses. Red bars show split-half reliability, approximating the ceiling for prediction accuracy.

strata $(x, y, \text{age}, \text{gender})$ and their optimal pairings for constructing training manipulation paths. This, along with all impression prediction models, that are sufficient to enable imposing additional controlled or manipulated variables for specific research cases are made available in section 3.6.

3.4 Validation studies design

All studies used between-subjects designs to minimize potential carryover or contrast effects and were conducted in accordance with ethical guidelines. Informed consent was obtained from each participant and they had the option to withdraw at any time. The sample size for each study was determined based on established guidelines, which indicate that stable face ratings typically occur with approximately 20 unique ratings per stimulus (Hehman et al., 2018). Respondents rated faces on specific traits using a dense (up to two decimal points) sliding scale ranging from 0 (lowest level of a given trait) to 100 (highest level). They saw each face only once in one of the experimental conditions, with face-to-condition assignment randomized across participants. The order in which the faces were presented was also randomized to control for order effects. To ensure data quality, we included several attention checks in all studies, including recall of the names of the dimensions being rated (all studies) and consistency checks between reported demographic information at different points in the study (Mturk study only).

3.5 Method variants and validation studies

The primary aims of the validation were to: (1) demonstrating the effectiveness of our flow-based manipulation in difficult valence-laden manipulation tasks, (2) compare the effectiveness of the flow-based method to vector-based manipulation obtained on the same impression data (3) show the robustness of our flow-based method under explicit constraints with controlled variables in the case of vector-based manipulation, (4) empirically evaluate the merits and limits of the gradient generalization of model-based manipulation. In all studies all participants rated saw each face in only one random manipulation-level condition, making all reported comparisons of observed manipulation effectiveness *between-subject*. We take advantage of this design to compare different methods, constructing the experimental conditions as to correspond to a linear change in ratings (between values of *manipulation strength*).

In Study 1, we tested the ability of gradient-based manipulation to effectively manipulate visual trait judgments of age and race-typicality. Using samples of US and Polish adults, we aimed to demonstrate the (cross-cultural) effectiveness of gradient-based manipulations for judgments with a stable perceptual basis.

In Study 2, we evaluated the ability of the gradient manipulations to increase or decrease the impressions of trustworthiness and dominance. (Oosterhof & Todorov, 2008; Peterson et al., 2022; Sobieszek et al., 2024; Sutherland et al., 2013). We hypothesized that the gradient-based manipulation will reproduce the observed pattern of a linear response under dominance manipulation and a non-linear response under trustworthiness manipulation, with a plateau for increasing this judgment.

In Studies 3 and 4 we focused on this hardest manipulation objective of increasing valence-laden impressions of trustworthiness, attractiveness, and perceived smartness. In Study 3, we tested our initial flow-based model trained for each of the 3 traits under additional control for the age and gender of the manipulated face. For the same initial faces we compared our approach with the existing linear model-based methods. We use the linear models of Peterson et al. (2022), which because they report no significant effect for increased trustworthiness, we combine with the initial stimuli sampling strategy of Sobieszek et al. (2024). They propose to trade-off stimuli diversity to increase the range within which linear manipulation remain effective by sampling initial faces only from the center of the feature distribution (what is called truncation in the StyleGAN model). Notably, in this study the flow matching was trained to control for age and gender, while the linear model method was allowed to altered perceived these traits to get an upper estimate of the amount of change that was possible.

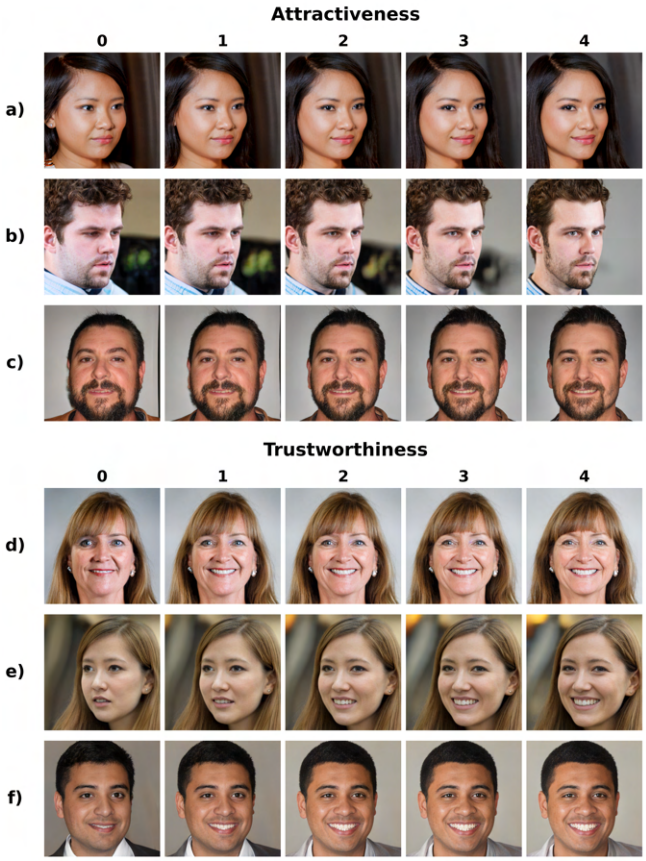
Finally, in Study 4, we refined our flow-matching approach and imposed controls on the age and gender of the manipulated face in the case of the linear model method (using the orthogonalization approach endorsed in Peterson et al. (2022)). This methodological refinement allowed for a more rigorous comparison, demonstrating the superior effectiveness of the flow matching technique. These results confirmed the ability of our approach to selectively manipulate target features while holding other facial characteristics unchanged.

Example stimuli manipulations rated in Study 4 are shown on Figure 5.

3.6 Data and code availability

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow JARS (Appelbaum et al., 2018). The designs of the studies and their analyses were not preregistered. Data analyses and model training were conducted using Python 3.10 and PyTorch 2.1. All stimuli, notebook with analysis code, and trained predictive models used in the reported studies are available in an open repository on the Open Science Framework: https://osf.io/2xsn7/?view_only=892c516b19654487a6e6f2bc2734d715. Datasets of age-gender stratified face coordinates for flow model training are available at: https://drive.google.com/drive/folders/13VhJ4diCHHA9kRC9Wp2bjhkGMLdKjg-n?usp=sharing. The tutorial code, which implements the core training algorithms used in the paper, can be accessed in text form at: https://wiki.deixis.ai/Papers/Sobieszek%2C+Siemi%C4%85tkowski/Traversal-Flow+Matching+Appendix and in interactive python notebook form at Google Colaboratory: https://colab.research.google.com/drive/13zPnxhcpcc7D8Gp5PHaDaaUNZRc_bi_R.

Figure 5: Example PGFM stimuli manipulations rated in Study 4



Note: The horizontal axis indicates the manipulation strength; 0 indicates the control condition. Additional examples of stimuli and their discussion can be found in the Supplementary Materials.

4 Studies

Study 1

Study 1 aimed to evaluate the effectiveness of gradient-based manipulation paths using our impression models. While linear methods depend on the same manipulation direction, affecting identical stimulus features, the gradient flow method enables manipulation of different features for each face, aligning more with the local gradient of our impression model. In this study, we specifically examined whether the gradient manipulation path approach could

effectively alter more stable perceptual features: race typicality, perceived age, and happiness. A key criterion for successful manipulation, besides a significant change in ratings, is the risk—the accuracy of the predicted effect across different levels of manipulation. For each face, the amount of feature space manipulation ℓ was chosen based on our impression model's predicted change in ratings—selecting manipulation strengths that should result in a linear rating change between conditions according to model predictions. This way, we can observe whether either manipulation type violates this prediction, indicating whether a manipulation path remains within the support of the impression model or leads to spurious efficiency predictions.

For the models of impressions reflecting racial categorizations, transforming faces of one group into another is of little scientific value. The scientific investigation of racial outgroup bias may find a method of varying intragroup typicality valuable. Specifically, tested whether gradient-based constrained manipulations with models predicting the impression 'black' applied to members of that category, can reliably influence the continuous categorization ratings, potentially addressing questions of inter-/intragroup heterogeneity that are both theoretically important and practically relevant for understanding real-world social perception processes.

To strengthen the results, we conducted parallel validation across two culturally distinct populations: One group comprised US participants from a multi-ethnic society with extensive intergroup contact; the other group comprised Polish participants from a predominantly ethnically homogeneous context with limited direct exposure to racial diversity. This cross-cultural design allowed us to establish whether gradient-based manipulations could be applied across different cultural contexts.

Participants A total of 129 participants were recruited through two sources. The first sample consisted of US adults recruited via Amazon Mechanical Turk (MTurk). After excluding individuals who failed attention or bot checks, this sample included 60 participants (63.3% women, 35.0% men, and 1.7% who identified as other gender, M_{age} = 27.77, SD_{age} = 4.54). The second sample consisted of Polish participants recruited via Facebook who participated voluntarily. After excluding those who failed the attention check, this sample included 69 participants (73.9% women, 21.7% men, and 4.4% who identified with other gender, M_{age} = 22.43, SD_{age} = 6.21).

Design and materials We initially generated 50 random StyleGAN faces per attribute and applied the exclusion criteria outlined in the Supplementary Materials, resulting in 60 faces used in this experiment. The final set included 15 faces each for the age, happiness, and race-typicality (different in each sample) manipulations. Each face was manipulated to create three versions with different trait intensities-decreased, neutral (unmanipulated, randomly generated face), and increased-on the target trait. Specifically for race typicality, initial faces were randomly selected from those predicted to have a typicality score of between 65 and 80 on a 100-point scale.

Results

To assess the effectiveness of our trait manipulation method, we conducted paired sample t-tests with Bonferroni correction for multiple comparisons across both samples.

US sample results For the US sample, paired sample t-tests revealed significant differences in perceived age between conditions. Participants rated faces in the increased age condition (M = 50.34, SD = 8.61) as significantly older than those in the neutral condition (M = 41.14, SD = 7.23, t = 5.96, $p < .001$, d = 1.54). Faces in the decreased age condition (M = 35.12, SD = 6.46) were perceived as younger than those in the neutral condition (t = -3.97, p = .001, d = -1.03).

For race typicality, paired sample t-tests demonstrated that faces in the increased Black typicality condition (M = 62.55, SD = 5.41) were rated as appearing more prototypically Black than those in the neutral condition (M = 56.92, SD = 6.69, t = 2.54, p = .023, d = 0.66). Faces in the decreased typicality condition (M = 43.32, SD = 8.39) were rated as significantly less Black-typical than those in the neutral condition (t = -5.79, $p < .001$, d = -1.49).

