PROBING PERCEPTUAL PRIORS OF MLLMS VIA GIBBS SAMPLING WITH INTERPRETABLE GENERATIVE CONTROLS

Anonymous authors
Paper under double-blind review

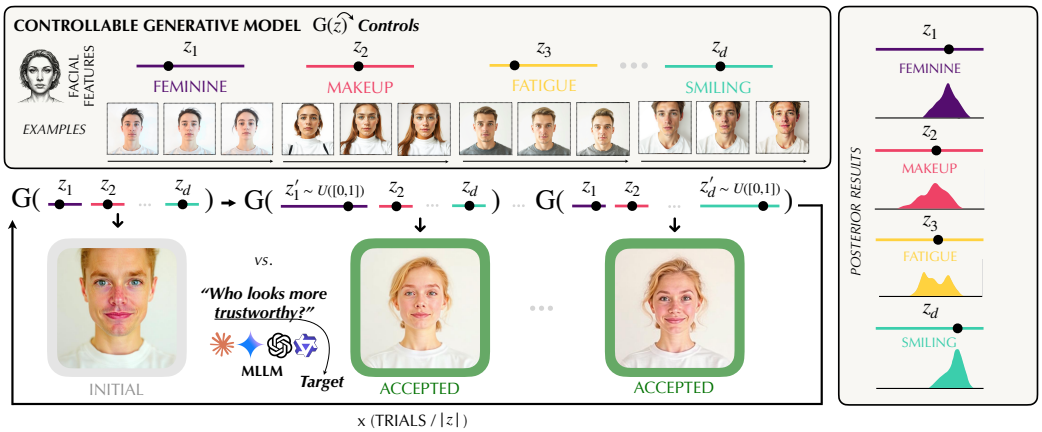


Figure 1: We introduce a method for probing the perceptual priors of multimodal language models by running a Metropolis–Hastings within Gibbs sampler over the interpretable latent space of a generative model. At each trial, the model is shown the current and a proposed generated stimulus (differing along one slider) and asked to choose the one that matches the target concept (i.e., the object under study) closely. Its binary choice realizes a Barker acceptance step, while iterating across slider dimensions yields Markov-chain samples from the model's implicit prior over the target. We apply the method across four domains (see examples in Appendix A) where model priors shape how these models perceive us and our world.

ABSTRACT

A model's behavior on a task is jointly determined by the input it receives and the prior it brings in, i.e. the distribution over stimuli it implicitly expects. Interpretability research has traditionally studied models by holding inputs fixed and examining model responses either mechanistically, probing how internal structure represents inputs, or behaviorally, measuring how variation in inputs leads to variation in outputs. Neither reconstructs the prior distribution itself, since internal structure shows what a model can represent, not what it expects, and any fixed stimulus set leaves most of the possible input space unseen. In particular, such an input space in real-world settings, such as images seen by VLMs, is extremely high-dimensional and diverse. These priors thus remain a poorly understood component of models that nonetheless influence real-world behavior. We propose a method to sample from models' perceptual prior distributions directly, by steering a generative model to produce stimuli along controllable axes and running Gibbs sampling over that space with the model under study as the judge. We apply this to a variety of categories and target variables (such as trustworthiness in faces and cheapness in art images) and recover both canonical biases and surprising novel priors invisible to direct prompting, warranting further investigation of their downstream effects.

1 INTRODUCTION

When humans make decisions, we draw on implicit representations of the world. Our priors shape what we notice, what we prefer, and ultimately how we behave. Thus, behaving intelligently depends

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

fundamentally on the quality of these representations. Models are no different, and their outputs are similarly shaped by representations learned from data. Yet while decades of work in cognitive science have developed techniques for characterizing human perceptual priors, we have comparatively few tools for probing the analogous representations of these models. This is particularly important for safety and alignment, as delegating to a model implicitly assumes its priors align with those of the humans it serves, an assumption that is rarely tested and, when it fails, can propagate biases invisibly through downstream systems.

The classical approach is to present participants with stimuli that span the space of interest and ask them to evaluate each one. In low-dimensional spaces, this works well. A small number of stimuli can densely cover the space, and the resulting evaluations directly map out the structure of the representation. But the number of stimuli needed grows combinatorially with the number of controllable dimensions, and any fixed set is implicitly a bet that the regions of interest lie within it. In high dimensions, this bet rarely pays off. Markov Chain Monte Carlo with People (MCMCP) (Sanborn & Griffiths, 2007; Sanborn et al., 2010; Sanborn & Griffiths, 2015) and its extension Gibbs Sampling with People (GSP) (Harrison et al., 2020) address this problem by letting people's own judgments determine which stimuli are presented next. In MCMCP, participants make a binary choice at each step, and these choices act as the acceptance step of a Metropolis-Hastings sampler over the parameter space. GSP replaces the binary choice with a continuous slider and updates one dimension at a time, effectively implementing a Gibbs sampler that converges faster in high-dimensional spaces. Both methods were developed to extract human semantic representations, but they can also inspire similar methods for models.

In this paper, we introduce a method that applies Gibbs sampling to study models rather than people. This requires a generative model with controllable, interpretable dimensions, allowing each one to be sampled in turn. Although the method is modality-agnostic, we validate it in the image domain and train sliders via SliderSpace (Gandikota et al., 2025), which provides LoRA-based sliders for the text-to-image model FLUX (Labs et al., 2025). We apply the method across four domains: faces, affordances, aesthetics, and authenticity. Each one probes priors that shape consequential decisions. Faces capture how models see us, the lens through which they increasingly screen applicants, defendants, or partners. Affordances test whether models share the physical intuitions required for robotics and embodied reasoning. Aesthetics expose the priors that govern what generative and recommendation systems produce. Finally, authenticity reveals the priors that determine what models accept as real or altered, with implications for misinformation and content moderation. Overall, this work contributes:

1. A Bayesian method for sampling the perceptual priors of multimodal language models over interpretable generative latent spaces. Unlike prior MCMC work that operates over simple words or closed-form numerical inputs (Zhu et al., 2024; Zhu & Griffiths, 2024; Marjieh et al., 2024b), our setup is validated with images as inputs over a complex text-to-image generator (FLUX + SliderSpace).
2. An open-source modality-agnostic framework: any controllable, interpretable generator paired with a binary preference judge plugs into the same scheme, so the method extends naturally to other modalities.
3. Experiments across four high-stakes image domains (faces, affordances, aesthetics, and authenticity), using four frontier VLMs (*Claude Sonnet 4.6*, *Gemini 3 Flash*, *GPT-5.4*, *Qwen3-VL 235B*), recovering priors as full posteriors that direct prompting fails to surface.

2 Related Work

2.1 Eliciting Human Perceptual Priors

Cognitive scientists have a long tradition of developing methods for recovering human perceptual priors. MCMCP (Sanborn & Griffiths, 2007; Sanborn et al., 2010; Sanborn & Griffiths, 2015) elicits these priors through adaptive Metropolis-Hastings sampling (Metropolis et al., 1953; Hastings, 1970). It casts the participant's binary choice as Barker's acceptance rule (Barker, 1965), equivalent to a Bradley-Terry model of paired comparison (Bradley & Terry, 1952; Christiano et al., 2017), valid under the assumption that humans probability-match their choices to the underlying preference (Vulkan, 2000; Shanks et al., 2002). GSP (Harrison et al., 2020) extends this with continuous slider

responses and dimension-wise updates, effectively implementing a Gibbs sampler (Geman & Geman, 1984) that converges faster in high-dimensional spaces. GSP has since been applied across diverse perceptual domains, including musical consonance (Marjieh et al., 2022), singing (Anglada-Tort et al., 2023), emotional prosody (Van Rijn et al., 2021; van Rijn et al., 2022), visual patterns (Kumar et al., 2022), musical chords (Marjieh et al., 2024a), and multidimensional visual aesthetics (Van Geert & Jacoby, 2024). We adapt this family of methods to multimodal language models (MLLMs) rather than people, taking the Gibbs structure from GSP and the Barker-rule binary acceptance (Barker, 1965) from MCMCP, leveraging categorical rather than continuous responses. The combination is a Metropolis-Hastings within Gibbs sampler (Tierney, 1994), with the MLLM playing the role of the acceptance function. By updating one dimension at a time, our method's efficiency is closer to GSP than MCMCP in high-dimensional spaces.

2.2 Probing Models

Existing model evaluations are largely functional, asking whether the model produced the right output. But behavioral systems, such as agents powered by these models, can have different decision processes that produce identical outputs. Therefore, behavioral tests are needed to surface why a model behaves the way it does, what conditions change its behavior, and how it differs from humans (Cherep et al., 2025b; 2026b; 2024). Such tests require new tools, methods, and frameworks across a range of domains. From counterfactual environments that enable the causal exploration of agent behavior (Cherep et al., 2025a), to multi-agent environments (Vezhnevets et al., 2023), to visual prompt optimization to study what visual features drive agentic decisions (Cherep et al., 2026a). Together, these efforts probe agents at multiple levels both internal and external. Our work adds another tool, recovering the perceptual priors that shape how models behave.

A growing body of work probes models in Bayesian terms. Zhu et al. (2024) use MCMC with LLMs to recover mental representations over text. Zhu & Griffiths (2024) elicit LLM priors over abstract numerical and causal quantities through iterated in-context learning. Marjieh et al. (2024b) compare LLM judgments to human sensory judgments across six modalities. Unlike these, our work probes priors directly within a controllable and multimodal generative space rather than over text. Because the same Gibbs protocol applies to both humans and models, the recovered posteriors are directly commensurable in the same latent space.

2.3 Controllable Generative Models

Our method requires a generative model with independently controllable dimensions. The original GSP experiments used StyleGAN (Karras et al., 2019; 2020) controlled via GANSpace (Härkönen et al., 2020), which discovers directions through PCA in the latent space. These directions are sometimes recognizable, but they are not designed to be interpretable, and many remain entangled. Interpretability matters for our purposes because it is what makes the recovered priors legible. Without interpretable axes, we cannot link a posterior over the latent space to meaningful concepts. Instead, we use SliderSpace (Gandikota et al., 2025), which extends Concept Sliders (Gandikota et al., 2024) (LoRA-based controls for diffusion models) by automatically decomposing a model's visual capabilities into named, interpretable axes. SliderSpace also can operate over FLUX (Labs et al., 2025), which produces substantially more realistic and diverse images than StyleGAN. Other approaches to disentangled control are compatible with our method, including dedicated semantic latent spaces in diffusion models (Kwon et al., 2022; Haas et al., 2024) and sparse autoencoders (Surkov et al., 2025; Singh et al., 2025).

3 Methods

We now turn to the mechanics of running Gibbs sampling with a multimodal LM as the judge. The natural design is to give the model a slider and ask it to pick a position, but this turns out to be impractical. The cost of every small slider move is high in this regime, and the perception-action loop is turn-based for VLMs which means they can expend an unbounded number of rounds making small moves in slider space by trial and error. We also observed VLMs' tendency to oscillate, explore the bounds, and take arbitrarily small steps back and forth. To mitigate this, we propose to use Metropolis-Hastings within Gibbs with the Barker acceptance function as a tractable replacement.

Setup. Let $G : \mathcal{Z} \to \mathcal{X}$ be a controllable text-to-image generator, where $\mathcal{Z} = \prod_{k=1}^{d} \mathcal{Z}_k$ is a product of d interpretable, mutually orthogonal slider axes (for example, as in our case, obtained via training SliderSpace (Gandikota et al., 2025) on top of FLUX.1-schnell (Labs et al., 2025)). Each $\mathcal{Z}_k$ we first reparameterize to $[0, 1]$. We use SliderSpace to learn $n = 64$ candidate axes per domain, and label all axes automatically (see Appendix D); the human interpreter then selects a subset of $d \leq n$ axes whose effects are semantically interpretable and potentially useful, which preserves the orthogonality property of the parent set.

Throughout, we fix a VLM judge and a textual criterion C (e.g. "more attractive"). The judge receives two stimuli and returns a binary choice. We assume choices are governed by an implicit, unobserved utility $\ell : \mathcal{Z} \to \mathbb{R}$ via the Bradley-Terry model (Bradley & Terry, 1952):

$$\Pr[\text{judge picks } z' \mid z, z'] = \sigma\big(\ell(z') - \ell(z)\big), \quad \text{where } \sigma(u) = (1 + e^{-u})^{-1}. \tag{1}$$

This is the same primitive that underlies much of the RLAIF literature (Bai et al., 2022; Lee et al., 2023). The induced target distribution is then:

$$\pi(z) \propto \exp\big(\ell(z)\big), \tag{2}$$

which is the GSP target distribution (Harrison et al., 2020) when the VLM's noise scale is absorbed into ℓ. Modes of π are accordingly the "prototypes" the VLM endorses for C, and the spread of π describes the variation it tolerates. Here, two methods from the GSP/MCMCP family are most directly useful: *joint MCMCP* (binary comparison with full-state proposals) and *direct GSP* (asking the agent to pick a slider position). Both have been adapted to language models on low-dimensional domains (Zhu et al., 2024), but (1) our combinatorial dimensionality is much higher here and queries are much costlier, and (2) our generator is much more complex than the number–to–color mapping, and thus closed-form numbers from LLMs would be uninterpretable as generation parameters.

As we noted earlier, the natural design (giving the VLM a slider and asking it to pick a position) is implausible in this regime. Zhu et al. (2024) avoid such issues in a 3-D HSL space by asking GPT-4 to emit a value for the missing dimension, which is effectively a closed-form numerical answer to a closed-form numerical question. Our setting has no such closed form, since the slider is over generated images, and the "answer" is thus a position whose meaning the VLM only reliably knows *after* another generation. This complicates the purported GSP information advantage in our setting.

A natural alternative is joint MCMCP, which proposes the entire z at once and still uses a single binary VLM comparison per step. The single-query primitive is intuitively appealing, but the joint proposal is not scalable. Roberts et al. (1997) show that for random-walk Metropolis on a d-dimensional target, the proposal scale must shrink as $\sigma_d \propto d^{-1/2}$ to maintain a non-degenerate acceptance rate, after which the convergence time grows proportionally. This becomes a high-query design, which is ruled out by the high cost of stimulus generation. Our scan must therefore amortize stimulus cost over as much information as possible.

3.1 MH-WITHIN-GIBBS WITH BARKER ACCEPTANCE

We propose to instead use a Metropolis-Hastings within Gibbs sampler (Tierney, 1994) that uses the binary VLM comparison as its acceptance step (see Algorithm 1). This construction has three pieces:

1. We update one coordinate at a time, conditional on the others. This is the structural device of GSP (Harrison et al., 2020).
2. Within each coordinate update, we use the Barker acceptance rule (Barker, 1965):

$$\alpha_B(z, z') = \frac{\pi(z')}{\pi(z) + \pi(z')} \tag{3}$$

Barker is the acceptance rule used by MCMCP (Sanborn & Griffiths, 2007). Substituting equation 2 into equation 3 gives

$$\alpha_B(z, z') = \frac{e^{\ell(z')}}{e^{\ell(z)} + e^{\ell(z')}} = \sigma\big(\ell(z') - \ell(z)\big) \tag{4}$$

which is exactly the BT comparison probability (Equation (1)). Every binary VLM query thus realizes a single Barker accept/reject step at no additional cost, the same property exploited by (Sanborn & Griffiths, 2007; Zhu et al., 2024) in the joint-proposal regime.

3. Given current state $z = (z_1, \ldots, z_d) \in \mathcal{Z}$, one full scan iterates $k = 1, \ldots, d$:
   (a) Propose $z'_k \sim q_k(\cdot \mid z_k)$ with q_k a symmetric kernel on $\mathcal{Z}_k$; let $z' = (z_1, \ldots, z'_k, \ldots, z_d)$
   (b) Generate $G(z)$ and $G(z')$. Present them to the VLM with criterion C in randomized order; receive choice $Y \in \{z, z'\}$
   (c) Set $z \leftarrow Y$

   We use $q_k \sim \mathrm{Uniform}([0,1])$[1], drawing each proposal independently of the current state (in Appendix F we use a random walk to generate each proposal with the Robbins–Monro rule (Robbins & Monro, 1951)). This is Independent Metropolis-Hastings (IMH) within Gibbs. One scan across dimensions thus costs $2d$ image generations and d VLM queries. The Gibbs decomposition then reuses one of the two generations from the previous step in each scan, reducing the amortized cost to $d + 1$ generations per scan.