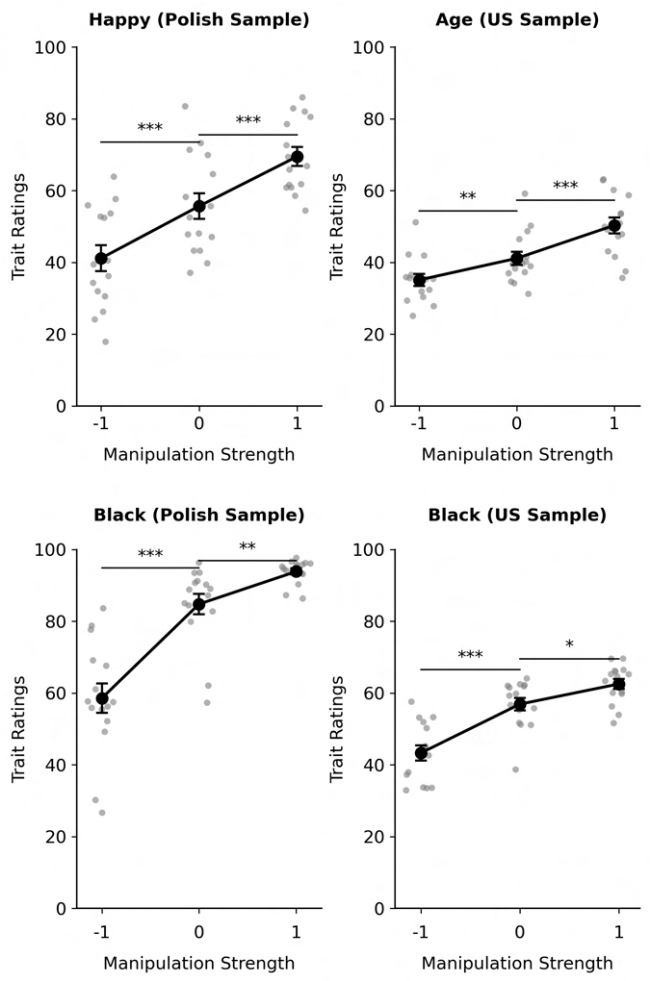
Polish sample results In the Polish sample, paired sample t-tests also showed significant differences across all manipulation conditions. Faces with increased Black typicality (M = 93.88, SD = 3.37) were rated as more prototypically Black than those in the neutral condition (M = 84.81, SD = 11.12, t = 3.69, p = .002, d = 0.95). Faces with decreased typicality (M = 58.59, SD = 15.98) were rated as significantly less Black-typical than neutral faces (t = -10.78, $p < .001$, d = -2.78).

For happiness ratings, faces in the increased intensity condition (M = 69.53, SD = 10.21) were perceived as significantly happier than those in the neutral condition (M = 55.70, SD = 13.96, t = 6.31, $p < .001$, d = 1.63). Faces in the decreased happiness condition (M = 41.17, SD = 14.02) were rated as significantly less happy than neutral faces (t = -6.00, $p < .001$, d = -1.55). The summarized results of these four analyses are presented in Figure 6.

Discussion

Our results showed that the gradient-based manipulation method successfully altered perceptible facial features in both the US and Polish samples (see the Supplementary

Figure 6: Gradient flow validation study 1 results.



Note: Error bars represent 95% Confidence Intervals. The black horizontal lines indicate significant differences in means, but note that we did not test the differences between the increased trait intensity and decreased trait intensity conditions. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Materials for the examples of the manipulated stimuli). The consistent effects observed for both demographic traits (age, racial typicality) and affective expressions (happiness) highlight the versatility of this approach in manipulating different dimensions.

Notable differences were observed in the responses of US and Polish participants to Black typicality manipulations, likely reflecting differences in social context and experiences of intergroup contact. Participants from the US, a multi-ethnic society with extensive racial diversity, showed more nuanced differentiation across manipulation levels. In contrast, participants from Poland, a predominantly ethnically homogeneous country, exhibited higher baseline ratings and more compressed responses. These patterns are consistent with prior research on cross-cultural face perception (e.g. Vingilis-Jaremko et al. 2020). This suggests that familiarity with facial variation within racial categories influences perceptual sensitivity to typicality gradations.

Notably, our manipulation framework was effective in both cultural contexts. The large effect sizes observed in both samples confirm that targeted modifications to visual traits can be made without compromising facial identity. This lays the groundwork for more complex manipulations, as implemented in subsequent studies.

Study 2

In this study, we evaluated the ability of direct gradient manipulation paths. This time, we focused on features strongly associated with affective impressions, such as dominance and trustworthiness. We trained a differentiable model tailored to each face, integrating person-specific gradients that account for local variations in latent space.

Participants For this study, Polish participants were recruited through Facebook and took part voluntarily. After excluding those who failed the attention check, the final sample consisted of 58 individuals (72.4% women and 27.6% men, $M_{age} = 20.41$, $SD_{age} = 1.65$). Upon completion of the study, participants were offered the opportunity to enter a prize draw for one of three monetary rewards of 50 PLN each.

Design and materials The final set consisted of 40 faces (20 faces for each dimension: dominance and trustworthiness. Each face was manipulated using both methods to create three versions (reduced intensity, neutral/ unmanipulated, increased intensity).

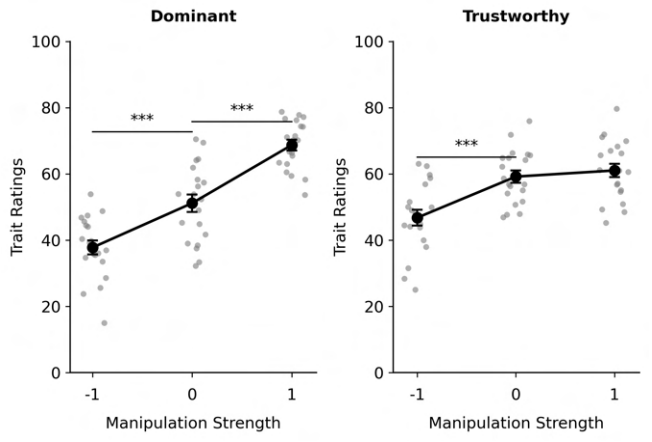
Results

To assess the effectiveness of our per-face manipulation approach, we conducted paired-sample t-tests with Bonferroni correction for multiple comparisons to analyze the effects of our manipulations on perceived dominance and trustworthiness.

The paired samples t-tests revealed significant differences in perceived dominance between the manipulation conditions. Faces in the increased dominance condition ($M = 68.72$, $SD = 7.35$) were rated as significantly more dominant than those in the neutral condition ($M = 51.16$, $SD = 11.79$; $t = 6.32$, $p < .001$, $d = 1.41$). Similarly, faces in the decreased dominance condition ($M = 37.77$, $SD = 9.46$) were rated as significantly less dominant than those in the neutral condition ($t = -4.51$, $p = < .001$, $d = -1.01$).

For trustworthiness, our analyses revealed an asymmetrical pattern of effects. While faces in the decreased trustworthiness condition ($M = 46.78$, $SD = 10.67$) were rated as significantly less trustworthy than those in the

Figure 7: Gradient flow validation study 2 results.



Note: Error bars represent 95% Confidence Intervals. The black horizontal lines indicate significant differences in means, but note that we did not test the differences between the increased trait intensity and decreased trait intensity conditions. ***$p < .001$.

neutral condition (M = 59.18, SD = 8.20, t = -5.48, p < .001, d = -1.22), the difference between the increased trustworthiness condition (M = 61.03, SD = 9.01) and the neutral condition did not reach statistical significance (t = 1.38, p = .183, d = 0.31). Figure 7 summarizes the results of these analyses.

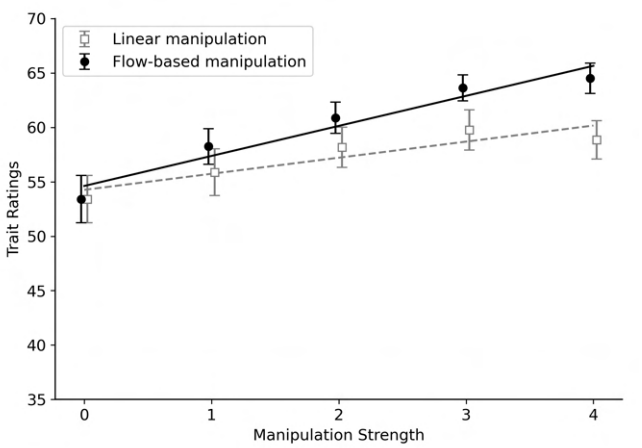
Discussion

The results highlight the major asymmetry between the risk along manipulation paths, decreasing versus increasing valence-laden traits, such as trustworthiness. While dominance These findings underscore that effective manipulation of such features requires methods that take into account the existing facial characteristics of individual faces, since both manipulations that are constant across all stimuli or are based purely on the model gradient may force faces to look in ways that undermine the intended feature attribution (e.g., increasing to an unnatural extent the magnitude of the smile or changing the image hue beyond what looks natural - examples and more detailed qualitative descriptions of obtained manipulation are provided in the Supplementary Materials).

Study 3

In Study 3, we evaluated the manipulation effectiveness for increasing valence-laden impressions for the flow-matching approach and a linear model approach similar to Peterson et al. (2022). Because increasing valenced trait intensity is more challenging than decreasing it, we investigated increasing these traits with four manipulated conditions.

Figure 8: Smartness validation study 3 results.



Note: Error bars represent 95% confidence intervals. Regression lines show the relationship between manipulation intensity and rated smartness for our flow-matching method (black) and the linear model method (gray).

Noting that the manipulation to increase trustworthiness in Peterson et al. (2022) study did not reach statistical significance (Sobieszek et al., 2024), we aimed to use the analysis to find suitable conditions under which a linear model-based manipulation may succeed in increasing valence-laden dimensions. We note how the only hitherto method to achieve a linear increase in trustworthiness judgments was the morphing method-of-constant-stimuli of Sutherland et al. (2013). The tradeoff that explains this success is that the stimuli used contain features that are average between two or many different stimuli. This necessarily decreases their variance, but also crucially makes sure that the magnitude of features stays within bounds of the natural covariation. We hypothesized that the guarantees this provides for support preservation (ecological validity) is the primary factor explaining the success of morphing methods in this respect. Based on this, we propose changing the initial sampling distribution of stimuli to simulate the favorable conditions that face averaging provides. The StyleGAN architecture includes a parameter called truncation $\lambda \in [0, 1]$, which moves initially sampled coordinates closer to the global average of features. Due to the high dimensionality, the naturally sampled faces are always approximately the same distance from this center. This fact provides an interpretation for our chosen truncation value of 0.6 in this study, which represents the sampling of faces with diminished feature magnitude, as if they were averages of three random faces. This way, the start of the manipulation is exceptionally far from the ends of the feature distribution and may allow a best-case estimate for linear methods. We return to the natural, "real-world" feature magnitude scenario in Study 4.

Participants

We recruited participants through an online panel administered by an external research company in Poland. After excluding participants who failed the attention check, the final sample consisted of 425 individuals (45.2% women, 54.4% men, 0.02% other gender identifications, and 0.02% that did not want to reveal their gender, $M_{age} = 44.06$, $SD_{age} = 15.16$).

Design and materials

The experimental stimuli consisted of 40 unique face identities per dimension, with manipulations targeting three valence-laden psychological attributes: trustworthiness, attractiveness, and smartness. Each dimension included nine experimental conditions: a baseline condition with randomly generated faces and eight manipulation conditions with different intensity levels. Four of these intensity levels were generated using our optimal transport-based flow matching technique, while the remaining four were generated using the linear model method, using the vectors of coefficients obtained by Peterson et al. (2022).

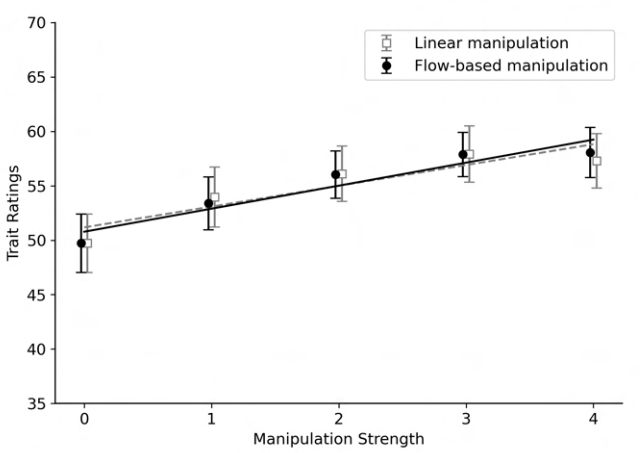
Results

To assess the effectiveness of both manipulation approaches, we conducted regression analyses comparing the flow matching technique with the linear model method. For the perceived smartness manipulations (Figure 8), the flow-matching approach showed notable effectiveness ($F(1, 198) = 111.90$, $p < .001$, $R^2 = .36$), outperforming the linear model method ($F(1, 198) = 21.59$, $p < .001$,$R^2 = .10$). For the trustworthiness manipulations (Figure 9), both methods again yielded significant effects. The flow-matching approach showed a slightly stronger effect ($F(1, 198) = 31.47$, $p < .001$, $R^2 = .14$) compared to linear model method ($F(1, 198) = 20.34$, $p < .001$, $R^2 = .09$). For attractiveness manipulations (Figure 10), both methods produced significant linear relationships. The flow-matching approach showed some extend of effectiveness ($F(1, 198) = 18.83$, $p < .001$, $R^2 = .09$), however, the linear model method showed a stronger effect ($F(1, 198) = 36.64$, $p < .001$, $R^2 = .16$).

Discussion

The results showed that the addition of truncated features made both methods effective in increasing valenced impressions. The linear model method showed stronger effects for attractiveness, while our approach performed better for perceptions of smartness. However, a closer examination of the manipulated stimuli revealed a critical limitation of the linear model method: it altered perceived age and gender while manipulating trustworthiness and attractiveness. These confounds likely affected its effectiveness, particularly for these two dimensions. The flow matching approach, through its optimal transport-based design, preserved these attributes while manipulating intended dimensions. This limitation motivated us to implement explicit age and gender control procedures for the vector-based approach in Study 4, which allowed us to make a more rigorous comparison of each method's ability to selectively manipulate target attributes while holding perceived age and gender constant.

Figure 9: Trustworthiness validation study 3 results.



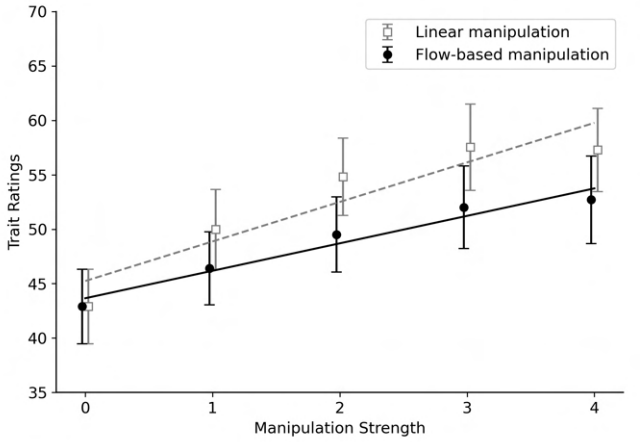
Note: Error bars represent 95% confidence intervals. Regression lines show the relationship between manipulation intensity and rated trustworthiness for our flow-matching method (black) and the linear model method (gray).

Study 4

In Study 4, we added two final components to the flow model trained with PGFM. Firstly, when drawing faces from the face feature distribution to manipulate e.g. trustworthiness, their initial trustworthiness rating follows the natural training set distribution. As a consequence, constructing the flow matching dataset from purely random sampling contains much fewer examples for very high-rated examples—we corrected this imbalance by deliberately sampling additional high-rated faces to include more information about how the flow can reach these stimuli. Secondly, we can additionally observe a fact about the solution vector field that the flow model is targeting, namely that if the manipulation paths are to have a positive dose-response correspondence, the solution can be decomposed into a sum of two orthogonal vector fields: the first being co-linear with the gradient (to increase the impression) and the second being the unknown distributional correction for the trajectory to remain within the data support. Thanks to this observation, we added the gradient field as the starting point for the PGFM model, enabling it to devote more training time to satisfying support preservation.

We observed that linear model manipulation in Study

Figure 10: Attractiveness validation study 3 results.



Note: Error bars represent 95% confidence intervals. Regression lines show the relationship between manipulation intensity and rated attractiveness for our flow-matching method (black) and the linear model method (gray).

3 exhibited a substantial change in perceived age and gender when modified along the dimensions of trustworthiness and attractiveness. To mitigate this confound and compare how both methods under constrained manipulation, we introduce the control by orthogonalization for age and gender as described in Peterson et al. (2022) paper, and constrain the flow models to stay within the same perceived gender and within ±1 the same predicted age.

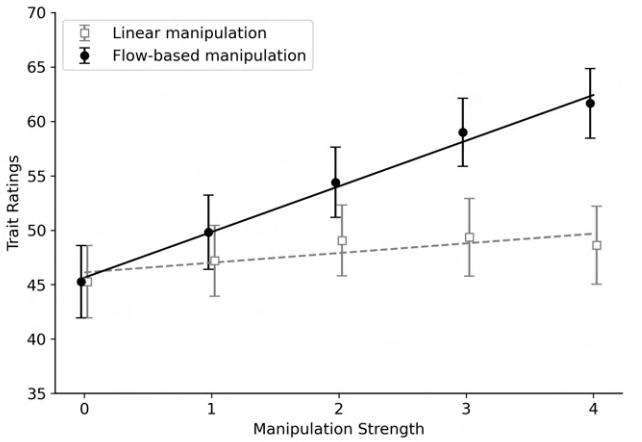
Participants A new sample of participants was recruited through the same external online panel as in Study 3. After all the exclusions the final sample consisted of 414 individuals (50.24% women, 49.52% men, and 0.24% other gender identifications, M_{age} = 50.79, SD_{age}= 15.90).

Design and materials The stimulus set consisted of 50 faces per dimension (trustworthiness and attractiveness). Each dimension included nine experimental conditions: one condition with randomly generated faces and eight manipulation conditions increasing intensity (four levels created using our flow matching method and four using the linear model method).

Results

To assess the effectiveness of both manipulation approaches, we conducted linear regression analyses comparing the flow matching technique with the linear model method (Peterson et al., 2022). For the attractiveness manipulations (Figure 11), we observed a significant linear relationship for the flow matching approach ($F(1, 248) = 64.24$, $p < .001$, $R^2 = .21$). In contrast, the linear model method did not produce significant changes in

Figure 11: Attractiveness validation study 4 results controlling for age and gender.



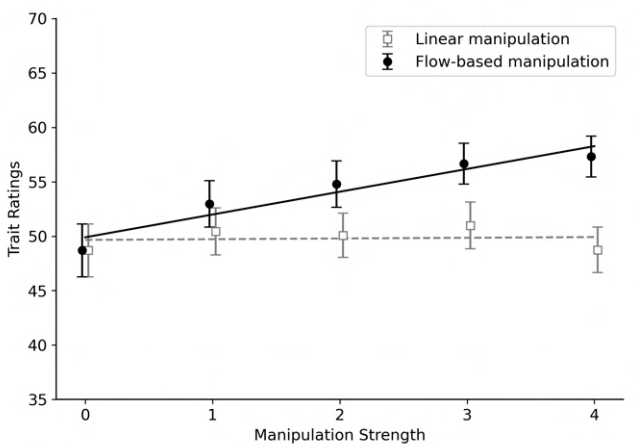
Note: Error bars represent 95% confidence intervals. Regression lines show the relationship between manipulation intensity and rated attractiveness for our flow-matching method (black) and the age- and gender-controlled linear model method (gray).

attractiveness ratings ($F(1, 248) = 2.65$, $p = .105$, $R^2 = .01$). The same pattern emerged for the trustworthiness manipulations (Figure 12). The flow matching technique changed trustworthiness ratings in a linear fashion ($F(1, 248) = 38.11$, $p < .001$, $R^2 = .13$), while the linear model method showed no significant effect ($F(1, 248) = 0.03$, $p = .853$, $R^2 < .001$).

Discussion

The results of this study demonstrated the significant advantage of our flow matching approach over the linear model method in the presence of controlled variables. While in Study 3, the flow methods controlling for age and gender were similarly effective to the linear manipulation not controlling for these confounds, in Study 4 we additionally observe this efficiency could not be achieved at all when controlling for confounds in the linear approach. This highlights, firstly, how critically the linear manipulation relied on influencing perceived gender and age, secondly, that the new algorithmic approach to control we perused—grounded in the consideration of feature distribution preservation—uniquely provides a powerful way of controlling confounding factors while retaining manipulation effectiveness. This consistency is noteworthy, as it also overcomes the limits on the increase in valence-laden impressions reported with previous methods. This result paves the way for successfully using deep learning techniques to achieve precise, orthogonal manipulation of stimuli. This can, in turn, allow researchers to isolate the functional effects of specific impressions and study the interactive effects of factors like age, gender, and race

Figure 12: Trustworthiness validation study 4 results controlling for age and gender.



Note: Error bars represent 95% confidence intervals. Regression lines show the relationship between manipulation intensity and rated trustworthiness for our flow-matching method (black) and the age- and gender-controlled linear model method (gray).

while using realistic stimuli that preserve their ecological validity under manipulation.

We will now aim to describe qualitatively the consequences of conditional support preservation during manipulation. An illustrative example from this study is shown in the top right and bottom right corners of Figure 2. It demonstrates qualitative differences in manipulation approaches for enhancing attractiveness, with additional example manipulations across various traits and demographic groups provided in the Supplementary Materials. The mentioned older male face exemplifies the complexity of the optimal manipulation path, as effective attractiveness manipulation for this image necessitates identifying features uniquely linked to variations in attractiveness ratings within the specific context of male faces of this age group. Without explicit control, vector-based manipulation priorities decreasing perceived age to resemble a demographic stratum of faces that are generally considered attractive, resulting in a loss of internal validity as the manipulation predominantly becomes an age-decreasing one (which can be seen in Figure 8 in the Supplementary Materials). In contrast, our flow-matching approach maintains focus on attractiveness-specific features while preserving age-related characteristics, thus illustrating the fundamental differences in how each method approaches the manipulation challenge. Understanding the high constraints imposed by correct manipulation with controlled factors illustrates the somewhat counterintuitive point that the type of control achieved by the full flow-based method enables the construction of fully orthogonal experimental designs for all manipulated and controlled variables, allowing researchers to disentangle and investigate complex interactions between factors at all levels with unprecedented precision.

The vector-based approach also demonstrates systematic limitations in feature selection and support preservation, which become apparent through an examination of the manipulated stimuli. The linear approach identified tanned skin as a feature that is, on average, positively associated with increased attractiveness judgments. However, when the initial face already exhibits a disproportionately flushed or reddened skin tone, further increases in this feature yield inverse results, creating an unnatural and unattractive appearance (see also Figure 9 in the Supplementary Materials). This limitation stems from the general inability of linear and gradient-based approaches to recognize when a particular face lies at the extreme end of the realistic range of values for a given feature. This prevents these methods from predicting that further increases will result in a sharp drop in actual ratings. The optimal manipulation should account for these distributional boundaries and select alternative features that would also increase predicted ratings while remaining within the feature distribution where those predictions are valid. A similar case can be seen in the female face in the middle row in Figure 2 (when trustworthiness is manipulated). This face begins in an extremely rotated pose and requires manipulation that is specifically calibrated to the feature distribution. The linear manipulation does not attempt to rotate the face further; consequently, it suffers from the dual limitation of being unable to learn different manipulations for different faces, and of failing to account for distributional constraints that would ensure realistic and effective trait enhancement (see also the similar case for manipulating perceived smartness in the Supplementary Materials). This highlights the superior adaptability of the flow-matching approach in handling diverse facial configurations and in maintaining ecological validity across varied starting conditions.

This ability to accurately control confounds *during* a manipulation has a significant impact on the experimental designs possible with the method. While stimuli databases broadly enable only statistical control of the means of confounds within experimental conditions, the ability to leave confounds unaffected during manipulations enables the construction of arbitrarily detailed, fully orthogonal designs that can enable the study of the disentangled direct effects and multivariate interaction effects of naturally highly correlated factors with unprecedented precision. The high correlations between impressions preclude the construction of such complex designs by matching stimuli from even extensive databases. An important research direction is thus further development and testing of such model-based controlled manipulations e.g. to disentangle emotional dimensions in the prosody of speech and factors like race in face stimuli. To aid the latter, in the limitations section, we show how the MCI (the Classification Image of the impression as captured

by a given impression model, derived in Appendix C and D) can be used to uncover potential biases captured in the models (the removal of which is necessary before introducing race as an explicit controlled factor).