Algorithm 1 Metropolis–Hastings within Gibbs with Barker acceptance

Require: Initial slider state $z^{(0)} \in [0,1]^d$, criterion C, initial proposal scale σ_0, target acceptance rate α^*, decay η, burn-in fraction b, number of trials T.

$\sigma \leftarrow \sigma_0$
for $\tau = 1, \ldots, T$ **do**
  $k_\tau \leftarrow ((\tau - 1) \bmod d) + 1$ ▷ Gibbs sweep over coordinates
  if Proposal is uniform **then**
    Sample $z^* \sim \mathcal{U}(0,1)$
  else if Proposal is Gaussian **then**
    Sample $z^* \sim \mathcal{N}(z_{k_\tau}^{(\tau-1)}, \sigma^2)$ reflected at $[0,1]$
  end if
  Set proposed state $z' \leftarrow (z_1^{(\tau-1)}, \ldots, z^*, \ldots, z_d^{(\tau-1)})$, replacing coordinate k_τ
  Generate current stimulus $G(z^{(\tau-1)})$
  Generate proposed stimulus $G(z')$
  Randomize image order & query VLM under criterion C ▷ Barker rule from VLM choice
  if the VLM chooses $G(z')$ **then**
    $z^{(\tau)} \leftarrow z'$
    $a \leftarrow 1$
  else
    $z^{(\tau)} \leftarrow z^{(\tau-1)}$
    $a \leftarrow 0$
  end if
  if Proposal is Gaussian and $\tau \leq bT$ **then**
    $\log \sigma \leftarrow \log \sigma + \tau^{-\eta}(a - \alpha^*)$ ▷ Robbins–Monro rule (Robbins & Monro, 1951)
  end if
end for

Importantly, we observe that the BT assumption (Equation (1)) need only hold along single SliderSpace axes for our chain to be valid:

Assumption 1 (Coordinate-wise Bradley-Terry)**.** *For every coordinate $k \in \{1, \ldots, d\}$, every $z_{-k} \in \prod_{j \neq k} \mathcal{Z}_j$, and every pair $z_k, z'_k \in \mathcal{Z}_k$:*

$$\Pr[\textit{VLM picks } (z'_k, z_{-k}) \mid (z_k, z_{-k}), (z'_k, z_{-k})] \;=\; \sigma\big(\ell(z'_k, z_{-k}) - \ell(z_k, z_{-k})\big)$$

Assumption 1 is weaker than the full BT assumption, since it only requires BT to hold for comparisons that vary on a single axis. In our regime, Assumption 1 is more plausible than full BT for a few reasons. First, SliderSpace explicitly seeks "semantic orthogonality" among the principal directions (Gandikota et al., 2025) (axes are initialized as principal components of the CLIP-embedding distribution induced

[1]This type of construction is variously known as independence chains/sampling (Tierney, 1994; Roberts & Rosenthal, 1998), Metropolized independent sampling (Liu, 1996), or independent Metropolis–Hastings (IMH) (Holden et al., 2009). These denote proposals independent of the current state. We use the latter (IMH), and study the special case $q_k \sim U(0,1)$.

by the prompt, and the LoRA adapter for each axis is trained to align $\Delta\phi_k = \phi(G(z+\delta e_k)) - \phi(G(z))$ with the corresponding principal direction v_k). Single-axis perturbations therefore produce perceptual changes confined to one orthogonal CLIP direction by construction. Second, SliderSpace also enforces that each axis induces consistent transformations across seeds and prompt variations (Gandikota et al., 2025). The variation along z_k at a fixed z_{-k} thus has stable semantic meaning, which is what allows $\ell(\cdot, z_{-k})$ to be a coherent function of z_k. Finally, if we human-select a subset of d axes from the n that SliderSpace produces, the orthogonality and consistency properties survive regardless. Orthogonality of $\{v_1, \ldots, v_n\}$ in CLIP space implies orthogonality of any subset, and per-axis consistency is independent of the other axes retained. The complementary assumption we make is that SliderSpace orthogonality maps to a regime where VLM judges are reliable. Empirically, prior work has found that LLM/VLM judgements can degrade in multi-attribute settings (Stureborg et al., 2024). Though not directly comparable to our setting, it further supports the value of reducing VLM responses to coordinate-wise comparisons. We therefore expect Assumption 1 to hold to a closer approximation than its full-BT counterpart would.

3.2 Controllable Generative Models

Our sampler requires a generator with two properties: (1) each controllable dimension corresponds to a single, semantically interpretable feature; and (2) different dimensions are at least approximately orthogonal, so the coordinate-wise updates licensed by Assumption 1 do not couple unrelated factors.

We instantiate this with FLUX.1-schnell (Labs et al., 2025) as the generator and SliderSpace (Gandikota et al., 2025) to train the orthogonal interpretable controls. We train a rank-1 LoRA for $n = 64$ candidate sliders per domain (faces, aesthetics, authenticity, and affordances) with 50,000 PCA samples over the CLIP-embedding distribution induced by domain-specific diverse prompts; compute requirements are reported in Appendix H. Effectively, perturbing the slider value translates the generated image's CLIP embedding along the corresponding principal axis. The procedure is agnostic to the embeddings used to construct these controllable coordinates. We use CLIP because it is domain-general, but domain-specific embeddings (which might capture better domain details) can be substituted without changing the sampler. Thus, CLIP-specific orthogonality is a property of our instantiation rather than the sampling framework, which can operate over any sufficiently disentangled controllable representation.

Because the n directions are PCA-orthogonal in CLIP space by construction, any subset is also orthogonal. Sampling all 64 dimensions is computationally prohibitive, so we select $d = 10$ sliders per domain based on semantic interpretability, visual effect strength, and nonredundancy (see Appendix B for examples). The selection is independent of the downstream target labels and is made before running the probing experiments. Omitted dimensions limit coverage, but do not bias selection toward the reported target-specific findings. Slider training details are in Appendix M, and validation that each selected axis behaves as a usable Gibbs coordinate, and that joint configurations actually steer the generator across distinct image distributions, is in Appendix K. We automatically label the sliders using an auto-interpretability procedure described in Appendix D.

3.3 Probing Experiments

We probe the priors of four frontier multimodal language models: *Claude Sonnet 4.6*, *Gemini 3 Flash*, *GPT-5.4*, and *Qwen3-VL 235B*. All models are queried with default temperature and sampling parameters. For each binary comparison, the order in which the current and proposed images are presented to the judge is independently randomized per query, so that the judge's preference is invariant to position effects (Pezeshkpour & Hruschka, 2024). For each (model, target) configuration, we run 10 independent chains of 2,000 trials each, with $d = 10$ sliders per domain. Seeds differ per chain and per target, so no two chains or targets share initial images.

We probe four domains, each with its own slider set and target list: **Faces** (8 targets): *attractive*, *criminal*, *fun*, *hardworking*, *intelligent*, *serious*, *trustworthy*, *youthful*; **Aesthetics** (5 targets): *amateur*, *beautiful*, *cheap*, *expensive*, *experimental*; **Authenticity** (2 targets): *authentic*, *manipulated*; **Affordances** (8 targets): *sittable*, *carryable*, *throwable*, *breakable*, *bounceable*, *resonant*, *rough*, *squeezable*. This gives 23 targets total. The full task and prompts per domain are listed in Appendix J, and compute requirements are in Appendix H. Before running the main experiments, we sanity-check the kernel on a controlled low-dimensional task with clear ground truth (HSL colors, Appendix G).

4 RESULTS

Across all domains and VLMs, our experiments span 2.5M+ trials and 1B+ tokens. Qualitative examples of the recovered priors across the four domains are shown in Appendix A. Unless otherwise stated, posterior summaries are computed over post-burn-in samples from the IMH-within-Gibbs chains. Because the sampled quantities are slider coordinates in a controllable generative model, the reported values should be interpreted as model-specific association signals within the reachable SliderSpace–FLUX stimulus manifold (i.e. not as direct measurements of real-world attributes implied by the labels). For reference, Appendix L shows the top-3 and bottom-3 sliders per target across the four models, and Appendix I shows posterior summaries for all (target, parameter, model) conditions in the experiment.

4.1 STRONGEST AND WEAKEST RECOVERED PRIORS

Figure 2 (left) reports the strongest recovered priors across the 23 targets, ordered by cross-model mean. For face traits with cognitive-positive valence (*Intelligent*, *Hardworking*, *Serious*) the top-recovered slider is the same: *Eyeglasses*. For face traits with affective-positive valence (*Fun*, *Trustworthy*) the top-recovered slider is *Smiling*. In the authenticity domain, models converge on visual edit-signatures (*Black and White*, *Crushed Blacks*, *Upscaling Noise*) as markers of *manipulated* images. For affordances, the top-recovered slider for *Rough* is *Paper/Material*. For art/aesthetics, the top slider is *Ornate Detail* for *Beautiful*.

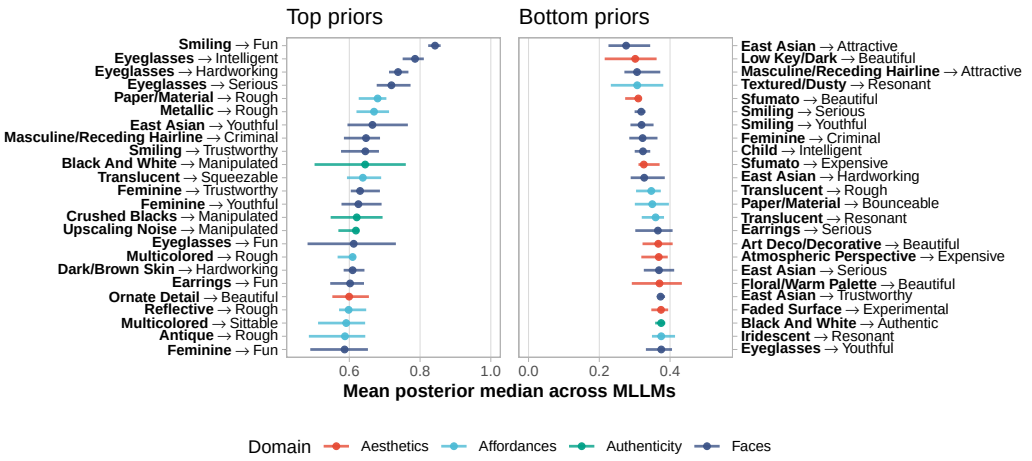


Figure 2: Strongest (left) and weakest (right) recovered priors across the 23 targets, pooled across the four MLLM judges. Each row is a (target, slider) condition ordered by mean across models; the point is the cross-model mean and the bar shows the cross-model interquartile range. Slider values are in $[0, 1]$; a value of 0.5 corresponds to no preference along that axis.

On the right of Figure 2 are results in the opposite direction, showing the most negative (below the 0.5 baseline) associations models appear to encode. For the non-identity targets, several recovered priors place below baseline: *Smiling* for *Serious* and *Youthful*, *Child* for *Intelligent*, *Sfumato* for *Beautiful* and *Expensive*, *Low Key/Dark* for *Beautiful*, and *Textured/Dusty* for *Resonant*. We next look at identity-related priors.

4.2 DEMOGRAPHIC ASSOCIATIONS

A central purpose of this proposed method is to make visible associations that direct prompting may not, so that they can be audited. As such models are increasingly used to screen, rank, or moderate content involving people, a directional preference linking a target concept to a demographic-linked visual axis is consequential even when a model might not state it when asked, as has been shown in prior work on covert biases (Hofmann et al., 2024). Among the recovered priors, several such

conditions appear. We report them as candidate implicit biases the method exposes, deserving of further scrutiny. However, we note that the chain only recovers which direction along an axis a judge prefers for a given criterion. Whether that preference originates in pre-training or post-training alignment is outside the scope of what these experiments can establish (Section 5). All values reported are medians over the full slider posteriors, where 0.5 is the no-preference point.

Concerningly, for *Attractive*, *Trustworthy*, *Intelligent*, and *Hardworking*, all 4 judges place the *East Asian* and *Receding Hairline* auto-labeled axes below the baseline. This association is target-dependent and not a uniform prior over the slider: for *Youthful* the same *East Asian* slider sits well above baseline for 3/4 judges, as does *Receding Hairline* for *Criminal*. Conversely, the *Feminine* axis places below baseline for *Criminal* across all 4 judges and, more weakly, for *Intelligent* across all four, while being above baseline for *Attractive* and *Trustworthy* in 3/4. Taken together across judges, the axes auto-interpreted as masculinity raise perceived criminality and lower perceived attractiveness, trustworthiness, and intelligence, and femininity inverts much of that ordering (though not on intelligence).

When split by model, dominant priors recur across judges, but the strength varies. *Gemini 3 Flash* produces the strongest extremes in both directions: for instance, it places *Receding Hairline* at 0.87 for *Criminal* (vs. pooled 0.65) but pushes *East Asian* for *Attractive* down to 0.16, the most extreme version of the aforementioned bias. *Claude Sonnet 4.6* and *GPT-5.4* are comparatively less extreme. Claude is most often out of step with the others: it puts low utility on *Glasses* for *Trustworthy* (0.28) and its low magnitudes are uniformly gentler (0.27–0.29) than Gemini's (0.16–0.23). We visualize these in Appendix L, Figure 20. The agreement structure quantified next is partly a consequence of these per-model styles.

4.3 Cross-model Rank Agreement

Figure 3 reports cross-model Spearman rank correlations on the recovered slider values, by domain. *Authenticity* is the highest-consensus domain (median $\rho \in [0.42, 0.69]$ across the six model pairs); *Faces* and *Affordances* are intermediate ($[0.30, 0.56]$); *Aesthetics* is the lowest ($[0.07, 0.30]$). When the concept has a sharp visual signature (e.g., an edited photograph), models agree on the slider it cues; when the concept is diffuse (e.g., *cheap* or *experimental* aesthetics), they pick up different correlates. *Claude Sonnet 4.6* is in the lowest-ρ pair for both *Faces* (Claude–Qwen, $\rho = 0.30$) and *Aesthetics* (Claude–Gemini, 0.07), consistent with per-model results, where Claude was the model most often out of step.

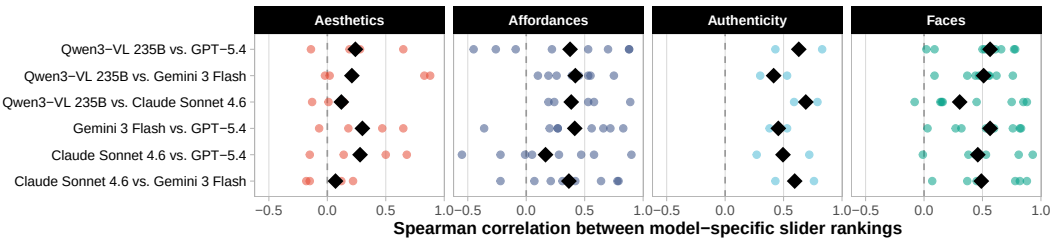


Figure 3: Cross-model rank agreement on recovered slider values, per domain, across the six model pairs. Higher ρ means more consensus on which sliders matter for the targets in that domain.

4.4 Top-slider Selectivity per Target

Figure 4 reports selectivity, the gap between the top slider and the rest of that target's sliders. Face sliders: *Intelligent/Eyeglasses* (0.33), *Serious/Eyeglasses* (0.32), *Fun/Smiling* (0.30), *Hardworking/Eyeglasses* (0.30). *Authentic*—with *Natural Edit* as its nominal top slider—has selectivity 0.05 across models, close to flat. *Cheap* is also diffuse (Floral Warm/Palette = 0.12). These results corroborate the absence of *Authentic* and *Cheap* from Figure 2.

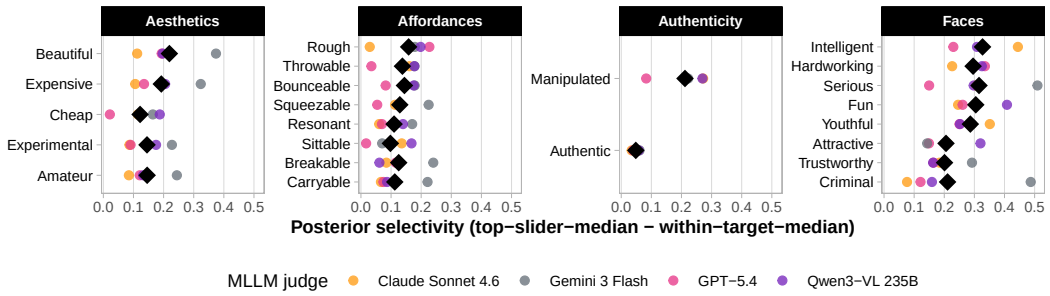


Figure 4: Per-target selectivity of the top-recovered slider, defined as the gap between the top slider's posterior median and the median posterior median across all 10 sliders for that target (per model, then averaged across models).