5 General discussion

The presented framework describes manipulation in feature spaces on three levels of complexity: as manipulation directions, paths, and transport of feature distributions. We identified these levels as corresponding formally to vector fields, integrating curves, and probability flows, which enables the application of modern deep learning estimation methods for flows and gradients. The methods developed in the article present the first publicly available method for constructing tightly controlled face stimuli for dimensions of the One Million Impressions dataset (Peterson et al., 2022), which, thanks to their ecological validity, can be used across areas of experimental psychology. Moreover, the formal groundwork laid in the article—connecting principles of experimental manipulation with modern generative models—is sufficiently general to extend reverse correlation beyond its current application domains. We have termed this formal framework, which utilizes flows, gradient fields, and continuous feature spaces to achieve valid, specific, ecologically valid, and orthogonally controlled manipulations, the Optimal Experimental Manipulation of Stimuli (OEMS). We will now map how the framework translates to the other (digitally representable) stimuli.

For other classes of image stimuli, the formal results translate directly, as StyleGAN-architecture models are available for abstract and portrait art, textures, animals, medical scans, cars, buildings, and interiors . For voice stimuli, OEMS can also be extended due to a fascinating link: successful controllable voice generators are based on combined image and flow methods. Voice samples are first transformed into a 2D image of their spectral frequencies and controlled via a flow-based vector model Eskimez et al. (2024), which has recently spurred work on manipulating psychological emotion dimensions (Cho et al., 2024). For video stimuli, work has begun on extending reverse correlation of psychologically relevant cues to facial expressions (Yan et al., 2023). The current fast development of new video generators may soon enable more general applications, the work on OEMS for which may draw from the neuroscience literature on temporal CIs via Spike-Triggered Averages (Sandler & Marmarelis, 2015). For text stimuli, our work was partially inspired by an existing method that utilizes word representations in selecting word stimuli for experimental conditions for studies manipulating emotional load, called Stimuli Descent (Plisiecki & Sobieszek, 2024). As distance in word representations corresponds to semantic similarity, the algorithm finds semantically matching words (specificity) via a gradient flow manipulation path of an emotion prediction model (validity), resulting in word stimuli for different experimental conditions (e.g., increasing in positivity, skeleton-fossil-wishbone, crass-cheeky-whimsical). The added benefits of flow-based paths translate directly to this setting, which, with the ever-rapid development of language models, could enable such precise text stimuli selection for more complex texts and dimensions. The key point for the methodologist is that these extensions do not require psychologists to train their own generative models, but take advantage of the parallel developments in AI of ever more accurate generative models to control the psychologically relevant aspects of stimuli via methods that extend naturally from the linear regression principles used in previous reverse correlation works.

By providing a unified theory, OEMS enables the transfer of methodological insights across these diverse areas of methods in AI, psychology, and neuroscience. The multidisciplinary nature of this endeavor stems from it not fitting uniquely into one methodological area, relying on and working in conjunction with methods of:

1. *Psychometrics* underlying the measurement of our dependent variables (the various impressions y).
2. *Statistical modeling* of impressions requiring proper validation procedures and amenable feature representations—the direct correspondence of which to stimulus factors of variation, is the subject of *disentangled representation learning* (Wang et al., 2024).
3. *Artificial Intelligence* enabling the modern generative models capable of learning well-conditioned feature spaces and generating realistic new stimuli.
4. *Causal inference* provides the necessary language of (controlled) interventions, operationalizing 'a manipulation to increase perceived trustworthiness holding age constant,' however, as the high dimensional continuous representations of interest in OEMS cannot be factorized further into individual variables, the problem quickly escapes classical theory, drawing more from developments in *heterogeneous treatment effects* estimation and its facilitation with generative models, recent work on which began in econometrics (Athey et al., 2024).
5. Mathematics of *Optimal Transport*—which solves minimal difference matching problems between two samples or distributions and provides mathematical guarantees for Conditional Support Preservation, as well as *flows* and *differential equations*—which capture how the stimulus is to change under a manipulation when modeling more accurate, heterogeneous treatment effects.
6. Finally, *Optimal design*—the complementary area of methodology concerning what happens right after and before the concerns of OEMS. Specifically,

as optimal design informs how to maximize statistical efficiency in the design of experiments, it can aid selection of candidate stimuli for a norm database creation studies—e.g., the important results by Lammert et al. (2022), showing how simply removing incidental feature correlations in the dataset can increase the power of reverse correlation estimates—or inform alternative procedures (e.g., the multiple choice CI of Schmitz et al., 2024). When deploying OEMS-generated stimuli, optimal design guides level spacing, factorial structure, and allocation to maximize efficiency for the substantive hypothesis.

A unique contribution of OEMS is in facilitating *fully orthogonal experimental designs* with ecologically valid digital stimuli. The novel *Conditional Support Preservation (CSP)* condition paves a way for achieving fully orthogonal manipulations by considering stimuli with the given combination of evoked impressions as a probability distribution in feature space. Because generative models have been trained on a vast amount of data, the prior information contained in how the stimuli are distributed in feature spaces acts as a statistical power multiplier, allowing the extraction of significantly more information than classical reverse correlation would assume under a multiple regression analysis. Being able to sample from the same sampling distribution that was used to generate the norms database enables overcoming the missing-not-at-random problem, in which predictive models trained on norm databases have no guarantees of generalizing to new stimuli of the same kind. The generalizability of impression model predictions afforded by knowledge of generative models' sampling distributions and application of formal tools like optimal transport and flow matching can lead researchers to disentangle highly correlated psychological factors with precise control over confounds. This would, in turn, enable examination of complex, multivariate interactions that were previously impractical: e.g., test how perceived trustworthiness modulates economic decisions and whether these effects vary systematically across age, gender, and race within a factorial design that preserves each factor at the level of the stimuli. Intersectional questions—such as whether competence stereotypes operate differently for older women than for younger men, holding age and gender constant while manipulating competence-related cues—become testable with precision.

More broadly, OEMS and neural network models of impressions can enable theory-forward tests of independence versus interaction among social-perception dimensions, with accurate dose–response functions being available at the experiment design stage of stimuli rather than inferred post hoc. Especially consequential is the observed efficiency on valence-laden impressions (trustworthiness, attractiveness) that have resisted reliable *increases* with earlier methods and that can confound investigations of other constructs (see Sobieszek et al., 2024). The result is not only stronger internal validity and clearer causal attribution, but also the potential for application to more complex psychological constructs in more challenging stimulus representations. Combining the endless supply of novel stimuli from generative models with the precise control afforded by OEMS can lead to increased reproducibility, validity, and scope of experimental research in psychology.

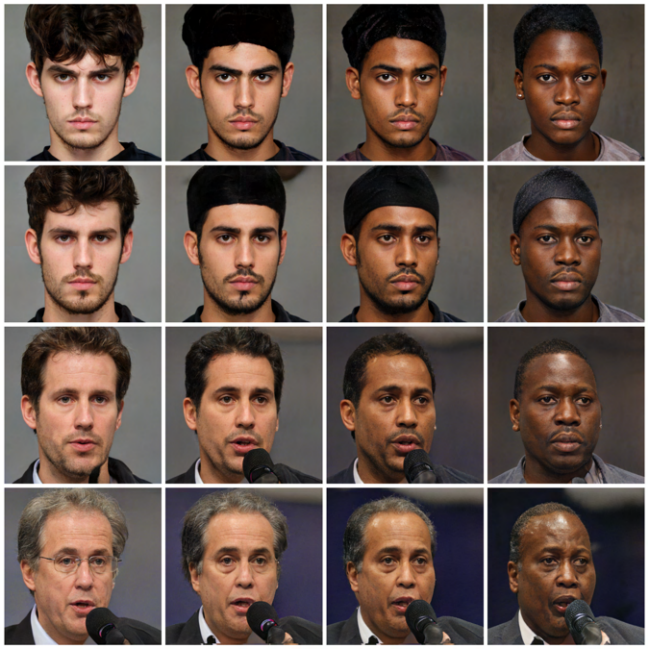


Figure 13: The Model Classification Images visualizing the stereotype of negative trustworthiness within 4x4 levels of the ratings of age and black within male faces.

6 Conclusion

The present work aimed to bridge experimental methodology in psychology and the rapidly growing field of generative models by building on and systematizing classical reverse correlation methods for the modern era of AI-generated stimuli. We established formal foundations for Optimal Experimental Manipulation of Stimuli, providing a unified framework for visualization and manipulation of reverse correlation approaches grounded in the geometric properties of feature spaces and the formalism of a psychometric gradient, as well as the probabilistic consideration of feature distributions and conditional support preservation. The proposed Psychometric Gradient Flow Matching algorithm connects for the first time problems of experimental manipulation methodology with tools of optimal transport theory and flow matching to solve long-standing challenges in reverse correlation-based manipulation, especially those involving the ma-

nipulation of valence-laden impressions and the control of confounding factors. Empirical validations across four studies ($N_{total} = 1026$) showed how the strengths and limitations of linear-, gradient-, and flow-based manipulation approaches, with our PGFM algorithm uniquely able to manipulate traits such as trustworthiness and attractiveness while maintaining control over confounding variables (e.g., age and gender). This represents a significant methodological advancement, enabling researchers to create experimental manipulations with unprecedented control and ecological validity.

We have shown how neural network prediction models can significantly increase the accuracy of impression models. In this general non-linear setting, we have explored the critical consideration of misalignment between model accuracy and manipulation efficiency metrics, causing the validation set model accuracy to not accurately reflect the obtained manipulation validation results. We have provided an explanation for, a measure of, and a solution to this issue, in the form of, respectively, the Missing-not-at-random problem, the expected risk along manipulation paths, and the conditional support preservation condition satisfied by flow-based methods (which we empirically found keeps the risk of model error constant across manipulation strengths). While this increases confidence in the applicability of the described techniques to experimental practice, validation checks or smaller validation studies are still necessary when introducing new manipulations for specific research questions.

The continued work on methodologically rigorous incorporation of Artificial Intelligence models into experimental practice opens new possibilities for research in experimental psychology, while also ensuring that psychological science stays ahead in studying the potential risks and societal harm these technologies may pose. By equipping researchers with open-source tools and specific models to generate precisely controlled stimuli, we aimed to facilitate this research program and enable the use of these methods across experimental psychology.

References

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Vision Research*, *42*(27), 2935–2951. https://doi.org/10.1016/S0042-6989(02)00397-9

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection [Introduces response-weighted noise averaging to reveal perceptual templates.]. *Proceedings of the IEEE*, PN–7.

Albohn, D. N., Martinez, J. E., & Todorov, A. (2024). Determinants of shared and idiosyncratic contributions to judgments of faces [Epub 2024 Sep 19]. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(11), 1117–1130. https://doi.org/10.1037/xhp0001239

Albohn, D. N., Uddenberg, S., & Todorov, A. (2022). A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Frontiers in Psychology*, *13*, 997498. https://doi.org/10.3389/fpsyg.2022.997498

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The apa publications and communications board task force report. *American Psychologist*, *73*(1), 3.

Athey, S., Imbens, G. W., Metzger, J., & Munro, E. (2024). Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, *240*(2), 105076.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 187–194. https://doi.org/10.1145/311535.311556

Bradley, M. M., & Lang, P. J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. https://api.semanticscholar.org/CorpusID:145474983

Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*(1), 333–361. https://doi.org/10.1080/10463283.2017.1381469

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, *12*(2), 199–213. https://doi.org/10.1088/0954-898X_12_2_301

Cho, D.-H., Oh, H.-S., Kim, S.-B., Lee, S.-H., & Lee, S.-W. (2024). Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *Interspeech 2024*, 1810–1814. https://doi.org/10.21437/interspeech.2024-398

Choi, J., Hwang, G., Cho, H., & Kang, M. (2023). Analyzing the latent space of gan through local dimension estimation. https://arxiv.org/abs/2205.13182

Chung, Y., & Jeglic, E. L. (2017). Detecting suicide risk among college students: A test of the predictive validity of the modified emotional stroop task. *Suicide and Life-Threatening Behavior*, *47*(4), 398–409. https://doi.org/https://doi.org/10.1111/sltb.12287

Crookes, K., Ewing, L., Gildenhuys, J.-d., Kloth, N., Hayward, W. G., Oxner, M., et al. (2015). How well do computer-generated faces tap face expertise? *PLoS ONE*, *10*(11), e0141353. https://doi.org/10.1371/journal.pone.0141353

Derous, E., Pepermans, R., & Ryan, A. M. (2017). Ethnic discrimination during résumé screening: Interactive effects of applicants' ethnic salience with job context. *Human Relations*, *70*(7), 860–882. https://doi.org/10.1177/0018726716676537

Dotsch, R., Wigboldus, D. H. J., Van Knippenberg, A., et al. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. https://doi.org/10.1111/j.1467-9280.2008.02186.x

Eckstein, M. P., & Ahumada, A. J. (2002). Classification images: A tool to analyze visual strategies. *Journal of Vision*, *2*(1), i. https://doi.org/10.1167/2.1.i

Eskimez, S. E., Wang, X., Thakker, M., Li, C., Tsai, C.-H., Xiao, Z., Yang, H., Zhu, Z., Tang, M., Tan, X., et al. (2024). E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. *2024 IEEE Spoken Language Technology Workshop (SLT)*, 682–689.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271. https://doi.org/10.1016/S0042-6989(01)00097-9

Hehman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. A. (2018, February). Assessing the point at which averages are stable: A tool illustrated in the context of person perception. https://doi.org/10.31234/osf.io/2n6jq

Jaeger, B. (2021). Consensual and idiosyncratic trustworthiness perceptions independently influence social decision-making. *European Journal of Social Psychology*, *51*(7), 1172–1180. https://doi.org/https://doi.org/10.1002/ejsp.2806

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lammert, A. C., Parrell, B., & Roop, B. W. (2022). Stimulus whitening improves the efficiency of reverse correlation. *Behavior research methods*, *54*(6), 2969–2978.

Lassalle, A., Pigat, D., O'Reilly, H., Berggren, S., Fridenson-Hayo, S., Tal, S., Elfström, S., Råde, A., Golan, O., Bölte, S., Baron-Cohen, S., & Lundqvist, D. (2019). The EU-Emotion Voice Database. *Behavior Research Methods*, *51*(2), 493–506. https://doi.org/10.3758/s13428-018-1048-1

Lavan, N., & Sutherland, C. A. (2024). Idiosyncratic and shared contributions shape impressions from voices and faces. *Cognition*, *251*, 105881. https://doi.org/https://doi.org/10.1016/j.cognition.2024.105881

Lipman, Y., Albergo, D., & Pacchiardi, L. e. a. (2023). Flow matching for generative modeling. *Transactions on Machine Learning Research*.

Ma, D. S., Kantner, J., & Wittenbrink, B. (2021). Chicago face database: Multiracial expansion. *Behavior Research Methods*, *53*(3), 1289–1300. https://doi.org/10.3758/s13428-020-01482-5

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). 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.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*(4), 1428–1444. https://doi.org/10.3758/s13428-019-01323-0

May, K. A., & Solomon, J. A. (2013). Four theorems on the psychometric function. *PLOS ONE*, *8*(10), 1–34. https://doi.org/10.1371/journal.pone.0074815

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Peebles, W., & Xie, S. (2023). Scalable diffusion models with transformers. https://arxiv.org/abs/2212.09748

Peszek, J., & Poyato, D. (2023). Heterogeneous gradient flows in the topology of fibered optimal transport. *Calculus of Variations and Partial Differential Equations*, *62*(9), 258.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*(17), e2115228119. https://doi.org/10.1073/pnas.2115228119

Plisiecki, H., & Sobieszek, A. (2024). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behavior Research Methods*, *56*(5), 4716–4731. https://doi.org/10.3758/s13428-023-02212-3

Ringach, D. L., Sapiro, G., & Shapley, R. (1997). A subspace reverse-correlation technique for the study of visual neurons. *Vision Research*, *37*(17), 2455–2464. https://doi.org/10.1016/S0042-6989(96)00313-X

Ringach, D. L., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166. https://doi.org/10.1207/s15516709cog2802_2

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. https://arxiv.org/abs/2112.10752

Samide, R., Cooper, R. A., & Ritchey, M. (2020). A database of news videos for investigating the dynamics of emotion and memory. *Behavior Research Methods*, *52*, 1469–1479. https://doi.org/10.3758/s13428-019-01327-w

Sandler, R. A., & Marmarelis, V. Z. (2015). Understanding spike-triggered covariance using wiener theory for receptive field identification. *Journal of Vision*, *15*(9), 16–16.

Schmitz, M., Rougier, M., & Yzerbyt, V. (2024). Introducing the brief reverse correlation: An improved tool to assess visual representations. *European Journal of Social Psychology*. https://doi.org/10.1002/ejsp.3100

Snefjella, B., & Blank, I. (2020). Semantic norm extrapolation is a missing data problem. https://doi.org/10.31234/osf.io/y2gav

Sobieszek, A., Siemiątkowski, M., & Imbir, K. K. (2024). Generative neural networks for experimental manipulation: Examining dominance-trustworthiness face impressions with data-efficient models. *British Journal of Psychology*, *115*, 1–20. https://doi.org/10.1111/bjop.12732

Summerell, E., Shuttleworth, L., Lin, C., & Denson, T. F. (2025). The effects of race, gender, and alcohol cues on anger perception in crowds. *Australian Journal of Psychology*, *77*(1). https://doi.org/10.1080/00049530.2024.2426661

Sutherland, C. A. M., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science*, *8*(5), 538–551. https://doi.org/10.1177/1948550617697176

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. https://doi.org/https://doi.org/10.1016/j.cognition.2012.12.001

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In B. Gawronski (Ed.), *Advances in experimental social psychology* (pp. 189–245). Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2020.11.004

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., & Bengio, Y. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport [arXiv:2302.00482v4]. *Transactions on Machine Learning Research*.

Villani, C. (2008). *Optimal transport: Old and new* (Vol. 338). Springer.

Vingilis-Jaremko, L., Kawakami, K., & Friesen, J. P. (2020). Other-groups bias effects: Recognizing majority and minority outgroup faces. *Social Psychological and Personality Science*, *11*(7), 908–916. https://doi.org/10.1177/1948550620919562

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745–10759. https://doi.org/10.1109/tpami.2023.3263585

Wang, X., Chen, H., Tang, S., Wu, Z., & Zhu, W. (2024). Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9677–9696. https://doi.org/10.1109/TPAMI.2024.3420937

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, *63*(8), 1293–1313.

Yan, S., Soladié, C., Aucouturier, J., & Seguier, R. (2023). Combining GAN with reverse correlation to construct personalized facial expressions. *PLOS ONE*, *18*(8), e0290612. https://doi.org/10.1371/journal.pone.0290612

Zhukova, M. A., Ovchinnikova, I. V., An, I., & Grigorenko, E. L. (2021). Altered neural processing of emotional words in adults with a history of institutionalization: Evidence from the emotional stroop task. *European Journal of Neuroscience*, *53*(5), 1487–1497. https://doi.org/https://doi.org/10.1111/ejn.15015

A Generative modeling

A.1 The stimulus space as a differentiable manifold

To define a psychometric gradient, we first require a well-behaved space. While a visual stimulus $\mathbf{s}$ can be represented as a vector of raw pixel values in $\mathbb{R}^{d_{\text{pix}}}$, this space is ill-suited for our purpose. Most points in pixel space do not correspond to valid stimuli (e.g., natural faces), and its Euclidean distance is a poor proxy for perceptual similarity.

However, a well-founded proposition of the generative modeling literature is the **manifold hypothesis**: the set of stimuli sampled from a perceptually consistent distribution $p(s)$, forms a lower-dimensional, non-linear manifold $\mathcal{S}$ embedded within the high-dimensional pixel space. A manifold is the specific mathematical object one needs to define a space on which gradients/integrals are well defined and such an object is equipped with an intrinsic measure of distance between two stimuli, that is not a straight line in the larger pixel space, but the length of the shortest path between them that stays entirely on the manifold—a path known as a **geodesic**. This corresponds to the most direct, perceptually smooth "perfect morph" between the two stimuli.

To perform calculus, we need a coordinate system for this manifold. Modern deep generative models provide a powerful solution by learning a mapping G from a low-dimensional latent *feature space* $\mathbf{x} \in \mathbb{R}^d$ to the stimulus manifold $\mathcal{S}$.

$$\underbrace{\mathcal{X}}_{\substack{\text{feature}\\ \text{space}}} \subset \mathbb{R}^d \xrightarrow{G} \underbrace{\mathcal{S}}_{\substack{\text{valid-stimuli}\\ \text{manifold}}} \subset \underbrace{\mathbb{R}^{d_{\text{pix}}}}_{\text{raw pixels}}$$

Definition A.1 (Well-Conditioned Feature Space)**.** A feature space $\mathcal{X}$ of generator G is **well-conditioned** if it satisfies:

1. **Differentiability:** The generator G is continuously differentiable (C^1), making its Jacobian J_G well-defined.
2. **Approximate Isometry:** The mapping G locally preserves approximate distances and angles, which is achieved when J_G can be represented with a diagonal matrix with equal entries (J_G is diagonalizable, and the diagonal entries approximate the metric scaling c). This property enables the flat Euclidean geometry of $\mathcal{X}$ to faithfully represent distances in the curved geometry of $\mathcal{S}$.
3. **Convex Support of Valid Features:** G maps its feature distribution $p_{\text{feature}}(x)$ to $p_{\text{valid}}(s)$ and (at least locally) the support of this feature distribution is (segment-)convex:

$$x_0 + t(x_1 - x_0) \in \text{supp}(p_{\text{feature}}) \quad \forall t \in [0, 1].$$

Hence, linear interpolation paths between valid feature points remain within the region of ecologically valid stimuli.

Of particular interest is an additional objective for GAN training called **Path Length Regularization** (PLR). The goal of PLR is to encourage the feature space $\mathcal{X}$ to be a length-preserving coordinate space for the stimulus manifold $\mathcal{S}$.

$$\mathcal{L}_{PLR} = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{N}(0,I)} \left[\left(\|J_G(\mathbf{x})^T \mathbf{y}\|_2 - c \right)^2 \right] \tag{1}$$

where $J_G(\mathbf{x})$ is the Jacobian of the generator w.r.t. the feature space, $\mathbf{y}$ is a random image-space vector, and a is a dynamically adjusting constant representing the running average of the norm. This loss penalizes deviations from a fixed-magnitude response, effectively encouraging the Jacobian to be orthogonal.

In their work introducing StyleGAN2, Karras et al. (2020) provided an analytical basis for this regularizer. They showed that encouraging the Jacobian to be orthogonal via PLR results in a smoother and more disentangled feature space. A key consequence is that a step of a given length in any direction in $\mathcal{X}$ tends to correspond to a step of a fixed perceptual magnitude in $\mathcal{S}$. They also provide measures for testing whether the model achieves the probabilistic and geometric modeling goals as defined in Section 2.1, which are the Frechet Inception Distance (for ecologically valid distribution matching) and Perceptual Path Length (for feature linearity and isometry; see Choi et al., 2023 for a geometric analysis of StyleGAN2 satisfying the manifold hypothesis).