4.5 Explicit-Prior Baselines

Our sampler is designed to recover a joint implicit prior over the selected slider space. A natural baseline is to ask whether simpler, explicit elicitation procedures recover the same structure. We therefore compare IMH against two direct baselines that query the model on fixed slider sweeps: a scalar rating task (how well does each sample match the target?) and a paired forced-choice task (between two samples, which better matches the target? This is the same binary choice primitive as our primary IMH method). The full set of experiments is described in Appendix C, which shows that these recover more uniform-like (less informative) priors, and that their misalignment from the IMH-Gibbs approach correlates with inter-slider correlations (which explicit methods assume are effectively zero).

Finally, we give additional convergence, predictive, and interaction diagnostics in Appendix E.

5 Limitations

Our method inherits its expressiveness from the underlying generative model: the recovered prior can only span what SliderSpace can learn and FLUX can express, and any concept absent from or poorly represented by the latent space is invisible to the chain. While SliderSpace is trained to produce orthogonal sliders, multi-attribute interactions could in principle violate this. However, this could be improved by using more specialized embeddings than CLIP. Finally, while the recovered priors expose biases that direct prompting hides, our method does not by itself attribute them; downstream causal analysis is needed to determine whether a recovered preference reflects pre-training data, post-training alignment, or in-context confounds.

6 Conclusion

Cognitive scientists have spent decades building methods to extract what humans cannot say, the priors their perceptual systems carry into every decision. With recent AI models, our traditional practice has been to prompt them. However, as these models start adjudicating consequential decisions, and direct prompting reveals little of what they actually perceive, we need methods that let the models' own choices, rather than their introspective reports, drive the search. Characterizing what models bring to a task, before they act on it, is increasingly a prerequisite for designing and governing them responsibly. We show that the same sampling tradition, a Gibbs sampler whose acceptance step is the model's own pairwise preference, transfers when run over inputs in the model's native perceptual modality, recovering full posteriors. Finally, while surfacing the priors of a model has positive impacts for auditing them, it also makes implicit priors legible to potential bad actors who may amplify or exploit them in downstream pipelines.

AI USE STATEMENT

In this work, we used generative AI tools for co-implementing methods, supporting qualitative data analysis, and providing feedback on research methodology. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, cleaning and reformatting datasets, interpreting results, proposing or refining hypotheses; and generating synthetic data sets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, brainstorming, and editing the paper to improve readability. We have reviewed all AI-assisted work: The agent-generated code was verified and tested for correctness by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The method described in this paper can reveal model priors that can be harmful to certain demographics, as discussed in more detail in Section 4.2. While it's important to develop methods to understand model priors, malicious actors can also exploit these to bias model decision-making.

REPRODUCIBILITY STATEMENT

All details to reproduce the results in this paper are in this manuscript (including the Appendix), and the code, which will be released as an open-source contribution.

REFERENCES

Sanket Agrawal, Dootika Vats, Krzysztof Łatuszyński, and Gareth O Roberts. Optimal scaling of mcmc beyond metropolis. *Advances in Applied Probability*, 55(2):492–509, 2023.

Manuel Anglada-Tort, Peter MC Harrison, Harin Lee, and Nori Jacoby. Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Current Biology*, 33(8):1472–1486, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Anthony Alfred Barker. Monte carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Manuel Cherep, Nikhil Singh, and Pattie Maes. Superficial alignment, subtle divergence, and nudge sensitivity in llm decision-making. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

Manuel Cherep, Chengtian Ma, Abigail Xu, Maya Shaked, Pattie Maes, and Nikhil Singh. A framework for studying AI agent behavior: Evidence from consumer choice experiments. *ICLR 2026*, 2025a.

Manuel Cherep, Nikhil Singh, and Patricia Maes. Behavioral systems require behavioral tests. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*, 2025b.

Manuel Cherep, Pranav M R, Pattie Maes, and Nikhil Singh. Visual persuasion: What influences decisions of vision-language models? In *International Conference on Machine Learning*, 2026a. URL https://arxiv.org/abs/2602.15278.

Manuel Cherep, Pattie Maes, and Nikhil Singh. AI agents are sensitive to nudges. *Proceedings of the National Academy of Sciences (PNAS)*, 2026b.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15994–16003, 2025.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 721–741, 1984.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Graßhof, Sami S Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2024.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. Gibbs sampling with people. *Advances in neural information processing systems*, 33:10659–10671, 2020.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024.

Lars Holden, Ragnar Hauge, and Marit Holden. Adaptive independent metropolis–hastings. 2009.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Sreejan Kumar, Carlos G Correa, Ishita Dasgupta, Raja Marjieh, Michael Y Hu, Robert Hawkins, Jonathan D Cohen, Karthik Narasimhan, Tom Griffiths, et al. Using natural language and program abstractions to instill human inductive biases in machines. *Advances in Neural Information Processing Systems*, 35:167–180, 2022.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, 6(2):113–119, 1996.

Raja Marjieh, PM Harrison, Harin Lee, Fotini Deligiannaki, and Nori Jacoby. Reshaping musical consonance with timbral manipulations and massive online experiments. *bioRxiv*, 2022.

Raja Marjieh, Peter MC Harrison, Harin Lee, Fotini Deligiannaki, and Nori Jacoby. Timbral effects on consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales. *Nature Communications*, 15(1):1482, 2024a.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445, 2024b.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Gareth O Roberts and Jeffrey S Rosenthal. Markov-chain monte carlo: some practical implications of theoretical results. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pp. 5–20, 1998.

GO Roberts, A Gelman, and WR Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, pp. 110–120, 1997.

Adam Sanborn and Thomas Griffiths. Markov chain monte carlo with people. *Advances in neural information processing systems*, 20, 2007.

Adam N Sanborn and Thomas L Griffiths. Exploring the structure of mental representations by implementing computer algorithms with people. In *Cognitive Modeling in Perception and Memory*, pp. 212–228. Psychology Press, 2015.

Adam N Sanborn, Thomas L Griffiths, and Richard M Shiffrin. Uncovering mental representations with markov chain monte carlo. *Cognitive psychology*, 60(2):63–106, 2010.

David R Shanks, Richard J Tunney, and John D McCarthy. A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, 15(3):233–250, 2002.

Nikhil Singh, Manuel Cherep, and Pattie Maes. Discovering and steering interpretable concepts in large generative music models. *arXiv preprint arXiv:2505.18186*, 2025.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders. In *Mechanistic Interpretability for Vision at CVPR 2025 (Non-proceedings Track)*, 2025.

Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pp. 1701–1728, 1994.

Eline Van Geert and Nori Jacoby. Using gibbs sampling with people to characterize perceptual and aesthetic evaluations in multidimensional visual stimulus space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Pol Van Rijn, Silvan Mertes, Dominik Schiller, Peter Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. Exploring emotional prototypes in a high dimensional tts latent space. *arXiv preprint arXiv:2105.01891*, 2021.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter Harrison, Elisabeth André, and Nori Jacoby. Voiceme: Personalized voice generation in tts. *arXiv preprint arXiv:2203.15379*, 2022.

Alexander Sasha Vezhnevets, John P Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A Duéñez-Guzmán, William A Cunningham, Simon Osindero, Danny Karmon, and Joel Z Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.

Nir Vulkan. An economist's perspective on probability matching. *Journal of economic surveys*, 14 (1):101–118, 2000.

Jian-Qiao Zhu and Thomas L Griffiths. Eliciting the priors of large language models using iterated in-context learning. *arXiv preprint arXiv:2406.01860*, 2024.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L Griffiths. Recovering mental representations from large language models with markov chain monte carlo. *arXiv preprint arXiv:2401.16657*, 2024.

A Examples

We show qualitative examples of the recovered priors across the four domains in Figure 5. Each column corresponds to a (target, model) pair.



Figure 5: Example of recovered priors across the four domains (faces, affordances, aesthetics, authenticity), with one domain per model.

B Sliders

For each domain, Figures 6 to 9 show qualitative sweeps of the $d = 10$ sliders selected from the $n = 64$ SliderSpace candidates: each starts from the same seed image and progressively increases the slider's value while holding the others fixed. The sweeps illustrate what each axis controls and that the selected sliders produce coherent, monotone semantic changes when varied in isolation.

C Comparison to Explicit Priors

The IMH sampler is designed to recover a joint implicit prior, i.e. it observes which images a model prefers during a sequence of adaptive pairwise comparisons and turns those choices into samples from each model's target-conditioned distribution over the slider space. A simpler alternative is to query the model explicitly on a fixed set of slider sweeps. We implement and compare two such baselines, described below.



Figure 6: Examples for the face sliders trained with SliderSpace. The top image is the original, progressively steered towards the bottom. Labels are automatically generated (see Appendix D).

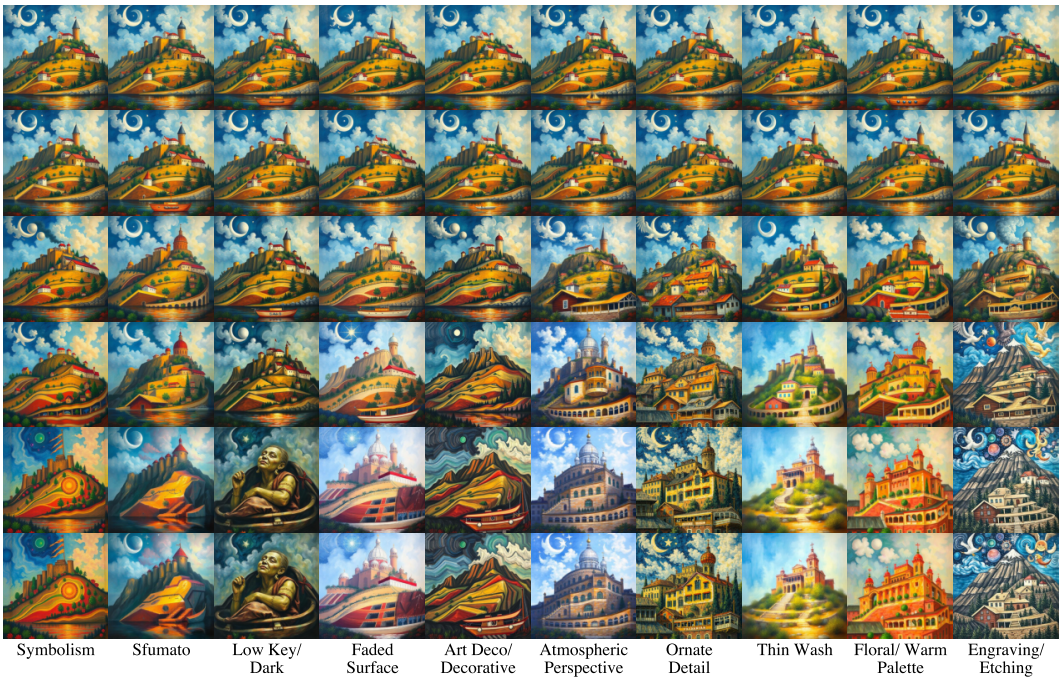


Figure 7: Examples for the aesthetics sliders trained with SliderSpace. The top image is the original, progressively steered towards the bottom. Labels are automatically generated (see Appendix D).

First, for each domain, we construct direct slider sweeps. For each chain, we first sample a base slider vector and a fixed generation seed. For each slider, we then vary only it while holding the other nine slider values and the generation seed fixed. Thus, within a sweep, visible changes are



Figure 8: Examples for the affordances sliders trained with SliderSpace. The top image is the original, progressively steered towards the bottom. Labels are automatically generated (see Appendix D).

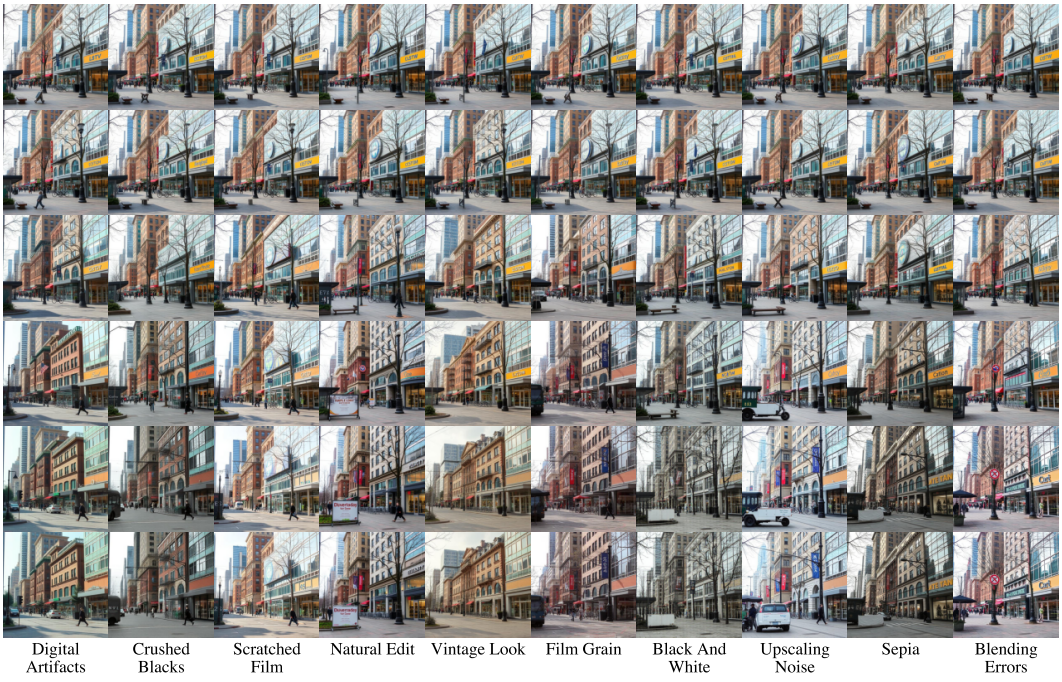


Figure 9: Examples for the authenticity sliders trained with SliderSpace. The top image is the original, progressively steered towards the bottom. Labels are automatically generated (see Appendix D).

causally attributable to one slider. This design gives the explicit priors baseline the cleanest possible one-dimensional evidence about each control.

Rating baseline. In the rating baseline, the model sees one image at a time and assigns a score from 1 to 10 indicating how well the image matches the target. For a given target and slider, each image contributes its slider value and its rating. We normalize the ratings into weights and compute weighted posterior summaries over the slider coordinate. This produces a one-dimensional explicit posterior for every target–slider pair and is a direct analog of asking the model "how much does this isolated slider setting match the target concept?"

Pairwise Bradley–Terry baseline. Here the model instead gets two images from the same slider sweep and chooses which image better matches the target (with randomized pair order). For each target and slider, we fit a Bradley–Terry model in which the latent utility of an image is linear in the value of the varied slider:

$$\Pr(a \succ b) = \frac{\exp(\beta z_a)}{\exp(\beta z_a) + \exp(\beta z_b)} \tag{5}$$

The fitted utility βz defines a set of importance weights over the swept slider values, from which we compute weighted posterior medians and intervals.

Magnitude and rank disagreement. Then, for each target, model, domain, and slider, we compare the explicit posterior median to the IMH posterior median. The signed difference gives the direction and magnitude of disagreement. Because the procedures can differ in calibration, we also compare centered profiles, subtracting the mean posterior median across the sliders before computing differences. Finally, we rank sliders within each target by posterior median and compute rank displacement between explicit and IMH rankings, to study whether these methods identify different sliders as important. Results are shown in Figures 10 and 11 for the top-10 disagreements by each method.