A.2 Agreement of Morphing and generative approaches

Proposition A.2 (Shown in Karras et al., 2020)**.** *When training a well-conditioned generator G, the model that minimizes $\mathcal{L}_{PLR}$ minimizes the Perceptual Path Length, that is, a linear interpolation path $\mathbf{x}_1 \to \mathbf{x}_2$ in the feature space $\mathcal{X}$ approximates the perfect morph (geodesic) path $\mathbf{s}_1 \xrightarrow{morph} \mathbf{s}_2$ on the stimulus manifold $\mathcal{S}$.*

A direct corollary of the above result is the distance-difference correspondence:

Corollary A.3. *Given how the stimulus perfectly in the middle along a perfect morph $\mathbf{s}_1 \xrightarrow{morph} \mathbf{s}_2$, is as similar to $\mathbf{s}_1$ as it is to s_2, if $G(x)$ has a well-conditioned feature space:*

$$\boxed{dist(x_1, x_2) \approx diff(s_1, s_2)}$$

Consequently, the x-space averaging of a a group of stimuli $\bar{\mathbf{x}}$ corresponds to finding a prototype stimulus $\mathbf{s} = G(\bar{\mathbf{x}})$ minimizing perceptual difference to the associated stimuli (equivalently to s-space averaging in the morphing method-of-constant-stimuli of Sutherland et al., 2017, 2013)

Graceful failure of assumptions An important note is that the strict mathematical conditions above are only approximately satisfied in practice, the methods described in the manuscript are very robust to their violations. If some direction in features results in a larger visual change than others (violating isometry), the 511 other dimensions diffuse drastically its impact on distance calculations. Formally we say that in $\mathbb{R}^d$ when $d >> 3$ the approximate isometry of G can be well approximated by a bi-Lipshitz G, which yields upper and lower bounds for distance-difference scaling in each direction. As for vialoations of convex support, if some interpolation paths move outside of generator support one can limit pairs to within some distance in features, making the learned manipulation path curve around the the hole in valid stimuli.

B Reverse correlation revisited

B.1 The impression as differentiable potential function

In physical sciences a vector field influencing the observed motion in a space X can often be described as a gradient of a function $X \to \mathbb{R}$, which is then called a *potential*. We posit that—largely independent of the response scale—between-subject regularities in impressions, together with the manifold hypothesis for valid stimuli, imply a differentiable $\Psi : \mathcal{S} \to \mathbb{R}$ that maps each stimulus $s \in \mathcal{S}$ to a population-average scalar value, with a well-defined gradient $\nabla\Psi$, which by analogy we call the *psychometric potential* . The reverse correlation is sometimes described as giving access to some hidden mental representations, while still being based on equally noisy and confounded scale-based measurements of the reported impression. In the following sections we posit that if the paradigm offers any privileged access, it comes from the estimand of reverse correlation studies being the gradient of this latent impression potential, whose direction in feature space, averaging measurements across individuals, exhibits a degree of invariance to measurement scale. Let:

- The measured individual be $j \sim$ population
- The latent impression: ψ (scalar random variable)
- Observed output: y (rating/choice on quantitative scale)

The covariation of the rating, impression, stimulus and features for participant j can be factorized as:

$$p_j(y, \psi, s, x) = \underbrace{p_j(y \mid \psi)}_{\text{Impression measurement}} \underbrace{p_j(\psi \mid s)}_{\text{Impression formation}} \underbrace{p(s \mid x)}_{\text{Generator}} \underbrace{p(x)}_{\text{Stimulus variation}}$$

Definition B.1 (Individual and Population Mean Impressions)**.** The internal impression strength $p_j(\psi \mid s)$ that a given stimuli is likely to evoke in participant j is unknown and inferences about its magnitude could only be made in expectation, within-subject, across-stimuli assuming higher ratings are associated with a higher *expected impression magnitude*—we label this theoretical ground-truth as:

$$\bar{\psi}_j(s) \triangleq \mathbb{E}[\psi_j \mid s]$$

We will see this function is largely inestimable, as ratings of the same stimuli are not i.i.id under repeated measurements and depend on individual answering style. However under proper randomization of participants and stimuli, the probabilistic aspects of the impression forming and measurement processes are uncorrelated across different individuals. Therefore, an empirical estimate of $\bar{\Psi}(s)$ understood to be the population mean of $\bar{\psi}_j(s)$ in first impressions is possible. The estimator:

$$\boxed{\widehat{\Psi}(s) = \frac{1}{J} \sum_{j=1}^{J} Y_j^{(1)}(s)} \tag{2}$$

from single ratings from many participants $Y_j^{(1)}$ is unbiased for the population mean and consistent as $J \to \infty$ (for its empirical convergence rates for different face impressions see Hehman et al. 2018). We now turn to whether its estimand meaningfully captures the gradient of the impression despite being based on scale ratings.

We will first establish the existence of the gradient of the psychometric potential.

Proposition B.2 (Discriminability $\implies$ Differentiability)**.** *Along any morph $s_1 \to s_2$ within valid stimuli, perceptual difference vanishes and discrimination approaches chance (50%). Hence the expected ratings for the two stimuli converge, implying $|\psi(s_1) - \psi(s_2)| \to 0$ as diff$(s_1, s_2) \to 0$. Thus ψ is continuous on $\mathcal{S}$ (for $x \in \text{supp}(p_{feature})$).*

By Rademacher's theorem: if individual mean impressions are locally Lipschitz, i.e. the rate at which the impressions approach each other under smaller changes ς to a stimuli is bounded,

$$\frac{\left|\bar{\psi}_j(s) - \bar{\psi}_j(s+\varsigma)\right|}{d(s, s+\varsigma)} \leq C_j \cdot,$$

and $C_j < \infty$ for all stimuli than the gradient $\nabla\bar{\psi}_j$ exists almost everywhere on $\mathcal{S}$.

Proposition B.3 (Chain-rule for feature space potential)**.** *Let $f_y(x)$ be the model based on ratings for generated stimuli $G(x) = s$. The psychometric potential gradient direction in feature space is given by*

$$\nabla_x f_y(x) \rightrightarrows J_G(x)^\top \nabla_s \Psi\big(s\big).$$

Under near-isometry, the generator preserves angles between directions*, thus justifying how the estimation of the direction of $\nabla_x f_y$ that agrees with directions of $\nabla_s \Psi$ depends on a well-conditioned feature space. The derivations in the following sections will thus hold under this assumption, which formally states that under the "visual distance metric" of the generator f is a so-called "Isoparametric Function".*

B.2 Estimating the Psychometric Potential

Within a measurement scale and experimental procedure the (i)-th response to the same stimulus s, can be modeled as

$$Y_j^{(i)}(s) \;=\; g_j\big(\psi_j^{(i)}(s)\big) \;+\; b_j \;+\; \varepsilon_{j,s}^{(i)}.$$

Proposition B.4 (Individual impressions are inestimable)**.** *Because $\psi_j^{(i)}(s)$ and $\varepsilon_j^{(i)}$ can vary with (i) (e.g., habituation, memory, anchoring), for an individual j:*

$$\mathbb{E}\big[Y_j^{(1)}(s)\big] \;\neq\; \mathbb{E}\Big[\frac{1}{K}\sum_{i=1}^{K} Y_j^{(i)}(s)\Big].$$

Where the first term is the true expectation of the first impression only.

We estimate it from empirical ratings $Y_j^{(1)}(s)$ (first impressions) aggregated over observers j. As the natural estimator $\widehat{\Psi}(s) = \frac{1}{J}\sum_{j=1}^{J} Y_j^{(1)}(s)$ is unbiased, if the chosen measurement scale is dense enough for the central limit theorem to hold for the residual errors for j collected per-stimulus ratings, we can fit the differentiable $f_y(x)$ with unbiased gradient estimates within the training data support.

B.3 Measurement Invariance of the Psychometric Gradient

Psychometric data live on arbitrary response scales and come with idiosyncratic answering strategies. When comparing means of such responses, the non-invariance of the measurement between two individuals or groups seriously threatens validity. Here we show how our the local *direction* of impression increase in features is on average robust to the two most common nuisances—additive response bias and different shapes of the monotone 'impression-to-measurement scale' mappings.

Proposition B.5 (Invariance of the gradient direction to answering style)**.** *If each participant j responds based on their internal probabilistic impression $\psi_j(\mathbf{s})$, and individual impression-to-scale mapping g_j, bias b_j, and measurement noise $\varepsilon_{j,s}$ $Y_j = g_j(\psi_j(\mathbf{s})) + b_j + \varepsilon_{j,s}$. The gradient direction of the expected rating agrees with the direction of the latent psychometric gradient:*

$$\nabla\mathbb{E}[Y_j^{(1)}(s)] \doteq \nabla\bar{\psi}_j(\mathbf{s}) \quad \textit{independently of } g_j, b_j.$$

Proof. Assume two individuals have the exact same idiosyncratic impression formation response $\psi_1 = \psi_2$ but differ in g_j, bias b_j and noise $\varepsilon_{j,s}$ (complete non-invariance). We can show that the expectation of measured responses follows:

$$a_2\nabla\mathbb{E}[g_1(\psi(\mathbf{s}))] = \; a_1\nabla\mathbb{E}[g_2(\psi(\mathbf{s}))].$$

Using our generic measurement model, the expectation for a stimulus $\mathbf{s}$ is a

$$\mathbb{E}[Y_j(s)] = \mathbb{E}[g_j(\psi_j(\mathbf{s})) + b_j + \varepsilon_j]$$

To find the expected rating, we average over many hypothetical repetitions of this rating process. $\mathbb{E}[\varepsilon_{j,s}] = 0$ averages out to zero, while taking the gradient of both sides

$$\nabla\mathbb{E}[Y_j(s)] = \nabla\left(\mathbb{E}[g_j(\psi_j(\mathbf{s})) + b_j]\right)$$

any constant b_j vanishes under differentiation. The gradient and expectation are linear operators. By the dominated convergence theorem, they can be interchanged. Applying the the chain rule for vector valued functions:

$$\nabla\mathbb{E}[Y_j(s)] = (\mathbb{E}[\nabla g_j(\psi_j(\mathbf{s}))]) = \mathbb{E}[g_j'(\psi_j(\mathbf{s})) \cdot \nabla\psi_j(\mathbf{s})]$$

Here, g_j' is the derivative of the participant's response function. Since g_j is increasing $g_j'(\psi_j(\mathbf{s}))$ is a positive value describing how sensitive the participant's rating scale is around an impression level. Its expectation for the given stimulus is a positive constant, which we label a_j:

$$\nabla\mathbb{E}[Y_j(s)] = a_j \cdot \mathbb{E}[\nabla\psi_j(\mathbf{s})] = a_j \cdot \nabla\bar{\psi}_j(\mathbf{s})$$

Which shows both that the gradient of a participant's expected rating, $\nabla\mathbb{E}[Y_j(s)]$ points in the same direction as the gradient of their underlying latent impression $\nabla\bar{\psi}_j(\mathbf{s})$, and that under dense enough sampling and enough power this result is measurement invariant comparing individuals differing in b_j, g_j and noise. □

While appealing, this hypothetical individual expectation invariance does not factor into practice until an actually estimable quantity is considered. Thus, the invariance has to be shown to positively affect also the valid estimator of the population average.

Theorem B.6 (Gradient of the Population Mean is the Mean of Idiosyncratic Gradients)**.** *The gradient of the empirical population mean estimator is in equal in expectation to the mean of individual impression gradients.*

$$\nabla\mathbb{E}[\widehat{\Psi}(\mathbf{s})] = \mathbb{E}_{j\sim population}[\nabla\bar{\psi}_j(\mathbf{s})] \tag{3}$$

Proof. The definition of the common first impression estimator was:

$$\nabla\mathbb{E}[\widehat{\Psi}(\mathbf{s})] = \mathbb{E}_j[\,Y_j^{(1)}(s)]$$

By the law of total expectation and proposition B.5:

$$= \mathbb{E}_j[\,\mathbb{E}_{\text{repeated meas.}}[Y_j^{(1)}(s)]] = \mathbb{E}_j[a_j\bar{\psi}_j(\mathbf{s}) + b_j]$$

By the dominated convergence theorem, gradient again can be interchanged:

$$\nabla \mathbb{E}_j[a_j \bar{\psi}_j(\mathbf{s}) + b_j] = \mathbb{E}_j[a_j \nabla \bar{\psi}_j(\mathbf{s})]$$

Then the population-level psychometric gradient satisfies

$$\nabla \bar{\Psi}(\mathbf{s}) \; = \mathbb{E}_j[a_j \nabla \bar{\psi}_j(\mathbf{s})] \quad \text{and is independent of } b_j.$$

Finally we can observe, that because we are concerned with effective manipulation directions (regardless of magnitude), instead of the significance tests of differences in the magnitude of raw ratings, the population average is also robust under differences in a_j, which is a positive, bounded value. □

Taken together, the results make a strong analytical case why even if individual idiosyncratic impressions are difficult to measure, the average of those gradients is captured by the gradient of the easily-estimable population mean function, $\bar{\Psi}(\mathbf{s})$ and is stable under the common forms of nuisance variation that adversely effect statistical tests using the mean impression estimate itself. As a consequence a precise enough estimate of $\nabla f_y(x)$ moves individual impressions correspondingly to the gradient of the mean latent average impression itself $\bar{\Psi}(\mathbf{s})$. We will thus use $\bar{\Psi}(\mathbf{s})$ in the following sections as the ground truth population impression and examine how different reverse correlation estimators relate to its gradient. The positive weighting factor a_j may at smaller sample sizes increase sensitivity to answers of a particular individual, but estimated gradient directions converge closer to a true population average with sample size (due to the small effect sizes of any hidden causal factors correlating idiosyncrasies in how s is perceived with specific answering styles).

Note, however, these invariance results do include invariance to differences in how the measured construct is understood by raters. This subject remains an important complementary area of research, which may itself obtain valuable information from the described possibility of translating self-reports of levels of a given psychological construct into an actionable manipulation plan, whose downstream causal effects may be measured and compared.

C Average- and Model-based Estimands of Reverse Correlation

Definition C.1 (Modes of Reverse Correlation)**.** We distinguish between two primary modes of reverse correlation based on their methodological approach:

- **Choice-Based Averaging Methods** aim to visualize the prototype stimulus associated with an impression of as a response-weighted average of stimulus perturbations. The result is a Classification Image (CI) (Dotsch et al., 2008; Ringach & Shapley, 2004), as we discussed this same estimation target can be used for manipulation with the morphing method-of-constant-stimuli (Sutherland et al., 2017).
- **Model-Based Methods** aim to find the manipulation vector by fitting a linear model of ratings from features (Oosterhof & Todorov, 2008).

C.1 Model-Based Estimation via Ordinary Least Squares

Definition C.2 (The OLS Direction Estimator)**.** We model the effect of features $x^{(i)}$ on the stimuli mean rating $y^{(i)}$ with the linear model $Y = X\beta + c + \epsilon_i$. For a finite dataset, the Ordinary Least Squares (OLS) estimator is:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \tag{4}$$

And is not biased to align with feature covariance $\Sigma_{\mathbf{x}}$, but, as we'll see, is influenced by $\Sigma_{\mathbf{x}}$ to capture gradient directions that lay within the distribution support.

In the **model-based paradigm** we generate stimuli by sampling points $x \in \mathbb{R}^d$ in feature space according to the stimuli feature distribution $\mathcal{N}(\mathbf{x}_0, \Sigma_{\mathbf{x}})$. We collect enough scalar impression ratings y per stimuli to have normal residuals ϵ and fit $Y = \bar{\Psi}(X) + \varepsilon$, where $\varepsilon \perp X$ with $\mathbb{E}[\varepsilon|X] = 0$. We now describe what vector $\beta^\star = \mathbb{E}[\hat{\beta}_{\text{OLS}}]$ the estimator targets in expectation in relation to the posed psychometric potential $\bar{\Psi}$.

OLS Estimates the Expected Psychometric Gradient The *population* OLS coefficient solves

$$\beta^\star = \operatorname*{argmin}_{\beta} \mathbb{E}\big[(\bar{\Psi}(\mathbf{x}) - \beta^\top \mathbf{x})^2\big].$$

Differentiating and using full-rank $\Sigma_{\mathbf{x}}$ gives the normal equation $\Sigma_{\mathbf{x}}\beta^\star = \mathbb{E}[\mathbf{x}\bar{\Psi}(\mathbf{x})]$. By Stein's identity for Gaussian x $\mathbb{E}[\mathbf{x}\bar{\Psi}(\mathbf{x})] = \Sigma_{\mathbf{x}}\mathbb{E}[\nabla\bar{\Psi}(\mathbf{x})]$, so

$$\boxed{\beta^\star = \mathbb{E}_{\mathbf{x}}[\nabla \bar{\Psi}(\mathbf{x})]} \qquad \text{(Average gradient)}$$

Where the averaging is the *expected gradient of* $\bar{\Psi}$ *over the natural feature distribution* $\mathbf{x} \sim \mathcal{N}(\mathbf{x}_0, \Sigma_{\mathbf{x}})$.

OLS Estimates the Gradient after Natural Covariation Smoothing A canonical way to define smoothing of a function is to define a Gaussian kernel $\varphi_{\Sigma_{\mathbf{x}}}(\mathbf{v}) = \text{pdf}_{\mathcal{N}(\mathbf{0},\Sigma_{\mathbf{x}})}(\mathbf{v})$, for which a spatially smoothed form of $\bar{\Psi}$ is defined as a convolution:

$$(\varphi_{\Sigma_{\mathbf{x}}} * \bar{\Psi})(\mathbf{x}) = \int \bar{\Psi}(\mathbf{x} + \mathbf{v})\, \varphi_{\Sigma_{\mathbf{x}}}(\mathbf{v})\, d\mathbf{v}$$

Which can be understood as replacing each output with a spatial weighted average:

$$(\varphi_{\Sigma_{\mathbf{x}}} * \bar{\Psi})(\mathbf{x}_0) = \mathbb{E}_{\mathbf{v}\sim\mathcal{N}(0,\Sigma_{\mathbf{x}})}[\bar{\Psi}(\mathbf{x}_0 + \mathbf{v})].$$

Differentiating under the integral (permissible by dominated convergence) yields $\nabla(\varphi_{\Sigma_\mathbf{x}} * \bar{\Psi})(\mathbf{x}_0) = \beta^\star$. Hence

$$\boxed{\beta^\star = \nabla\big(\varphi_{\Sigma_\mathbf{x}} * \bar{\Psi}\big)(\mathbf{x}_0)} \quad \text{(Gradient of smoothed } \bar{\Psi}\text{)}$$

The causal effect interpretation of the OLS-obtained coefficients given its $\Sigma_\mathbf{x}$ weighting of local gradients is that the implied effect of each feature is a mixture of local effects, marginalized over the natural covariation of all other features. Thus entries of the obtained β^* faithfully capture causal effects if either (a) no interactions exists between the magnitude of other features and the size of the effect the given feature or (b) the scope of the feature distribution is so local, as to not capture significant interactions within its support (formally this corresponds to regions in which the probability measure curvature is negligible).

C.2 Choice-Based Averaging Estimators

We now turn to **choice-based averaging**, which is the original reverse correlation formulation. Presenting paired stimuli perturbed with $\pm\mathbf{v}$, the observer chooses the one evoking a greater impression; the perturbation direction chosen as more positive is recorded as $c\mathbf{v}$ with $c \in \{-1, +1\}$.

Assume perturbations $\mathbf{v} \sim \mathcal{N}(0, \sigma^2 I)$ in stimuli space, with base stimulus $G(\mathbf{x}_0) = \mathbf{s}_0$. Since valid stimuli p_{valid} lie on a much lower-dimensional curved manifold, in high-resolution stimuli spaces perturbations $\mathbf{s}_0 + \mathbf{v}$ typically fall off-manifold. Thus in noise-based 2 Alternatives Forced Choice (2AFC) protocols, non-random contribution to choices c are driven by similarity to the nearest valid stimulus $\mathbf{s}'$. Following Gosselin and Schyns (2001), visual discrimination attenuates to specific (low/high spatial frequency) features (a lower-dimensional subspace of x), allowing participants to extract signal from noise beyond the immediate neighborhood of $\mathbf{x}_0$. While a well-conditioned feature space x captures the covariance $\Sigma_\mathbf{x}$ of the attended features, CIs constructed in noise-based reverse correlation miss those unattenuated features. Noting that, we model the choice-based averaging modeling perturbation with the full feature covariance (corresponding to the full CI from method-of-constant-stimuli or choice-based averaging in features).

Definition C.3 (The CI direction vector)**.** Let $\mathbf{x}_0$ be the base stimulus, $\mathbf{v}$ be centered feature perturbations with covariance $\Sigma_\mathbf{x}$, and $c_j = \pm 1$ be the participant j response to the stimuli pair $\mathbf{x}_0 \pm \vec{v}$.

The **Classification Image (CI) direction** in feature space will be the response-weighted average of stimulus perturbations:

$$\mathbf{v}_{\text{CI}} \triangleq \mathbb{E}_{j,\,\mathbf{v}}[c_j \cdot \mathbf{v}] \tag{5}$$

Estimated from N trials as $\mathbf{v}_{\text{CI}} = \frac{1}{N}\sum_{i=1}^{N} c_i \mathbf{v}_i$.

For a given participant j, if a perturbation is along directions that do not covary with their impression (the central tendency of which we described as the idiosyncratic potential $\bar{\psi}_j(s)$), we expect the probability of having $c = \pm 1$ to be 50/50 (corresponding to $\bar{\psi}_j(s_1) = \bar{\psi}_j(s_2)$) and the probability increases if s_1 tends to elicit in the participant a stronger impression than s_2. The non-random contributions to choice probability thus come from the discrimination ability of an individual to differentiate their own impression levels.

To model discrimination ability from choices the psychometric literature uses psychometric (link) functions such as the probit, Weibull or logistic functions (e.g., Wichmann and Hill, 2001, c.f. May and Solomon, 2013).

A shared characteristics of these models are that (a) the estimated psychometric link for participant j is a strictly increasing function $g_j : \mathbb{R} \to [0, 1]$ that converts the discriminated signal difference Δ, to choice probability $P_>(\cdot)_j$, (b) g is linear at the point of subjective equality $g(0) = 0.5$. This assumption has been justified in various ways as the activation of a linear neuronal receptive fields, the comparison of stimuli to a prototype, or the strong local linearity of the choice function (May & Solomon, 2013).