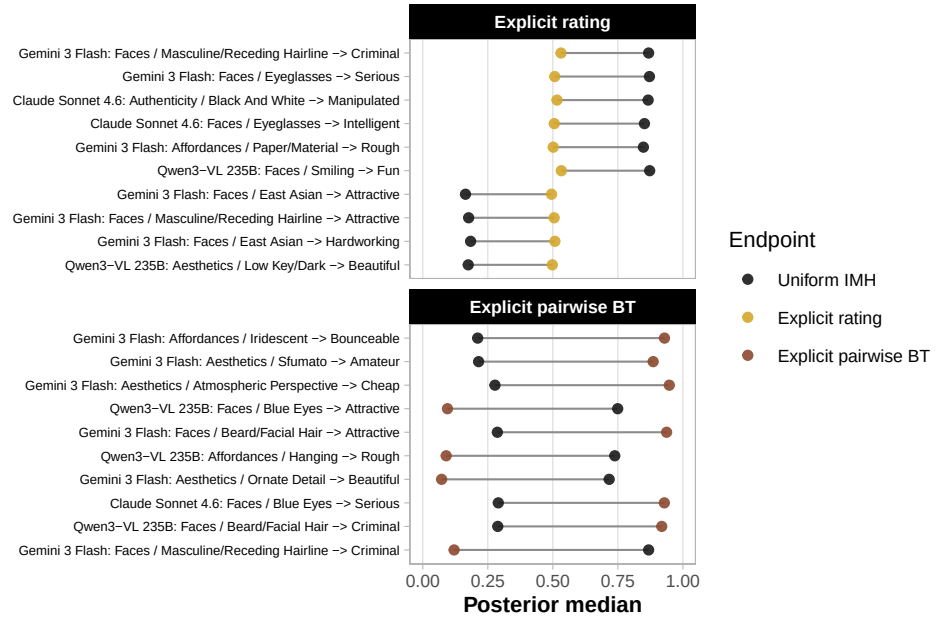


Figure 10: Magnitude of misalignment between explicit-prior and implicit Gibbs representations. Bars summarize how strongly each model–domain setting departs from the explicit slider coordinate system, aggregating across sliders and chains.

The explicit baselines are by construction insensitive to posterior dependence among sliders. IMH, by contrast, samples the joint distribution and can show that a model prefers combinations of features beyond independent marginal changes. To quantify this difference, we estimate slider–slider correlations from IMH draws and summarize each target by the mean absolute off-diagonal correlation. We then ask whether targets with stronger slider interactions are also the targets where explicit and implicit priors disagree most. Figure 12 shows the off-diagonal correlations ignored by the explicit methods, and Figure 13 shows that misalignment and interaction strength are correlated with Spearman $\rho = 0.63, p = 0.02$.

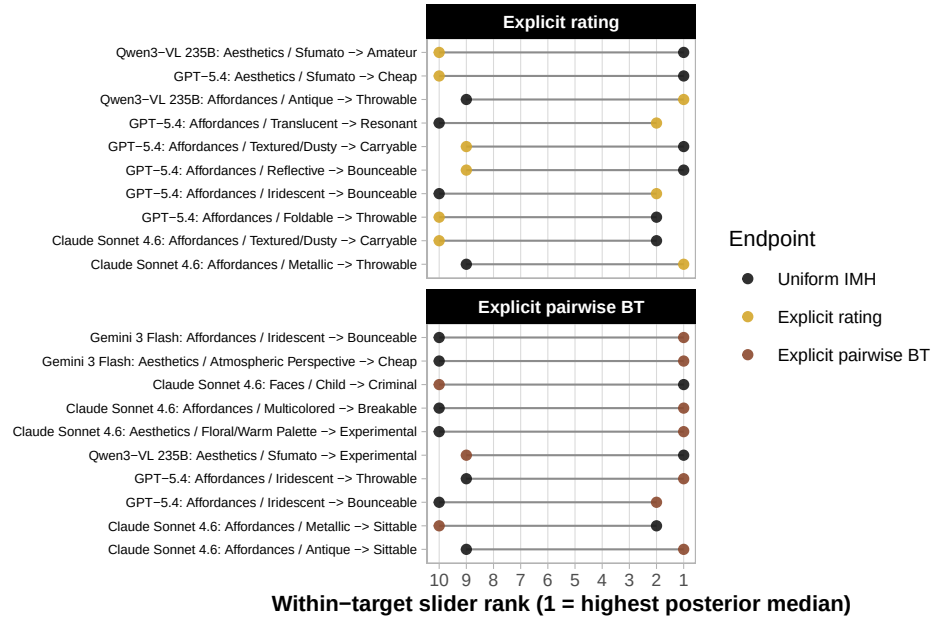


Figure 11: Rank misalignment between explicit-prior sliders and implicit Gibbs sliders across domains and models. For each setting, we compare the ordering induced by explicit slider values to the ordering recovered from implicit samples, with higher values indicating larger disagreement.

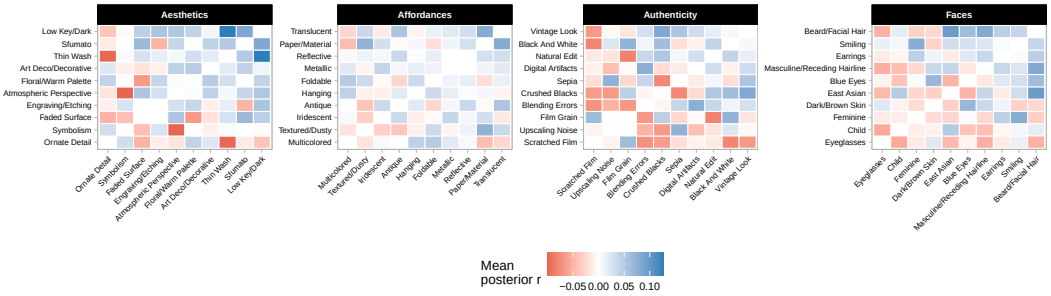


Figure 12: Correlation structure among learned slider dimensions. Each condition reports the empirical correlation between a pair of slider coordinates across sampled images, revealing whether sliders vary independently or are entangled. Explicit methods assume all off-diagonal entries are zero.

Informativeness. Finally, we quantify how much each elicitation procedure moves beyond the uninformative baseline of a uniform slider prior. For this, we compute the KL divergence from the recovered slider posterior to the uniform prior. For IMH, this is estimated from binned post-burn-in posterior draws. For the explicit baselines, it is computed as the discrete KL divergence of the induced weights over their fixed sampled design. This quantity should be interpreted as posterior concentration (however, note that a concentrated posterior may still be wrong if the elicitation procedure is biased or misspecified). Nevertheless, it estimates the value added over assuming a blank uniform posterior.

D Auto-Interpretability Pipeline

We interpret each latent slider by matching image trajectories to a vocabulary of visual descriptors. The base candidate label set was generated by Claude 5 Fable, given high-level descriptions of the task and application domains. This resulted in a total of 867 candidate labels, across Affordances

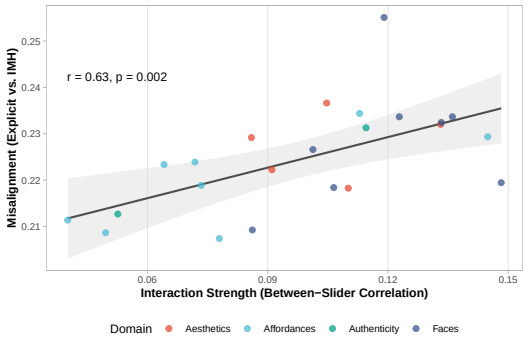


Figure 13: Relationship between slider interaction strength (X-axis) and explicit-implicit misalignment (Y-axis), showing that misalignments are larger as sliders are more correlated.

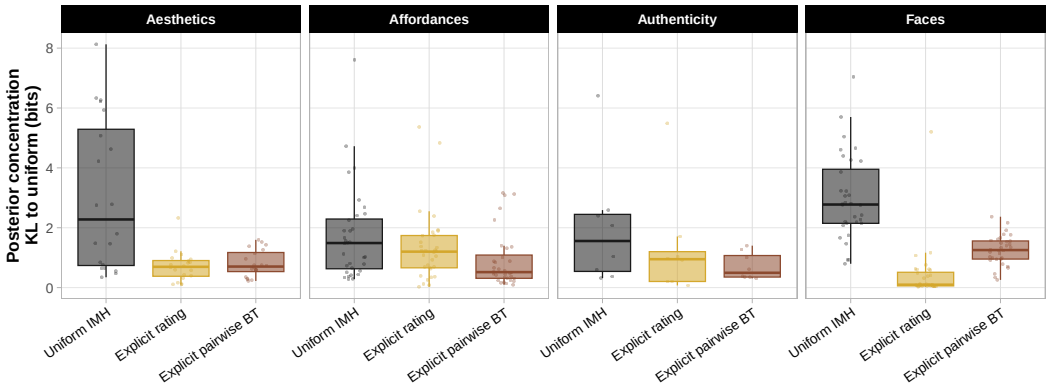


Figure 14: Posterior concentration relative to a uniform slider prior. Each point is a model–target combination, and the y-axis reports KL divergence from the recovered posterior to the corresponding uniform baseline. Larger values indicate that the elicitation procedure recovered a posterior that departs more strongly from an uninformative uniform prior.

(227), Art/Aesthetics (206), Authenticity (173), and Faces (261). Candidate labels and sampled images are embedded into a shared multimodal embedding space using `Gemini Embedding 2`.

Image sampling. We use a variant of the explicit-prior baseline images (described in Appendix C) where instead all sliders are set to 0 except the focal slider for a given chain, and we run 5 chains (here, different seeds/starting images) for each slider. We sort images within each (domain, chain, slider) by the corresponding slider value.

Slider scoring. For each slider j, chain c, and candidate label ℓ, we compute image–text similarities:

$$s_{ic\ell} = \langle f_{\text{img}}(x_{ic}), f_{\text{text}}(\ell) \rangle \tag{6}$$

To reduce generic labels, similarities are residualized by subtracting the per-image mean similarity across candidate labels:

$$\tilde{s}_{ic\ell} = s_{ic\ell} - \frac{1}{|\mathcal{L}|} \sum_{\ell' \in \mathcal{L}} s_{ic\ell'} \tag{7}$$

We score labels using endpoint contrast. Let L and H denote the low- and high-slider endpoint sets, respectively. By default, each endpoint contains the outer 25% of samples along the slider. The per-chain endpoint score is:

$$S_{cj\ell} = \max\left(0,\ \frac{1}{|H|} \sum_{i \in H} \tilde{s}_{ic\ell} - \frac{1}{|L|} \sum_{i \in L} \tilde{s}_{ic\ell}\right) \tag{8}$$

Thus, labels are rewarded when their image-text similarity increases from the low end to the high end of the slider, and negative contrasts are zeroed out.

Chain weighting. Some chains produce more visually salient slider sweeps than others, because the generation seed and the nuisance slider settings can make the same control more or less visible. We therefore weight chains by the amount of image-embedding movement induced by the current slider. However, we need a way to estimate the overall movement in a coherent direction.

For each chain c and slider j, let $\mathbf{e}_{ic} = f_{\text{img}}(x_{ic})$ be the image embedding for sample i, and let z_{icj} be its value on the current slider. We fit:

$$\mathbf{e}_{ic} = \boldsymbol{\alpha}_{cj} + z_{icj}\boldsymbol{\beta}_{cj} + \boldsymbol{\varepsilon}_{ic} \tag{9}$$

and use the slope norm

$$m_{cj} = \left\| \hat{\boldsymbol{\beta}}_{cj} \right\|_2 \tag{10}$$

as a chain-level movement score. These movement scores are normalized within slider and converted into softmax-squared aggregation weights, so chains with stronger coherent embedding movement contribute more to the final label score.

Global label assignment. To avoid duplicate final labels within a domain, we solve a domain-level assignment problem. For each domain, we build a slider $\times$ label score matrix using the top candidate labels per slider, using the Hungarian algorithm to select one label per slider and uses each canonical label at most once within the domain.

LM refinement. Finally, we perform a post-hoc label refinement step over the assigned label and top runner-up labels, with five candidate labels per slider by default. The refinement prompt includes the endpoint scores. The refinement is performed in a single call to `Gemini 3.1 Flash Lite`. The model is instructed to summarize compatible candidate labels, resolve conflicts using scores and visual plausibility, and enforce semantic uniqueness within each domain. The refined label is used in downstream outputs. Table 1 shows the results across all 40 sliders.

You are summarizing slider labels for an image analysis project.
For each slider, write one concise, visually plausible final label that summarizes the common visual concept across the candidate labels.
Candidate scores are nonnegative evidence strengths from the selected image-embedding scoring method; use them to resolve conflicts.
Do not simply copy the top-ranked label when several high-ranked labels express the same broader idea.
Combine compatible labels into a natural summary, for example {sad, unhappy, frowning, grumpy} -¿ Sad.
When candidate labels conflict, defer to labels that are more visually reasonable, general enough to use as a slider interpretation, and supported by stronger scores.
Before returning labels, review all sliders within each domain together and enforce semantic uniqueness across that domain.
Final labels must be semantically unique within each domain, not just textually different.
Do not assign near-duplicate concepts to separate sliders, for example Woman/Feminine/Female, Man/Masculine/Male, Happy/Smiling, Old/Aged, Young/Youthful, Dark/Black, or Large/Big.
If two sliders point to overlapping concepts, keep the concept only for the slider where it is most strongly and reasonably supported, then choose a distinct runner-up interpretation for the other slider.
Avoid overly narrow/specific or awkward labels unless the candidate set clearly supports them.
Use a brief noun phrase or adjective phrase, not a sentence.
Return only JSON with this schema:

```
{
    "labels": [
        {
```

```
            "domain": "faces",
            "slider_idx": 0,
            "final_label": "Sad/Frowning"
        }
    ]
}
```


Table 1: Top candidate labels and post-hoc refined labels used for slider interpretation.

	§	Final	Top 1	Top 2	Top 3	Top 4	Top 5
Affordances	v1	Foldable	Folded	Foldable	Metallic	Pristine	Bronze
	v2	Multicolored	Multicolored	Stained Glass	Red	Frayed	Hanging
	v3	Translucent	White	Cream	Frosted Glass	Translucent	Beige
	v4	Metallic	Bronze	Gold Colored	Copper Colored	Antique	Copper
	v5	Antique	Antique	Dusty	Tarnished	Old	Hammered
	v6	Textured/Dusty	Gray	Dusty	Granite	Plaster	Silver Colored
	v7	Hanging	Hangable	Multicolored	Stained Glass	Hanging	Pink
	v8	Paper/Material	Foldable	Folded	Paper	Cardboard	Antique
	v9	Reflective	Glossy	Bronze	Reflective	Chrome	Copper Colored
	v10	Iridescent	Translucent	Iridescent	Pink	Frosted Glass	Transparent
Aesthetics	v1	Symbolism	Pre-Raphaelite	Symbolic Imagery	Symbolism	Allegorical	Analogous Colors
	v2	Sfumato	Sfumato	Digital Painting	Mythological Scene	Dramatic Mood	Drybrush
	v3	Low Key/Dark	Low Key	Dark Tonality	Gothic	Mythological Scene	Tenebrism
	v4	Faded Surface	Faded Surface	Architectural Subject	Atmospheric Perspective	Thin Wash	Futurism
	v5	Art Deco/Decorative	Art Deco	Decorative Pattern	Futurism	Distorted Forms	Linocut
	v6	Atmospheric Perspective	Atmospheric Perspective	Tonalism	Seascape	High Key	Luminous
	v7	Ornate Detail	Ornate Detail	Dense Detail	Encaustic	Art Nouveau	Gold Leaf
	v8	Thin Wash	Thin Wash	Atmospheric Perspective	Tonalism	Architectural Subject	Underpainting Visible
	v9	Floral/Warm Palette	Floral Subject	Warm Palette	Architectural Subject	Encaustic	Pre-Raphaelite
	v10	Engraving/Etching	Dense Detail	Engraving	Etching	Ornate Detail	Pen And Ink
Authenticity	v1	Digital Artifacts	AI Generated Look	Face Swap Artifacts	Photorealistic Render	Scratched Film	Green Tint
	v2	Crushed Blacks	Crushed Blacks	Scratched Film	Black And White	Smoky	Teal And Orange Grade
	v3	Scratched Film	Scratched Film	Sepia	Reflections Present	Faded Film	Edge Halos
	v4	Natural Edit	Natural Edit	Vintage Look	Black And White	Dramatic Lighting	Cool White Balance
	v5	Vintage Look	Vintage Look	Sepia	Object Removal Traces	Incorrect Scale	Faded Film
	v6	Film Grain	Film Grain	Faded Film	Rephotographed Screen	Scratched Film	Scanned Print
	v7	Black And White	Black And White	Sepia	Rainy	Faded Film	Dodging And Burning
	v8	Upscaling Noise	Upscaling Artifacts	Digital Noise	JPEG Artifacts	Edge Halos	Deepfake Blending Seams
	v9	Sepia	Sepia	Black And White	Crushed Blacks	Vintage Look	Dust Spots
	v10	Blending Errors	Hair Blending Errors	Black And White	Purple Fringing	Selective Color Edit	Oversharpened
Faces	v1	Blue Eyes	Blue Eyes	Dry Skin	Green Eyes	Dyed Hair	Crow's Feet
	v2	Child	Child	East Asian Ethnicity	Unnatural Hair Color	Dreadlocks	Dyed Hair
	v3	East Asian	East Asian Ethnicity	Bloodshot Eyes	Freckled Skin	Monolid Eyes	Southeast Asian Ethnicity
	v4	Dark/Brown Skin	Dark Brown Skin	Brown Skin	South Asian Ethnicity	Tan Skin	Black Ethnicity
	v5	Feminine	Earrings	Female Presenting	Feminine	No Makeup	Ponytail

v6	Smiling	Smiling	East Asian Ethnicity	Southeast Asian Ethnicity	Grinning	Smirking
v7	Receding Hairline	Male Presenting	Receding Hairline	Blue Eyes	Stubble	Crow's Feet
v8	Eyeglasses	Eyeglasses	Full Beard	Long Beard	Trimmed Beard	Stubble
v9	Earrings	Female Presenting	Feminine	Earrings	Southeast Asian Ethnicity	Pacific Islander Ethnicity
v10	Beard/Facial Hair	Stubble	Male Presenting	Long Beard	Receding Hairline	Full Beard

E Additional Diagnostics

The main posterior summaries are marginal summaries of sampled slider values. We add two additional predictive diagnostics that use the accept/reject decisions to test whether the sampled marginal density has explanatory power for future proposals. For each proposal, we extract the current value and proposed value along the dimension being updated. We then estimate the marginal density of post-burn-in chain states along that same dimension using a kernel density estimate on the logit-transformed slider value. A proposal is predicted to be accepted when the estimated density at the proposed value is at least as large as the estimated density at the current value (chance = 0.5).