For participant j, the idiosyncratic *signal* that is being discriminated in the 2AFC reverse correlation task is thus well modeled as:

$$\Delta\psi_j(\mathbf{v}) \triangleq \mathbb{E}[\psi_j(\mathbf{s}_0+\mathbf{v}) - \psi_j(\mathbf{s}_0-\mathbf{v})] = \bar{\psi}_j(\mathbf{s}_0+\mathbf{v}) - \bar{\psi}_j(\mathbf{s}_0-\mathbf{v})$$

A *psychometric (link) function* for binary choice is a strictly increasing function $\mathcal{F}_j : \mathbb{R} \to [0, 1]$ that converts signal to choice probability:

$$p_j(\mathbf{v}) \triangleq \mathcal{F}_j\big(\Delta\psi_j(\mathbf{v})\big)\,.$$

The *population choice probability* is the expectation across participants,

$$P_>(\mathbf{v}) \triangleq \mathbb{E}_j[p_j(\mathbf{v})] = \mathbb{E}_j\Big[\frac{1}{2}\Big(c_j + 1\Big)\,\Big|\,\mathbf{v}\Big],$$

where the last equality states the probability of choice p_j in terms of its values $c_j = \pm 1$.

Theorem C.4 (Collinearity of the Classification Image and the Covariance-Aligned Gradient)**.** *The direction of the expected Classification Image vector, $\mathbf{v}_{CI}$, coincides with the transformation of the expected psychometric gradient $\nabla\bar{\Psi}$ by the feature covariance $\Sigma_\mathbf{x}$.*

$$\mathbf{v}_{CI} \rightrightarrows \Sigma_\mathbf{x}\beta^* \tag{6}$$

Where $\beta^* = \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{x}_0,\Sigma_\mathbf{x})}[\nabla\bar{\Psi}(\mathbf{x})]$.

Proof. By law of total expectation:

$$\begin{aligned}\mathbf{v}_{\text{CI}} = \mathbb{E}_{j,\mathbf{v}}[c_j \cdot \mathbf{v}] &= \mathbb{E}_\mathbf{v}\Big[\mathbb{E}_j[2\,p_j(\mathbf{v}) + 1] \cdot \mathbf{v}\Big] \\ &= \mathbb{E}_\mathbf{v}\Big[2\mathbb{E}_j[\mathbf{v} \cdot\, p_j(\mathbf{v})] + \mathbf{v}\Big]\end{aligned}$$

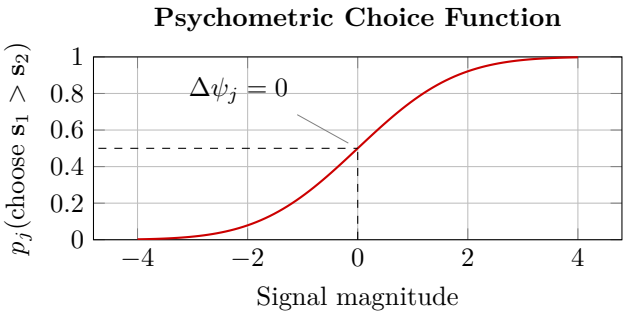


Figure 14: **Shape of a Psychometric Function of Choice Probability.** The nonlinear choice probability is modeled as depending on the discriminated signal strength, with $p_j = 0.5$ indicating the point of subjective equality.

For perturbation distributions that can be modeled as $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{x})$, the second term averages to 0 and to the first one we apply **Stein's Lemma**:

$$\mathbb{E}_{j,\mathbf{v}}[\mathbf{v} \cdot p_j(\mathbf{v})] = \Sigma_\mathbf{x} \mathbb{E}_{j,\mathbf{v}}[\nabla p_j(\mathbf{v})]$$

The psychometric function $\mathcal{F}$ is a monotone function of the potential difference, $p_j(\mathbf{v}) = \mathcal{F}\big(\Delta\psi_j(\mathbf{v})\big)$. By Proposition B.5, the gradient direction of the probability $\nabla p_j(\mathbf{v})$ is not influenced by the monotone link and collinear with:

$$\mathbb{E}_j[\nabla p_j(\mathbf{v})] \propto \nabla \mathbb{E}_j[\bar{\psi}_j(\mathbf{x}_0 + \mathbf{v}) - \bar{\psi}_j(\mathbf{x}_0 - \mathbf{v})]$$

Evaluating both expectations:

$$\begin{aligned}\mathbb{E}_{j,\mathbf{v}}\big[\nabla p_j(\mathbf{v})\big] &\doteq \mathbb{E}_\mathbf{v}[\nabla \mathbb{E}_j[\bar{\psi}_j(\mathbf{x}_0+\mathbf{v}) - \bar{\psi}_j(\mathbf{x}_0-\mathbf{v})]] \\ &= \mathbb{E}_\mathbf{v}[\nabla\big(\,\bar{\Psi}(\mathbf{x}_0+\mathbf{v}) - \bar{\Psi}(\mathbf{x}_0-\mathbf{v})\,\big)] \\ &= \mathbb{E}_\mathbf{v}[(\nabla\bar{\Psi})(\mathbf{x}_0+\mathbf{v})] + \mathbb{E}_\mathbf{v}[(\nabla\bar{\Psi})(\mathbf{x}_0-\mathbf{v})]\end{aligned}$$

The first term, $\mathbb{E}_\mathbf{v}[(\nabla\bar{\Psi})(\mathbf{s}_0 + \mathbf{v})]$, is the definition of the gradient of $\bar{\Psi}$ convolved with the Gaussian kernel φ_Σ, evaluated at $\mathbf{s}_0$. This is precisely the quantity defined in the theorem as $\beta^* = \mathbb{E}_v[\nabla\bar{\Psi}(\mathbf{s}_0 + \mathbf{v})]$. Due to the symmetry of the Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma_\mathbf{x})$, the second term evaluates to the exact same quantity. Therefore:

$$\mathbb{E}_{j,\mathbf{v}}\big[\nabla p_j(\mathbf{v})\big] \doteq 2\beta^*$$

Substituting this exact result into the expression for $\mathbf{v}_{\text{CI}}$:

$$\mathbf{v}_{\text{CI}} \doteq \Sigma_\mathbf{x} \cdot \mathbb{E}\big[\nabla p_j(\mathbf{v})\big] \doteq \Sigma_\mathbf{x} \cdot 2\beta^*$$

We arrive at the final relationship of proportionality:

$$\boxed{\mathbf{v}_{\text{CI}} \propto \Sigma_\mathbf{x}\beta^*} \tag{7}$$

□

In words, the Classification Image averages the same expected gradient as OLS but multiplies it by $\Sigma_\mathbf{x}$, thereby tilting the estimate toward high-variance stimulus directions. Equivalently, the CI is proportional to the gradient of the *smoothed potential* $\bar{\Psi} * \varphi_\Sigma$, but expressed in the *covariance-induced* geometry.

In words, CI averages the same expected gradient as OLS but multiplies it by $\Sigma_\mathbf{x}$, thereby tilting the estimate toward high-variance stimulus directions. Equivalently, CI is the gradient of the *same smoothed potential* expressed in the *covariance-induced* geometry.

D Bridging Estimators via Regularization

The CI and the OLS Manipulation Vector appear distinct. We now demonstrate a formal, continuous link between them using Ridge Regression, showing that the model-based framework can be tuned to recover the CI.

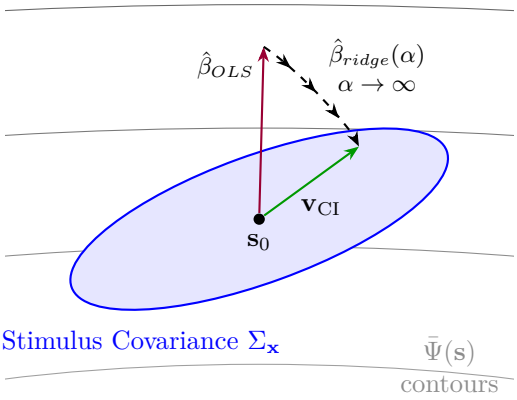


Figure 15: **Bridging Estimators via the Ridge Regression Path.** The directions learned by estimators of the impression $\bar{\Psi}(\mathbf{s})$. The CI vector, $\mathbf{v}_{\text{CI}}$, is biased toward the principal axes of the stimulus covariance $\Sigma_\mathbf{x}$. The OLS (manipulation) vector, $\hat{\beta}_{OLS}$, aligns with the impression gradient. The dashed line shows the expected ridge estimator direction (under the theorem below), as the regularization parameter α increases.

Definition D.1 (The Ridge Regression Manipulation Vector)**.** Ridge Regression modifies the OLS procedure by adding an L2 penalty on the magnitude of the Manipulation Vector, controlled by a hyperparameter $\alpha \geq 0$. The solution is:

$$\hat{\beta}_{ridge}(\alpha) = (V^T V + \alpha I)^{-1} V^T R \tag{8}$$

Note that for $\alpha = 0$, this simplifies to the OLS estimator, $\hat{\beta}_{OLS}$.

Theorem D.2 (Classification Image of a Regression Model)**.** *The rescaled expectation of a Ridge Linear Regression estimator, in the limit of infinite regularization,*

converges to the Classification Image (CI) vector.

$$\lim_{\alpha\to\infty} \mathbb{E}[\alpha \cdot \hat{\beta}_{ridge}(\alpha)] = \mathbf{v}_{CI} \tag{9}$$

And this limit gives a closed-form solution for the Classification Image of a regression model:

$$\mathbf{v}_{CI} = \Sigma_{\mathbf{x}} \cdot \hat{\beta}_{OLS} \tag{10}$$

Proof. The expected value of the Ridge estimator is $\mathbb{E}[\hat{\beta}_{ridge}(\alpha)] = (\mathbb{E}[V^T V] + \alpha I)^{-1}\mathbb{E}[V^T R]$. In the limit of large N, the first term gives $(\Sigma_{\mathbf{x}} + \alpha I)^{-1}$ while the second $\mathbf{v}_{\text{CI}}$. To analyze the limit as $\alpha \to \infty$, we multiply by α:

$$\begin{aligned}
\mathbb{E}[\alpha \cdot \hat{\beta}_{ridge}(\alpha)] &= \alpha(\Sigma_{\mathbf{x}} + \alpha I)^{-1}\mathbf{v}_{\text{CI}} \\
&= \alpha \left[\alpha\left(\frac{1}{\alpha}\Sigma_{\mathbf{x}} + I\right)\right]^{-1}\mathbf{v}_{\text{CI}} \\
&= \left(I + \frac{1}{\alpha}\Sigma_{\mathbf{x}}\right)^{-1}\mathbf{v}_{\text{CI}}
\end{aligned}$$

As $\alpha \to \infty$, the term $\frac{1}{\alpha}\Sigma_{\mathbf{x}} \to \mathbf{0}$. Therefore, the operator approaches the identity matrix:

$$\lim_{\alpha\to\infty} \left(I + \frac{1}{\alpha}\Sigma_{\mathbf{x}}\right)^{-1} = I$$

This proves the first part of the theorem:

$$\lim_{\alpha\to\infty} \mathbb{E}[\alpha \cdot \hat{\beta}_{ridge}(\alpha)] = I \cdot \mathbf{v}_{\text{CI}} = \mathbf{v}_{\text{CI}}$$

To prove the second part, we use the established relationship between the expected values of the estimators, $\mathbb{E}[\hat{\beta}_{OLS}] = \Sigma_{\mathbf{x}}^{-1}\mathbf{v}_{\text{CI}}$, which rearranges to $\mathbf{v}_{\text{CI}} = \Sigma_{\mathbf{x}}\mathbb{E}[\hat{\beta}_{OLS}]$. Substituting this into our proven limit gives:

$$\lim_{\alpha\to\infty} \mathbb{E}[\alpha \cdot \hat{\beta}_{ridge}(\alpha)] = \Sigma_{\mathbf{x}} \cdot \mathbb{E}[\hat{\beta}_{OLS}]$$

As the finite-sample estimator $\hat{\beta}_{OLS}$ is an unbiased estimate of its expectation, we can state the relationship in terms of proportionality, completing the proof. □

Figure 16: Ridge regression direction between $\alpha = 0.001$ and $\alpha = 1.0$ converging to the *CI* for real trustworthiness ratings in feature space.