Within-chain predictive check. First, we ask whether one part of a chain predicts held-out decisions from the same chain. For each chain and updated dimension, we split proposals into contiguous folds, fit the marginal density on the training folds, and evaluate accept/reject prediction on the held-out fold. High accuracy means the chain's own posterior mass is aligned with the model's later choices. Results, shown in Figure 15, point to consistent

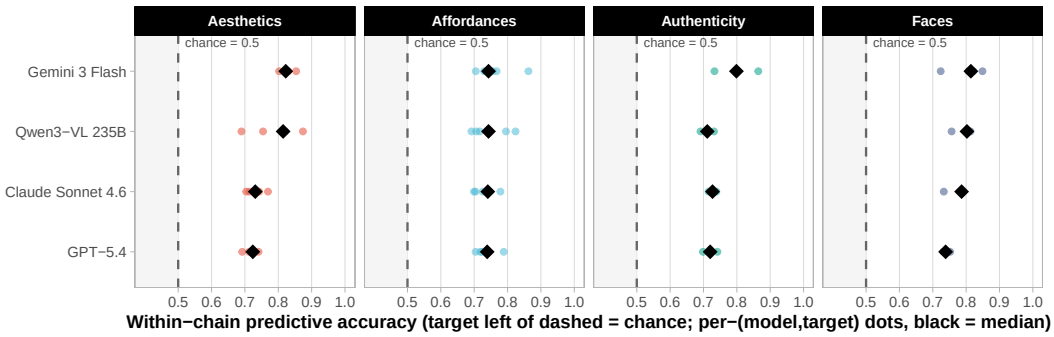


Figure 15: Within-chain predictive accuracy for slider-based models.

Cross-chain predictive check. We then apply a stricter cross-chain check. For each target and updated dimension, we fit the marginal density using all chains except one and evaluate proposals from the held-out chain to test whether different seeds recover compatible posterior structure. Results are shown in Figure 16.

E.1 Convergence

Table 2 reports per-(domain, model) convergence diagnostics aggregated over slider parameters, and Table 3 breaks the same chains down into the fraction of parameters that cross standard $\hat{R}$ and acceptance thresholds. *GPT-5.4* and *Claude Sonnet 4.6* produce the most-converged chains across all four domains (median split-$\hat{R} \in [1.02, 1.20]$); *Qwen3-VL 235B* and *Gemini 3 Flash* are systematically harder to mix, with median $\hat{R}$ reaching 1.6 on *Aesthetics*. *Affordances* mixes comparably well for all models: even *Qwen3-VL 235B* and *Gemini 3 Flash* sit near median $\hat{R} = 1.15$ there, well below their *Faces* and *Aesthetics* values. Parameter-level flags in Table 3 show that for *GPT-5.4* and *Claude Sonnet 4.6*, almost no parameters exceed $\hat{R} = 1.5$ in any domain ($\leq 1\%$), and none fall below 5%

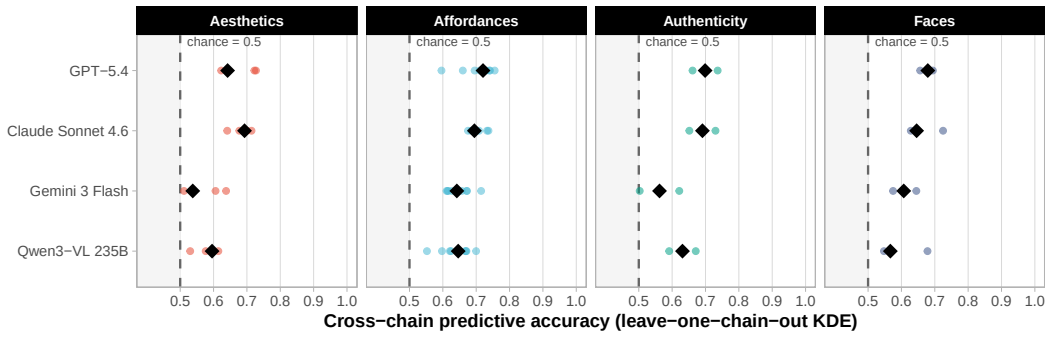


Figure 16: Cross-chain predictive accuracy for slider-based models.

acceptance. *Qwen3-VL 235B* and *Gemini 3 Flash* concentrate their poorly-mixing parameters in *Aesthetics* (50–58% above $\hat{R} = 1.5$ and 16–18% below 5% acceptance) and to a lesser extent *Faces* (22–32% above $\hat{R} = 1.5$); their *Affordances* chains, by contrast, rarely cross either threshold despite a high share above the looser $\hat{R} = 1.1$ cutoff. Acceptance rates align with these observations: *GPT-5.4* and *Claude Sonnet 4.6* accept the proposal 20–40% of the time on average while *Qwen3-VL 235B* and *Gemini 3 Flash* drop to 7–15% on the harder domains. This suggests that the latter two may have more decisive implicit priors that make a random proposal harder to satisfy.

Our experimental budget was chosen to prioritize breadth across models, targets, and domains rather than driving every individual chain to a stringent asymptotic convergence threshold. Consequently, $\hat{R}$ should be interpreted here primarily as a measure of estimation precision, in which conditions with high $\hat{R}$ provide less reliable estimates of posterior magnitude. Our main qualitative conclusions instead emphasize effects that replicate across independently initialized chains and, where applicable, across models. We report all convergence diagnostics for transparency.

Table 2: Per-(domain, model) convergence diagnostics, aggregated across the slider parameters of every target in the domain. `med. max` $\hat{R}$ is the median across targets of the worst-mixing slider's $\hat{R}$; `med.` $\hat{R}$ is the median across all (target, slider) pairs; `lag-1` is the mean lag-1 autocorrelation post-burn-in.

Domain	Model	med. max $\hat{R}$	worst $\hat{R}$	med. $\hat{R}$	lag-1
Aesthetics	Claude Sonnet 4.6	1.08	1.22	1.03	0.63
	GPT-5.4	1.09	1.29	1.06	0.69
	Qwen3-VL 235B	2.10	2.74	1.60	0.86
	Gemini 3 Flash	1.96	3.20	1.56	0.88
Authenticity	Claude Sonnet 4.6	1.11	1.20	1.06	0.65
	GPT-5.4	1.04	1.05	1.02	0.58
	Qwen3-VL 235B	1.28	1.45	1.08	0.76
	Gemini 3 Flash	1.66	2.07	1.39	0.81
Affordances	Claude Sonnet 4.6	1.08	1.20	1.03	0.60
	GPT-5.4	1.04	1.55	1.02	0.61
	Qwen3-VL 235B	1.31	2.10	1.15	0.77
	Gemini 3 Flash	1.33	2.44	1.15	0.79
Faces	Claude Sonnet 4.6	1.32	1.78	1.20	0.77
	GPT-5.4	1.25	1.42	1.11	0.77
	Qwen3-VL 235B	1.59	2.31	1.33	0.81
	Gemini 3 Flash	1.68	2.79	1.33	0.85

Table 3: Parameter-level diagnostic checks by perceptual domain and MLLM judge (Acc. = acceptance rate $\in [0, 1]$).

Domain	Model	$\hat{R} > 1.1$	$\hat{R} > 1.5$	Acc. $< 5\%$	med. Acc.
Aesthetics	Claude Sonnet 4.6	16%	0%	0%	0.36
Aesthetics	GPT-5.4	14%	0%	0%	0.28
Aesthetics	Qwen3-VL 235B	94%	50%	16%	0.08
Aesthetics	Gemini 3 Flash	100%	58%	18%	0.07
Authenticity	Claude Sonnet 4.6	15%	0%	0%	0.33
Authenticity	GPT-5.4	0%	0%	0%	0.41
Authenticity	Qwen3-VL 235B	35%	0%	0%	0.21
Authenticity	Gemini 3 Flash	85%	40%	5%	0.14
Affordances	Claude Sonnet 4.6	5%	0%	0%	0.37
Affordances	GPT-5.4	16%	1%	0%	0.42
Affordances	Qwen3-VL 235B	61%	12%	0%	0.21
Affordances	Gemini 3 Flash	61%	10%	0%	0.20
Faces	Claude Sonnet 4.6	62%	1%	0%	0.20
Faces	GPT-5.4	49%	0%	0%	0.23
Faces	Qwen3-VL 235B	81%	22%	0%	0.14
Faces	Gemini 3 Flash	90%	32%	0%	0.15

F Random Walk Metropolis-Hastings

In the main experiments, we drew the proposals independently from a $\mathrm{Uniform}([0, 1])$ distribution. However, this can also be done with a random walk by using a symmetric Gaussian proposal centered on the current value, reflected at the boundaries. Symmetry is required because the MLLM only provides a binary choice. The initial proposal standard deviation is $\sigma_0 = 0.2$. We adapt σ during the burn-in fraction of each chain (30% by default) via Robbins–Monro (Robbins & Monro, 1951) to target a desired acceptance rate (default 0.27 (Agrawal et al., 2023)); once burn-in ends, σ is frozen for the remainder of the chain.

In practice, the random-walk variant underperforms the uniform proposal of the main experiments, as the median $\hat{R}$ (see Table 4) frequently exceeds 1.5. The Robbins–Monro update often collapses σ to $\lesssim 0.1$ in search of the 0.27 target rate, but no σ achieves it. Large moves leave the high-π region and small moves still cross the local gradient sharply enough to be rejected, so acceptance plateaus at 0.05–0.15 and the chain barely moves. It is possible to adapt one scale per slider per chain, which might yield better results, but at the cost of a longer burn-in and thus more trials.

Table 4: Diagnostics for the random-walk variant on the faces experiment, aggregated across the $d = 10$ sliders per (agent, target). Each condition reports the chain-mean post-burn-in proposal scale σ followed by the median split-$\hat{R}$. Collapsed $\sigma \lesssim 0.1$ generally coincides with $\hat{R} > 2$, while stable $\sigma \in [0.4, 0.95]$ coincides with $\hat{R} \leq 1.5$.

Target	Qwen3-VL 235B	Claude Sonnet 4.6	Gemini 3 Flash	GPT-5.4
attractive	0.05 | 2.95	0.30 | 1.65	0.13 | 2.19	0.67 | 1.31
fun	0.03 | 4.67	0.18 | 1.87	0.15 | 2.11	0.27 | 1.43
intelligent	0.36 | 1.77	0.25 | 1.56	0.25 | 1.70	0.32 | 1.69
serious	0.02 | 4.36	0.46 | 1.60	0.19 | 2.24	0.43 | 1.38
trustworthy	0.48 | 1.17	0.06 | 3.44	0.08 | 3.02	0.52 | 1.35
youthful	0.08 | 2.59	0.10 | 2.04	0.02 | 7.28	0.34 | 1.44
criminal	0.13 | 2.56	0.67 | 1.19	0.07 | 3.38	0.85 | 1.13
hardworking	0.46 | 1.47	0.94 | 1.06	0.33 | 1.37	0.61 | 1.46

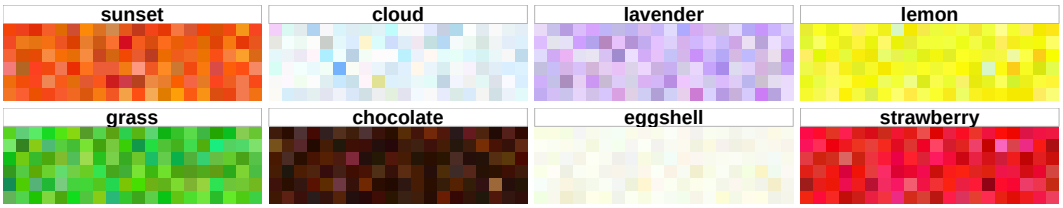
G Validation with Color Experiment

To verify that our Gibbs+Barker kernel correctly samples from the model's induced prior, we run a validation experiment in a low-dimensional space where ground truth is unambiguous: HSL colors. The state $z \in [0, 1]^3$ corresponds directly to the (hue, saturation, lightness) coordinates of a flat color tile, so the stimulus is fully determined by z without any generative model in between. We pick the same eight color-name targets as Harrison et al. (2020); Zhu et al. (2024)—*sunset*, *eggshell*, *lavender*, *chocolate*, *lemon*, *cloud*, *strawberry*, and *grass*—to compare against their human data. We run 10 independent chains of 2,000 trials each, using only *Qwen3-VL 235B*. Because the color associated with each target is human-recognizable, deviations between the recovered posterior mode and the canonical color flag kernel-level errors that would be hard to spot in higher-dimensional settings. This experiment is intentionally not a validation of our full pipeline with a complex generative model. Its sole purpose is to isolate Gibbs+Barker in a setting with known ground truth and verify that the procedure recovers the correct target distribution.

The recovered posteriors are visualized as 2D density overlays in Figure 17a. Figure 17b shows posterior samples for each named color. For every target, the central tendency falls within the canonical color associated with the name: saturated red-orange for *sunset*, saturated yellow for *lemon*, green for *grass*, light purple for *lavender*, dark red-brown for *chocolate*, near-white with high lightness for *cloud* and *eggshell*, and saturated red for *strawberry*. The kernel therefore recovers the correct region of the color space for every target, purely from binary preferences.



(a) 2D density overlays of the posterior over HSL coordinates for each target.



(b) Posterior samples rendered as color swatches for each named target.

Figure 17: Posterior color recovery for all eight named targets.

H Compute Resources

SliderSpace sliders were trained (see Appendix M for details) on a single NVIDIA H200 GPU (141 GB VRAM). Per domain, data generation and PCA over the 50,000 CLIP-encoded samples take approximately 3 hours, and slider training takes approximately 6 hours; total ≈9 hours/domain.

For the sampling experiments reported in the paper, we run FLUX.1-schnell inference on 40 parallel Amazon SageMaker endpoints, each backed by a single NVIDIA L40S GPU (48 GB VRAM). Each endpoint serves the same FLUX.1-schnell pipeline with the trained SliderSpace LoRAs, compiled and highly optimized to run inference at ∼0.7s per image.

I Posterior Summaries

Tables 5 to 8 show posterior summaries (medians with 50% credible intervals) for all conditions across the primary IMH experiments reported in the main paper.

Table 5: Posterior medians with 50% credible intervals for Affordances.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
Bounceable	Antique	0.40 [0.19, 0.68]	0.52 [0.27, 0.85]	0.39 [0.18, 0.66]	0.50 [0.24, 0.83]
	Foldable	0.53 [0.31, 0.77]	0.53 [0.24, 0.67]	0.48 [0.16, 0.86]	0.49 [0.31, 0.77]
	Hanging	0.54 [0.29, 0.71]	0.54 [0.29, 0.74]	0.40 [0.20, 0.79]	0.40 [0.31, 0.78]
	Iridescent	0.47 [0.25, 0.74]	0.41 [0.15, 0.68]	0.21 [0.06, 0.49]	0.46 [0.14, 0.75]
	Metallic	0.47 [0.20, 0.70]	0.48 [0.23, 0.68]	0.34 [0.13, 0.53]	0.42 [0.16, 0.56]
	Multicolored	0.43 [0.25, 0.76]	0.50 [0.29, 0.76]	0.36 [0.11, 0.71]	0.45 [0.26, 0.78]
	Paper/Material	0.37 [0.19, 0.67]	0.49 [0.26, 0.67]	0.33 [0.10, 0.63]	0.22 [0.09, 0.69]
	Reflective	0.53 [0.26, 0.75]	0.58 [0.30, 0.82]	0.49 [0.23, 0.76]	0.31 [0.15, 0.65]
	Textured/Dusty	0.56 [0.25, 0.75]	0.46 [0.18, 0.73]	0.58 [0.37, 0.79]	0.46 [0.28, 0.74]
	Translucent	0.64 [0.34, 0.85]	0.49 [0.26, 0.73]	0.53 [0.22, 0.72]	0.63 [0.32, 0.88]
Breakable	Antique	0.43 [0.21, 0.68]	0.45 [0.23, 0.70]	0.49 [0.27, 0.76]	0.43 [0.22, 0.75]
	Foldable	0.42 [0.21, 0.65]	0.42 [0.23, 0.67]	0.35 [0.22, 0.60]	0.35 [0.19, 0.62]
	Hanging	0.51 [0.27, 0.73]	0.60 [0.30, 0.82]	0.60 [0.29, 0.81]	0.50 [0.28, 0.77]
	Iridescent	0.44 [0.24, 0.63]	0.44 [0.23, 0.69]	0.37 [0.21, 0.55]	0.47 [0.24, 0.68]
	Metallic	0.47 [0.20, 0.73]	0.53 [0.27, 0.76]	0.70 [0.42, 0.84]	0.45 [0.24, 0.71]
	Multicolored	0.38 [0.16, 0.64]	0.53 [0.28, 0.77]	0.48 [0.27, 0.76]	0.45 [0.26, 0.71]
	Paper/Material	0.52 [0.24, 0.76]	0.53 [0.25, 0.79]	0.68 [0.37, 0.84]	0.51 [0.23, 0.75]
	Reflective	0.46 [0.23, 0.71]	0.51 [0.30, 0.73]	0.39 [0.19, 0.65]	0.52 [0.32, 0.74]
	Textured/Dusty	0.44 [0.21, 0.72]	0.45 [0.22, 0.67]	0.33 [0.16, 0.55]	0.49 [0.24, 0.81]
	Translucent	0.54 [0.28, 0.80]	0.42 [0.21, 0.68]	0.44 [0.19, 0.69]	0.45 [0.23, 0.67]
Carryable	Antique	0.50 [0.23, 0.75]	0.51 [0.24, 0.75]	0.55 [0.30, 0.72]	0.51 [0.25, 0.78]
	Foldable	0.48 [0.21, 0.73]	0.41 [0.20, 0.70]	0.47 [0.21, 0.80]	0.27 [0.15, 0.56]
	Hanging	0.45 [0.19, 0.72]	0.41 [0.20, 0.71]	0.36 [0.20, 0.66]	0.29 [0.13, 0.57]
	Iridescent	0.50 [0.28, 0.75]	0.45 [0.24, 0.68]	0.27 [0.12, 0.44]	0.49 [0.27, 0.67]
	Metallic	0.45 [0.21, 0.70]	0.46 [0.23, 0.69]	0.38 [0.16, 0.67]	0.40 [0.22, 0.66]
	Multicolored	0.50 [0.25, 0.78]	0.45 [0.21, 0.68]	0.39 [0.17, 0.69]	0.44 [0.22, 0.75]
	Paper/Material	0.45 [0.23, 0.71]	0.45 [0.21, 0.74]	0.64 [0.35, 0.86]	0.44 [0.16, 0.65]
	Reflective	0.50 [0.24, 0.75]	0.52 [0.23, 0.78]	0.44 [0.21, 0.74]	0.42 [0.21, 0.70]
	Textured/Dusty	0.54 [0.29, 0.80]	0.53 [0.28, 0.79]	0.44 [0.19, 0.63]	0.45 [0.23, 0.78]
	Translucent	0.57 [0.33, 0.79]	0.44 [0.21, 0.67]	0.29 [0.16, 0.46]	0.34 [0.16, 0.49]
Resonant	Antique	0.50 [0.24, 0.74]	0.49 [0.23, 0.72]	0.37 [0.16, 0.60]	0.43 [0.17, 0.72]
	Foldable	0.56 [0.28, 0.78]	0.51 [0.28, 0.73]	0.52 [0.22, 0.80]	0.48 [0.22, 0.71]
	Hanging	0.53 [0.28, 0.75]	0.55 [0.30, 0.76]	0.61 [0.46, 0.82]	0.59 [0.34, 0.81]
	Iridescent	0.47 [0.24, 0.72]	0.39 [0.19, 0.64]	0.25 [0.10, 0.52]	0.38 [0.16, 0.66]
	Metallic	0.50 [0.28, 0.78]	0.51 [0.25, 0.75]	0.57 [0.34, 0.81]	0.45 [0.29, 0.72]
	Multicolored	0.47 [0.23, 0.75]	0.55 [0.29, 0.83]	0.40 [0.17, 0.80]	0.58 [0.29, 0.75]
	Paper/Material	0.52 [0.28, 0.74]	0.58 [0.32, 0.82]	0.60 [0.35, 0.83]	0.52 [0.26, 0.76]
	Reflective	0.50 [0.25, 0.75]	0.51 [0.23, 0.77]	0.48 [0.35, 0.75]	0.60 [0.32, 0.73]
	Textured/Dusty	0.46 [0.20, 0.74]	0.36 [0.16, 0.61]	0.16 [0.10, 0.38]	0.26 [0.13, 0.51]
	Translucent	0.49 [0.24, 0.71]	0.34 [0.16, 0.58]	0.26 [0.11, 0.46]	0.35 [0.14, 0.62]
Rough	Antique	0.60 [0.36, 0.84]	0.47 [0.27, 0.75]	0.49 [0.30, 0.82]	0.80 [0.41, 0.89]
	Foldable	0.46 [0.23, 0.77]	0.49 [0.30, 0.78]	0.66 [0.32, 0.84]	0.35 [0.14, 0.70]
	Hanging	0.48 [0.24, 0.75]	0.36 [0.24, 0.72]	0.69 [0.55, 0.91]	0.74 [0.24, 0.83]
	Iridescent	0.58 [0.32, 0.80]	0.38 [0.20, 0.60]	0.60 [0.18, 0.83]	0.53 [0.35, 0.74]
	Metallic	0.60 [0.38, 0.84]	0.70 [0.46, 0.85]	0.76 [0.48, 0.88]	0.63 [0.42, 0.78]
	Multicolored	0.58 [0.32, 0.83]	0.57 [0.32, 0.83]	0.72 [0.55, 0.94]	0.57 [0.32, 0.90]
	Paper/Material	0.57 [0.28, 0.79]	0.66 [0.20, 0.86]	0.85 [0.59, 0.95]	0.65 [0.43, 0.84]
	Reflective	0.60 [0.31, 0.79]	0.47 [0.22, 0.77]	0.68 [0.30, 0.81]	0.64 [0.47, 0.85]
	Textured/Dusty	0.53 [0.26, 0.77]	0.40 [0.22, 0.60]	0.28 [0.08, 0.44]	0.47 [0.23, 0.58]
	Translucent	0.44 [0.20, 0.65]	0.29 [0.14, 0.53]	0.35 [0.16, 0.42]	0.31 [0.13, 0.54]
Sittable	Antique	0.44 [0.17, 0.74]	0.53 [0.30, 0.77]	0.52 [0.28, 0.76]	0.43 [0.21, 0.74]
	Foldable	0.50 [0.21, 0.72]	0.45 [0.25, 0.71]	0.57 [0.25, 0.75]	0.68 [0.43, 0.84]
	Hanging	0.43 [0.19, 0.73]	0.54 [0.26, 0.78]	0.53 [0.25, 0.81]	0.73 [0.44, 0.87]

Continued on next page

Table 5: Posterior medians with 50% credible intervals for Affordances.

target	agent parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
	Iridescent	0.48 [0.32, 0.74]	0.50 [0.24, 0.73]	0.55 [0.25, 0.79]	0.40 [0.20, 0.64]
	Metallic	0.63 [0.33, 0.76]	0.52 [0.29, 0.74]	0.54 [0.26, 0.70]	0.55 [0.28, 0.74]
	Multicolored	0.50 [0.35, 0.73]	0.52 [0.29, 0.75]	0.61 [0.44, 0.86]	0.74 [0.45, 0.88]
	Paper/Material	0.49 [0.26, 0.66]	0.53 [0.27, 0.81]	0.39 [0.17, 0.72]	0.54 [0.21, 0.83]
	Reflective	0.52 [0.26, 0.68]	0.53 [0.26, 0.77]	0.58 [0.36, 0.81]	0.58 [0.36, 0.75]
	Textured/Dusty	0.57 [0.25, 0.71]	0.51 [0.26, 0.74]	0.62 [0.43, 0.82]	0.56 [0.30, 0.74]
	Translucent	0.64 [0.31, 0.86]	0.45 [0.25, 0.70]	0.48 [0.26, 0.72]	0.70 [0.41, 0.84]
Squeezable	Antique	0.46 [0.17, 0.73]	0.48 [0.23, 0.73]	0.55 [0.33, 0.90]	0.61 [0.31, 0.78]
	Foldable	0.40 [0.17, 0.68]	0.38 [0.17, 0.62]	0.39 [0.24, 0.59]	0.34 [0.17, 0.60]
	Hanging	0.57 [0.27, 0.76]	0.54 [0.28, 0.82]	0.39 [0.25, 0.74]	0.70 [0.40, 0.89]
	Iridescent	0.61 [0.34, 0.85]	0.54 [0.29, 0.80]	0.57 [0.39, 0.92]	0.62 [0.40, 0.88]
	Metallic	0.45 [0.23, 0.68]	0.45 [0.23, 0.70]	0.38 [0.20, 0.60]	0.35 [0.16, 0.59]
	Multicolored	0.56 [0.25, 0.78]	0.50 [0.28, 0.76]	0.52 [0.32, 0.68]	0.66 [0.32, 0.89]
	Paper/Material	0.41 [0.24, 0.67]	0.48 [0.26, 0.75]	0.51 [0.14, 0.75]	0.40 [0.21, 0.73]
	Reflective	0.44 [0.18, 0.72]	0.44 [0.25, 0.74]	0.46 [0.31, 0.79]	0.47 [0.20, 0.75]
	Textured/Dusty	0.55 [0.35, 0.77]	0.51 [0.27, 0.73]	0.63 [0.40, 0.83]	0.53 [0.31, 0.79]
	Translucent	0.62 [0.38, 0.87]	0.52 [0.25, 0.74]	0.74 [0.42, 0.86]	0.67 [0.37, 0.82]
Throwable	Antique	0.48 [0.22, 0.77]	0.52 [0.22, 0.76]	0.47 [0.26, 0.70]	0.38 [0.23, 0.71]
	Foldable	0.37 [0.16, 0.61]	0.51 [0.23, 0.73]	0.32 [0.12, 0.54]	0.32 [0.17, 0.42]
	Hanging	0.51 [0.26, 0.74]	0.50 [0.27, 0.74]	0.44 [0.20, 0.81]	0.49 [0.17, 0.82]
	Iridescent	0.46 [0.23, 0.75]	0.45 [0.21, 0.68]	0.32 [0.14, 0.60]	0.48 [0.17, 0.61]
	Metallic	0.41 [0.21, 0.69]	0.46 [0.21, 0.68]	0.36 [0.12, 0.53]	0.38 [0.17, 0.65]
	Multicolored	0.53 [0.26, 0.74]	0.48 [0.23, 0.76]	0.54 [0.30, 0.83]	0.57 [0.27, 0.86]
	Paper/Material	0.42 [0.17, 0.72]	0.48 [0.24, 0.70]	0.60 [0.35, 0.76]	0.66 [0.37, 0.90]
	Reflective	0.48 [0.27, 0.73]	0.48 [0.22, 0.74]	0.50 [0.26, 0.78]	0.66 [0.38, 0.88]
	Textured/Dusty	0.49 [0.25, 0.77]	0.51 [0.27, 0.74]	0.40 [0.20, 0.61]	0.40 [0.18, 0.64]
	Translucent	0.64 [0.31, 0.84]	0.43 [0.19, 0.70]	0.29 [0.11, 0.50]	0.51 [0.18, 0.86]

Table 6: Posterior medians with 50% credible intervals for Art.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
Amateur	Art Deco/Decorative	0.50 [0.23, 0.78]	0.45 [0.21, 0.74]	0.75 [0.25, 0.94]	0.56 [0.24, 0.88]
	Atmospheric Perspective	0.51 [0.28, 0.79]	0.59 [0.30, 0.84]	0.59 [0.16, 0.84]	0.58 [0.32, 0.83]
	Engraving/Etching	0.49 [0.21, 0.79]	0.42 [0.20, 0.71]	0.44 [0.22, 0.72]	0.48 [0.23, 0.73]
	Faded Surface	0.47 [0.21, 0.72]	0.51 [0.26, 0.77]	0.61 [0.22, 0.86]	0.55 [0.20, 0.71]
	Floral/Warm Palette	0.54 [0.28, 0.78]	0.63 [0.34, 0.86]	0.60 [0.23, 0.90]	0.54 [0.28, 0.74]
	Low Key/Dark	0.47 [0.25, 0.70]	0.41 [0.22, 0.73]	0.49 [0.14, 0.69]	0.54 [0.25, 0.79]
	Ornate Detail	0.47 [0.25, 0.70]	0.45 [0.23, 0.70]	0.40 [0.17, 0.68]	0.51 [0.23, 0.81]
	Sfumato	0.48 [0.23, 0.71]	0.54 [0.24, 0.75]	0.21 [0.13, 0.37]	0.67 [0.30, 0.88]
	Symbolism	0.47 [0.22, 0.77]	0.56 [0.26, 0.78]	0.53 [0.22, 0.75]	0.48 [0.19, 0.71]
	Thin Wash	0.57 [0.25, 0.81]	0.50 [0.29, 0.75]	0.30 [0.20, 0.74]	0.49 [0.25, 0.84]
Beautiful	Art Deco/Decorative	0.48 [0.24, 0.70]	0.38 [0.24, 0.63]	0.27 [0.10, 0.52]	0.34 [0.16, 0.59]
	Atmospheric Perspective	0.59 [0.28, 0.75]	0.39 [0.21, 0.64]	0.42 [0.31, 0.65]	0.53 [0.19, 0.68]
	Engraving/Etching	0.53 [0.23, 0.76]	0.39 [0.20, 0.74]	0.37 [0.22, 0.53]	0.46 [0.26, 0.74]
	Faded Surface	0.48 [0.14, 0.76]	0.38 [0.19, 0.58]	0.42 [0.10, 0.78]	0.45 [0.32, 0.76]
	Floral/Warm Palette	0.55 [0.37, 0.82]	0.39 [0.18, 0.63]	0.32 [0.13, 0.57]	0.22 [0.09, 0.45]
	Low Key/Dark	0.48 [0.32, 0.76]	0.32 [0.17, 0.66]	0.23 [0.06, 0.41]	0.17 [0.07, 0.46]
	Ornate Detail	0.47 [0.31, 0.63]	0.58 [0.30, 0.79]	0.72 [0.46, 0.79]	0.63 [0.40, 0.72]
	Sfumato	0.42 [0.34, 0.72]	0.28 [0.14, 0.52]	0.28 [0.21, 0.41]	0.26 [0.13, 0.61]

Continued on next page

Table 6: Posterior medians with 50% credible intervals for Art.

target	agent parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
	Symbolism	0.43 [0.27, 0.70]	0.40 [0.21, 0.70]	0.31 [0.17, 0.56]	0.42 [0.19, 0.68]
	Thin Wash	0.51 [0.42, 0.69]	0.45 [0.28, 0.67]	0.55 [0.39, 0.72]	0.60 [0.29, 0.76]
Cheap	Art Deco/Decorative	0.55 [0.26, 0.73]	0.42 [0.23, 0.72]	0.34 [0.19, 0.71]	0.56 [0.29, 0.79]
	Atmospheric Perspective	0.61 [0.36, 0.84]	0.52 [0.23, 0.79]	0.28 [0.11, 0.74]	0.67 [0.22, 0.88]
	Engraving/Etching	0.50 [0.23, 0.72]	0.45 [0.22, 0.71]	0.40 [0.22, 0.73]	0.28 [0.13, 0.75]
	Faded Surface	0.54 [0.30, 0.81]	0.53 [0.32, 0.78]	0.39 [0.14, 0.68]	0.46 [0.18, 0.85]
	Floral/Warm Palette	0.50 [0.29, 0.74]	0.48 [0.23, 0.75]	0.60 [0.31, 0.87]	0.72 [0.45, 0.91]
	Low Key/Dark	0.39 [0.23, 0.70]	0.51 [0.23, 0.77]	0.32 [0.11, 0.76]	0.57 [0.16, 0.75]
	Ornate Detail	0.49 [0.28, 0.75]	0.47 [0.23, 0.71]	0.58 [0.41, 0.86]	0.62 [0.24, 0.88]
	Sfumato	0.44 [0.24, 0.67]	0.53 [0.29, 0.79]	0.47 [0.18, 0.76]	0.50 [0.15, 0.78]
	Symbolism	0.45 [0.25, 0.74]	0.52 [0.28, 0.78]	0.47 [0.14, 0.71]	0.47 [0.14, 0.77]
	Thin Wash	0.58 [0.31, 0.84]	0.51 [0.22, 0.77]	0.52 [0.27, 0.74]	0.41 [0.25, 0.77]
Expensive	Art Deco/Decorative	0.47 [0.27, 0.75]	0.46 [0.27, 0.65]	0.53 [0.23, 0.79]	0.66 [0.42, 0.82]
	Atmospheric Perspective	0.49 [0.22, 0.72]	0.33 [0.16, 0.58]	0.30 [0.15, 0.55]	0.36 [0.13, 0.50]
	Engraving/Etching	0.57 [0.30, 0.76]	0.65 [0.36, 0.84]	0.52 [0.28, 0.75]	0.49 [0.45, 0.78]
	Faded Surface	0.53 [0.35, 0.77]	0.54 [0.34, 0.79]	0.33 [0.15, 0.54]	0.43 [0.15, 0.66]
	Floral/Warm Palette	0.59 [0.36, 0.81]	0.54 [0.30, 0.76]	0.36 [0.24, 0.62]	0.39 [0.11, 0.54]
	Low Key/Dark	0.48 [0.32, 0.72]	0.54 [0.30, 0.77]	0.45 [0.30, 0.84]	0.61 [0.33, 0.70]
	Ornate Detail	0.43 [0.23, 0.74]	0.48 [0.26, 0.72]	0.75 [0.42, 0.84]	0.68 [0.42, 0.92]
	Sfumato	0.35 [0.18, 0.60]	0.40 [0.20, 0.63]	0.36 [0.21, 0.49]	0.19 [0.10, 0.41]
	Symbolism	0.58 [0.30, 0.83]	0.59 [0.32, 0.79]	0.58 [0.46, 0.85]	0.48 [0.22, 0.89]
	Thin Wash	0.45 [0.22, 0.72]	0.47 [0.25, 0.68]	0.40 [0.28, 0.63]	0.47 [0.20, 0.78]
Experimental	Art Deco/Decorative	0.54 [0.33, 0.79]	0.51 [0.27, 0.73]	0.61 [0.22, 0.78]	0.33 [0.11, 0.59]
	Atmospheric Perspective	0.58 [0.30, 0.81]	0.51 [0.29, 0.71]	0.36 [0.19, 0.59]	0.52 [0.29, 0.71]
	Engraving/Etching	0.48 [0.22, 0.75]	0.42 [0.18, 0.66]	0.62 [0.33, 0.87]	0.44 [0.26, 0.80]
	Faded Surface	0.48 [0.23, 0.73]	0.37 [0.19, 0.66]	0.29 [0.06, 0.59]	0.37 [0.17, 0.50]
	Floral/Warm Palette	0.42 [0.20, 0.71]	0.49 [0.23, 0.78]	0.39 [0.20, 0.78]	0.62 [0.39, 0.79]
	Low Key/Dark	0.49 [0.27, 0.74]	0.49 [0.26, 0.68]	0.42 [0.25, 0.71]	0.35 [0.26, 0.79]
	Ornate Detail	0.53 [0.22, 0.77]	0.51 [0.24, 0.73]	0.51 [0.20, 0.80]	0.46 [0.27, 0.79]
	Sfumato	0.56 [0.36, 0.79]	0.48 [0.25, 0.74]	0.34 [0.19, 0.52]	0.63 [0.35, 0.80]
	Symbolism	0.50 [0.23, 0.75]	0.58 [0.33, 0.80]	0.63 [0.38, 0.81]	0.61 [0.43, 0.74]
	Thin Wash	0.48 [0.21, 0.76]	0.49 [0.25, 0.75]	0.33 [0.24, 0.52]	0.38 [0.16, 0.74]

Table 7: Posterior medians with 50% credible intervals for Authenticity.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
Authentic	Black And White	0.37 [0.18, 0.65]	0.42 [0.23, 0.69]	0.34 [0.12, 0.60]	0.36 [0.20, 0.57]
	Blending Errors	0.52 [0.22, 0.76]	0.53 [0.23, 0.75]	0.51 [0.21, 0.65]	0.55 [0.28, 0.73]
	Crushed Blacks	0.49 [0.22, 0.71]	0.46 [0.20, 0.69]	0.48 [0.22, 0.77]	0.30 [0.13, 0.47]
	Digital Artifacts	0.48 [0.23, 0.71]	0.46 [0.21, 0.72]	0.35 [0.19, 0.58]	0.38 [0.18, 0.63]
	Film Grain	0.52 [0.28, 0.75]	0.50 [0.26, 0.73]	0.41 [0.17, 0.67]	0.53 [0.24, 0.71]
	Natural Edit	0.54 [0.30, 0.77]	0.49 [0.26, 0.74]	0.55 [0.29, 0.72]	0.57 [0.28, 0.80]
	Scratched Film	0.53 [0.28, 0.77]	0.53 [0.29, 0.77]	0.50 [0.32, 0.71]	0.52 [0.26, 0.75]
	Sepia	0.51 [0.29, 0.75]	0.53 [0.28, 0.75]	0.49 [0.25, 0.74]	0.57 [0.26, 0.80]

Continued on next page

Table 7: Posterior medians with 50% credible intervals for Authenticity.

target	agent parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
	Upscaling Noise	0.50 [0.25, 0.75]	0.48 [0.24, 0.73]	0.41 [0.15, 0.73]	0.51 [0.23, 0.74]
	Vintage Look	0.50 [0.27, 0.76]	0.50 [0.24, 0.74]	0.55 [0.29, 0.77]	0.41 [0.21, 0.64]
Manipulated	Black And White	0.87 [0.56, 0.94]	0.51 [0.25, 0.75]	0.72 [0.27, 0.87]	0.48 [0.35, 0.80]
	Blending Errors	0.66 [0.32, 0.88]	0.47 [0.25, 0.74]	0.37 [0.25, 0.70]	0.69 [0.31, 0.85]
	Crushed Blacks	0.71 [0.33, 0.93]	0.55 [0.30, 0.80]	0.69 [0.53, 0.88]	0.54 [0.29, 0.76]
	Digital Artifacts	0.41 [0.23, 0.68]	0.44 [0.25, 0.74]	0.60 [0.42, 0.81]	0.45 [0.31, 0.70]
	Film Grain	0.58 [0.32, 0.81]	0.49 [0.25, 0.75]	0.42 [0.14, 0.79]	0.46 [0.16, 0.71]
	Natural Edit	0.61 [0.28, 0.77]	0.52 [0.32, 0.80]	0.51 [0.40, 0.86]	0.52 [0.33, 0.79]
	Scratched Film	0.40 [0.14, 0.70]	0.49 [0.24, 0.76]	0.37 [0.15, 0.73]	0.33 [0.22, 0.51]
	Sepia	0.46 [0.21, 0.70]	0.51 [0.28, 0.72]	0.25 [0.10, 0.66]	0.38 [0.20, 0.75]
	Upscaling Noise	0.57 [0.31, 0.74]	0.58 [0.28, 0.80]	0.56 [0.46, 0.80]	0.76 [0.48, 0.90]
	Vintage Look	0.62 [0.43, 0.84]	0.49 [0.25, 0.74]	0.49 [0.24, 0.64]	0.51 [0.31, 0.76]

Table 8: Posterior medians with 50% credible intervals for Faces.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
Attractive	Beard/Facial Hair	0.44 [0.13, 0.77]	0.40 [0.19, 0.63]	0.29 [0.12, 0.52]	0.39 [0.22, 0.55]
	Blue Eyes	0.47 [0.16, 0.76]	0.61 [0.31, 0.82]	0.45 [0.21, 0.82]	0.75 [0.19, 0.89]
	Child	0.40 [0.18, 0.73]	0.53 [0.29, 0.72]	0.49 [0.24, 0.72]	0.47 [0.31, 0.58]
	Dark/Brown Skin	0.55 [0.26, 0.79]	0.48 [0.22, 0.74]	0.57 [0.27, 0.80]	0.59 [0.40, 0.74]
	Earrings	0.48 [0.16, 0.81]	0.51 [0.30, 0.78]	0.51 [0.34, 0.75]	0.39 [0.17, 0.71]
	East Asian	0.35 [0.18, 0.44]	0.25 [0.08, 0.46]	0.16 [0.08, 0.37]	0.34 [0.15, 0.54]
	Eyeglasses	0.44 [0.24, 0.67]	0.29 [0.15, 0.52]	0.46 [0.29, 0.64]	0.36 [0.19, 0.71]
	Feminine	0.65 [0.36, 0.80]	0.49 [0.22, 0.74]	0.59 [0.35, 0.84]	0.59 [0.21, 0.91]
	Masculine/Receding Hairline	0.30 [0.12, 0.52]	0.38 [0.14, 0.59]	0.18 [0.08, 0.36]	0.37 [0.26, 0.69]
	Smiling	0.38 [0.14, 0.68]	0.45 [0.17, 0.69]	0.33 [0.21, 0.62]	0.64 [0.34, 0.81]
Criminal	Beard/Facial Hair	0.53 [0.33, 0.81]	0.41 [0.21, 0.58]	0.32 [0.14, 0.66]	0.29 [0.20, 0.49]
	Blue Eyes	0.55 [0.21, 0.83]	0.52 [0.32, 0.77]	0.34 [0.22, 0.55]	0.57 [0.33, 0.81]
	Child	0.58 [0.22, 0.77]	0.50 [0.27, 0.73]	0.44 [0.15, 0.55]	0.36 [0.24, 0.76]
	Dark/Brown Skin	0.45 [0.33, 0.68]	0.51 [0.30, 0.79]	0.72 [0.47, 0.91]	0.64 [0.20, 0.84]
	Earrings	0.52 [0.19, 0.79]	0.59 [0.31, 0.82]	0.58 [0.23, 0.88]	0.34 [0.15, 0.73]
	East Asian	0.55 [0.28, 0.82]	0.39 [0.15, 0.65]	0.23 [0.08, 0.52]	0.55 [0.24, 0.83]
	Eyeglasses	0.48 [0.19, 0.77]	0.62 [0.33, 0.83]	0.43 [0.16, 0.80]	0.63 [0.45, 0.86]
	Feminine	0.41 [0.21, 0.69]	0.35 [0.15, 0.57]	0.30 [0.13, 0.44]	0.23 [0.10, 0.39]
	Masculine/Receding Hairline	0.47 [0.31, 0.75]	0.63 [0.33, 0.79]	0.87 [0.47, 0.93]	0.62 [0.44, 0.89]
	Smiling	0.40 [0.27, 0.67]	0.46 [0.23, 0.71]	0.23 [0.06, 0.49]	0.40 [0.20, 0.69]
Fun	Beard/Facial Hair	0.46 [0.20, 0.70]	0.36 [0.20, 0.53]	0.35 [0.22, 0.68]	0.38 [0.21, 0.72]
	Blue Eyes	0.62 [0.31, 0.78]	0.50 [0.20, 0.71]	0.54 [0.28, 0.71]	0.41 [0.21, 0.75]
	Child	0.49 [0.29, 0.63]	0.55 [0.30, 0.80]	0.43 [0.13, 0.60]	0.62 [0.33, 0.69]
	Dark/Brown Skin	0.59 [0.24, 0.75]	0.58 [0.29, 0.81]	0.56 [0.37, 0.72]	0.56 [0.22, 0.69]
	Earrings	0.56 [0.24, 0.82]	0.74 [0.44, 0.84]	0.61 [0.39, 0.82]	0.50 [0.31, 0.74]
	East Asian	0.66 [0.39, 0.87]	0.67 [0.43, 0.84]	0.41 [0.22, 0.74]	0.37 [0.21, 0.79]
	Eyeglasses	0.80 [0.62, 0.89]	0.49 [0.25, 0.71]	0.71 [0.23, 0.88]	0.45 [0.12, 0.83]

Continued on next page

Table 8: Posterior medians with 50% credible intervals for Faces.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
	Feminine	0.49 [0.14, 0.63]	0.62 [0.30, 0.72]	0.76 [0.42, 0.95]	0.48 [0.22, 0.80]
	Masculine/Receding Hairline	0.54 [0.35, 0.71]	0.37 [0.22, 0.58]	0.53 [0.32, 0.69]	0.38 [0.30, 0.80]
	Smiling	0.82 [0.68, 0.95]	0.82 [0.47, 0.95]	0.85 [0.71, 0.95]	0.87 [0.61, 0.92]
Hardworking	Beard/Facial Hair	0.41 [0.22, 0.65]	0.40 [0.22, 0.66]	0.39 [0.17, 0.48]	0.48 [0.25, 0.73]
	Blue Eyes	0.38 [0.19, 0.66]	0.48 [0.23, 0.68]	0.49 [0.29, 0.62]	0.36 [0.18, 0.62]
	Child	0.47 [0.18, 0.76]	0.35 [0.11, 0.69]	0.25 [0.12, 0.56]	0.44 [0.22, 0.70]
	Dark/Brown Skin	0.60 [0.30, 0.81]	0.64 [0.35, 0.81]	0.65 [0.40, 0.89]	0.55 [0.27, 0.81]
	Earrings	0.48 [0.23, 0.69]	0.43 [0.18, 0.64]	0.46 [0.28, 0.76]	0.43 [0.29, 0.65]
	East Asian	0.43 [0.18, 0.71]	0.32 [0.15, 0.60]	0.18 [0.10, 0.43]	0.37 [0.17, 0.63]
	Eyeglasses	0.68 [0.36, 0.87]	0.79 [0.45, 0.93]	0.72 [0.56, 0.86]	0.76 [0.53, 0.88]
	Feminine	0.50 [0.25, 0.72]	0.47 [0.24, 0.72]	0.41 [0.15, 0.69]	0.52 [0.27, 0.70]
	Masculine/Receding Hairline	0.41 [0.24, 0.65]	0.35 [0.23, 0.62]	0.44 [0.29, 0.60]	0.33 [0.15, 0.67]
	Smiling	0.44 [0.28, 0.67]	0.48 [0.27, 0.79]	0.31 [0.21, 0.60]	0.39 [0.20, 0.65]
Intelligent	Beard/Facial Hair	0.59 [0.27, 0.78]	0.54 [0.33, 0.76]	0.50 [0.30, 0.78]	0.52 [0.32, 0.81]
	Blue Eyes	0.39 [0.17, 0.73]	0.52 [0.27, 0.74]	0.57 [0.23, 0.76]	0.40 [0.17, 0.63]
	Child	0.34 [0.17, 0.55]	0.30 [0.15, 0.49]	0.30 [0.16, 0.44]	0.36 [0.21, 0.65]
	Dark/Brown Skin	0.64 [0.42, 0.84]	0.64 [0.39, 0.84]	0.59 [0.35, 0.81]	0.46 [0.31, 0.80]
	Earrings	0.27 [0.11, 0.54]	0.50 [0.26, 0.75]	0.33 [0.15, 0.59]	0.41 [0.21, 0.80]
	East Asian	0.29 [0.15, 0.48]	0.43 [0.15, 0.66]	0.41 [0.17, 0.56]	0.43 [0.26, 0.70]
	Eyeglasses	0.85 [0.71, 0.94]	0.74 [0.57, 0.88]	0.80 [0.68, 0.90]	0.75 [0.24, 0.88]
	Feminine	0.42 [0.20, 0.76]	0.46 [0.26, 0.70]	0.34 [0.10, 0.59]	0.49 [0.31, 0.82]
	Masculine/Receding Hairline	0.39 [0.24, 0.56]	0.35 [0.18, 0.64]	0.46 [0.26, 0.76]	0.39 [0.18, 0.65]
	Smiling	0.44 [0.19, 0.73]	0.56 [0.24, 0.81]	0.49 [0.24, 0.70]	0.51 [0.29, 0.84]
Serious	Beard/Facial Hair	0.45 [0.27, 0.70]	0.55 [0.32, 0.72]	0.37 [0.16, 0.54]	0.49 [0.16, 0.63]
	Blue Eyes	0.29 [0.15, 0.61]	0.50 [0.24, 0.74]	0.48 [0.25, 0.75]	0.32 [0.11, 0.72]
	Child	0.42 [0.21, 0.65]	0.54 [0.27, 0.71]	0.28 [0.14, 0.67]	0.41 [0.17, 0.60]
	Dark/Brown Skin	0.58 [0.24, 0.83]	0.64 [0.39, 0.84]	0.66 [0.24, 0.86]	0.47 [0.30, 0.80]
	Earrings	0.33 [0.14, 0.60]	0.56 [0.24, 0.67]	0.36 [0.15, 0.63]	0.22 [0.08, 0.52]
	East Asian	0.40 [0.24, 0.60]	0.34 [0.15, 0.54]	0.28 [0.12, 0.59]	0.46 [0.15, 0.72]
	Eyeglasses	0.72 [0.36, 0.89]	0.54 [0.32, 0.82]	0.87 [0.69, 0.96]	0.74 [0.56, 0.90]
	Feminine	0.43 [0.20, 0.79]	0.37 [0.15, 0.54]	0.60 [0.42, 0.78]	0.56 [0.26, 0.74]
	Masculine/Receding Hairline	0.42 [0.17, 0.76]	0.69 [0.40, 0.83]	0.25 [0.10, 0.51]	0.43 [0.14, 0.64]
	Smiling	0.35 [0.14, 0.60]	0.32 [0.14, 0.59]	0.30 [0.14, 0.49]	0.30 [0.11, 0.57]
Trustworthy	Beard/Facial Hair	0.51 [0.16, 0.80]	0.52 [0.30, 0.71]	0.57 [0.30, 0.81]	0.48 [0.29, 0.75]
	Blue Eyes	0.36 [0.11, 0.65]	0.37 [0.12, 0.66]	0.36 [0.18, 0.69]	0.42 [0.20, 0.67]
	Child	0.51 [0.24, 0.75]	0.65 [0.40, 0.76]	0.40 [0.18, 0.68]	0.40 [0.17, 0.68]
	Dark/Brown Skin	0.48 [0.29, 0.76]	0.45 [0.20, 0.72]	0.50 [0.21, 0.78]	0.61 [0.33, 0.77]
	Earrings	0.49 [0.28, 0.71]	0.59 [0.37, 0.85]	0.53 [0.32, 0.80]	0.52 [0.27, 0.79]
	East Asian	0.34 [0.19, 0.54]	0.39 [0.17, 0.58]	0.39 [0.18, 0.62]	0.38 [0.15, 0.64]
	Eyeglasses	0.28 [0.14, 0.47]	0.34 [0.13, 0.63]	0.65 [0.25, 0.83]	0.56 [0.25, 0.82]
	Feminine	0.67 [0.41, 0.91]	0.65 [0.29, 0.85]	0.73 [0.56, 0.90]	0.48 [0.23, 0.68]
	Masculine/Receding Hairline	0.32 [0.13, 0.73]	0.34 [0.15, 0.70]	0.41 [0.16, 0.64]	0.44 [0.21, 0.64]
	Smiling	0.54 [0.27, 0.88]	0.59 [0.33, 0.84]	0.81 [0.63, 0.92]	0.64 [0.36, 0.84]
Youthful	Beard/Facial Hair	0.46 [0.14, 0.76]	0.46 [0.18, 0.72]	0.74 [0.39, 0.93]	0.44 [0.22, 0.72]
	Blue Eyes	0.38 [0.21, 0.72]	0.49 [0.23, 0.73]	0.46 [0.35, 0.71]	0.41 [0.08, 0.64]

Continued on next page

Table 8: Posterior medians with 50% credible intervals for Faces.

Target	Parameter	Claude Sonnet 4.6	GPT-5.4	Gemini 3 Flash	Qwen3-VL 235B
	Child	0.42 [0.24, 0.63]	0.48 [0.25, 0.73]	0.56 [0.30, 0.79]	0.35 [0.23, 0.63]
	Dark/Brown Skin	0.40 [0.23, 0.68]	0.42 [0.29, 0.67]	0.53 [0.17, 0.80]	0.58 [0.32, 0.79]
	Earrings	0.28 [0.13, 0.50]	0.50 [0.24, 0.68]	0.43 [0.12, 0.75]	0.52 [0.28, 0.69]
	East Asian	0.76 [0.51, 0.89]	0.63 [0.44, 0.80]	0.79 [0.65, 0.85]	0.49 [0.22, 0.78]
	Eyeglasses	0.38 [0.21, 0.56]	0.30 [0.16, 0.52]	0.48 [0.24, 0.67]	0.34 [0.26, 0.68]
	Feminine	0.61 [0.46, 0.89]	0.72 [0.44, 0.87]	0.49 [0.43, 0.61]	0.68 [0.46, 0.90]
	Masculine/Receding Hairline	0.49 [0.21, 0.81]	0.31 [0.18, 0.60]	0.49 [0.26, 0.69]	0.30 [0.23, 0.83]
	Smiling	0.36 [0.13, 0.67]	0.28 [0.13, 0.54]	0.29 [0.23, 0.46]	0.35 [0.15, 0.58]

J PROMPTS

The prompts used for the task instructions are the ones sent to the MLLMs to complete the task (i.e. match the target). The prompts passed to the generative image model FLUX.1-schnell are fixed within each domain, and constrain the visual format so that the slider perturbations are the dominant source of variation across stimuli.

Tasks

Faces: *Which face better matches this adjective: {target}?*
Affordances: *Which object better matches this property: {target}?*
Aesthetics: *Which artwork better matches this description: {target}?*
Authenticity: *Which image looks more {target}?*

Image Generation

Faces: *id photo of a person from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face*
Affordances: *a single decorative colored object with a unique shape, isolated photograph, large in the frame, light gray seamless backdrop, soft studio lighting, centered, soft shadow, contrast, photorealistic*
Aesthetics: *a painting*
Authenticity: *a photograph of an outdoor city scene*

K SLIDER VALIDATION

Beyond per-slider behavior, we need to verify that joint configurations of the sliders can actually steer the generative model toward diverse image distributions. This is a fundamental constraint our method inherits from the underlying generator (as indicated in Section 5), and the reason why we trained with diverse prompts (details in Appendix M). As in Harrison et al. (2020) and their StyleGAN, we can only recover priors over the regions of stimulus space that SliderSpace+FLUX can reach. If the latent does not admit configurations depicting, say, a particular demographic, the absence of that demographic from the recovered posterior of a target is uninformative—not evidence that the model does not associate the two.

To probe the reachable subspace in our trained sliders, we run separate chains targeting a small set of demographic descriptors—*woman*, *man*, *black person*, *white person*, *asian person*, *elderly person*, *young person*, *hispanic person*, *indigenous person*, *middle eastern person*—and inspect whether each chain converges to images recognizable as instances of the descriptor. We use *Qwen3-VL 235B*, 1 chain per target, and 1,000 trials per chain for the experiments. Whether each chain reaches a region

matching its descriptor is in turn judged automatically by *Gemini 3 Flash*: each of the top-k accepted images is shown to the autointerpretability model alongside the target descriptor, and the per-target match rate is the fraction of the k images judged as matches. A positive outcome shows that the corresponding region is reachable, but we make no claim that all reachable regions are equally easy to reach, and demographic configurations that dominate FLUX's training data are likely more accessible than those that do not.

Eight of the ten descriptors are reached with perfect agreement; the exceptions are *hispanic person* (20%) and *elderly person* (0%). While FLUX could still generate images following those distributions, our limited number of sliders might not be able to steer the generation in that direction. This means concepts the sliders cannot reach are systematically absent from any recovered posterior regardless of whether the model holds the association. Including more sliders that provide finer control over the generation could solve this issue, but it comes at a computational cost that is prohibitive for our large-scale experiments. This does not invalidate results for reachable concepts; it only means that no conclusion can be drawn about concepts that the available controls cannot reliably reach.

L Top & Bottom Sliders per Domain and Model

Figures 18 and 19 expand every target into its top-3 and bottom-3 sliders across models, giving a per-target read of the priors that the pooled view averages over. This shows: (i) Face targets often have a clear single dominant slider plus secondary modifiers (ii) Other targets show weaker single-slider dominance and more cross-cutting cues (iii) The two authenticity targets are mostly antipodal: *Manipulated* loads on *Black and White*/*Crushed Blacks*/*Upscaling Noise*, *Authentic* on *Natural Edit*/*Sepia*/*Blending Errors*, and the directions are roughly inverse.

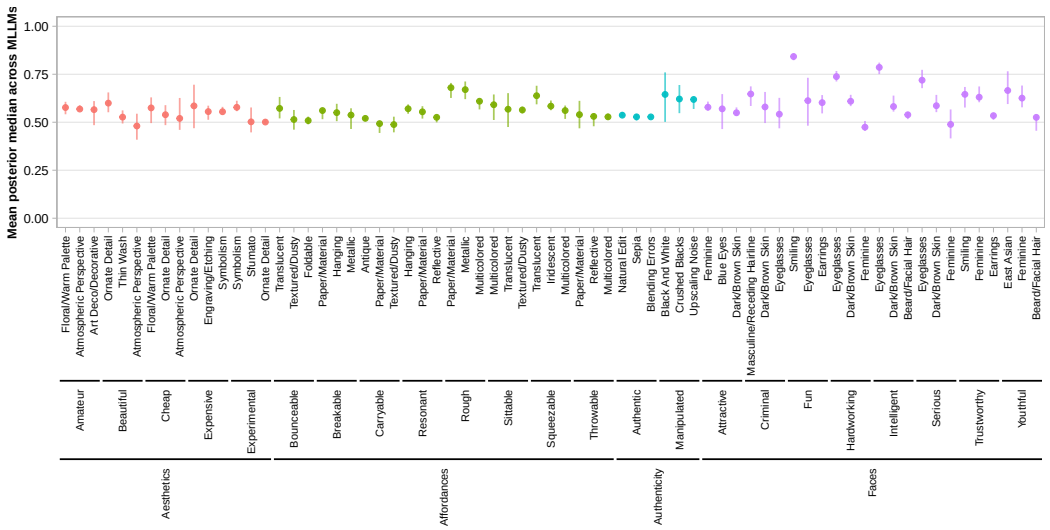


Figure 18: Top-3 (**top**) sliders per target across the four models. Points are cross-model means; bars are cross-model interquartile ranges.

Figure 20 splits the results into per-model top-5 and bottom-5 (target, slider) conditions, as discussed in Section 4.3.

M Slider Training

We train SliderSpace (Gandikota et al., 2025) sliders for each domain over FLUX.1-schnell (Labs et al., 2025). For each domain, 64 candidate sliders are trained using 50,000 CLIP-encoded samples for PCA with a batch size of 32. Training runs for 3000 iterations per slider with batch size 1, learning rate 2×10^{-3}, LoRA rank 1, $\alpha = 1$, two denoising steps per training step, guidance scale 0, and bfloat16 precision. At inference time, we use 4 denoising steps, 0 guidance scale, 512x512 resolution, and 256 maximum sequence length.

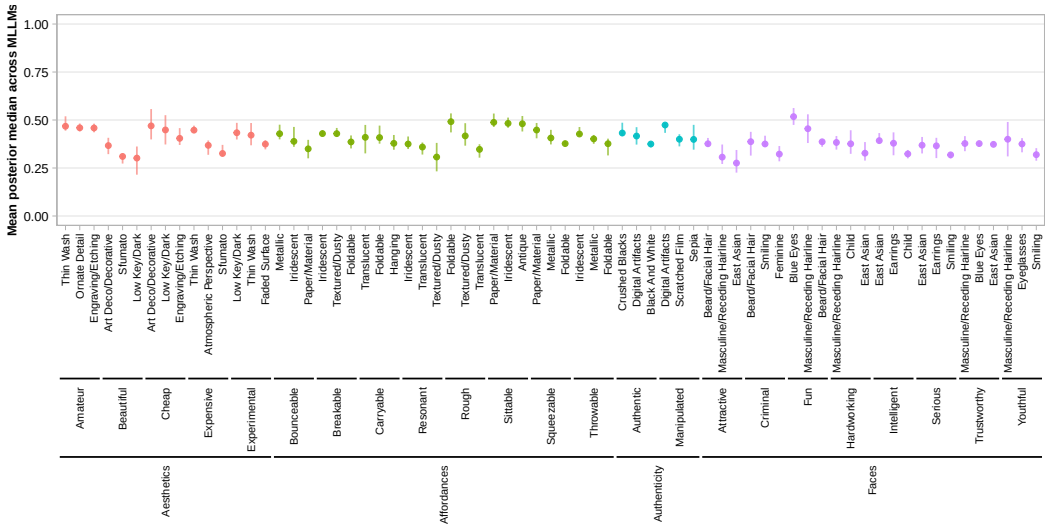


Figure 19: Bottom-3 (**bottom**) sliders per target across the four models. Points are cross-model means; bars are cross-model interquartile ranges.

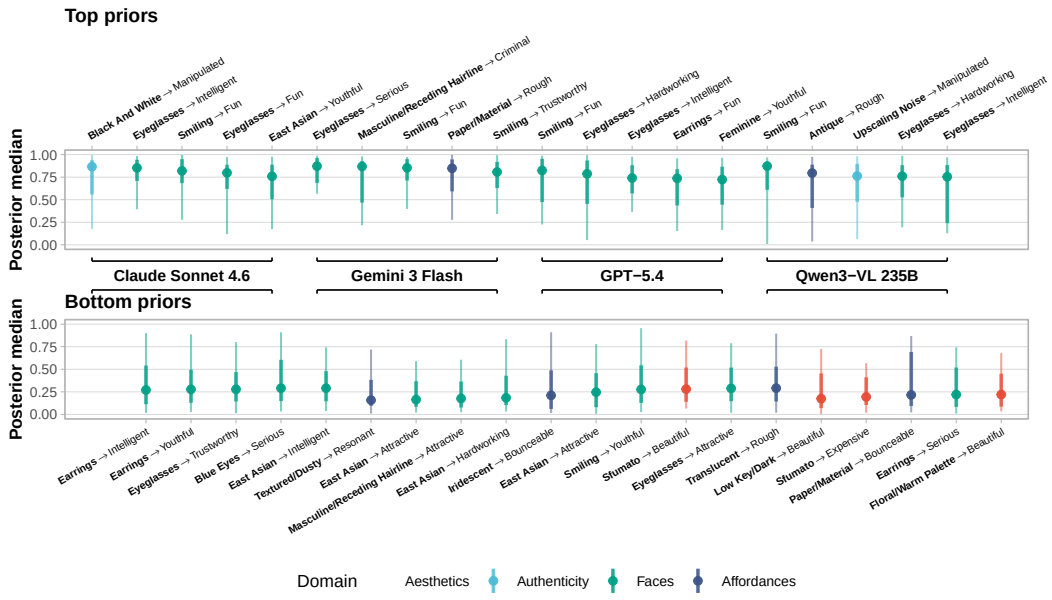


Figure 20: Top-5 (**top 2 rows**) and bottom-5 (**bottom 2 rows**) (target, slider) conditions per model. Points are per-model posterior medians, bars the 50% credible intervals.

For training, we use diverse prompts to learn diverse sliders, but prompt to isolate features that matter (e.g. face variation) while keeping other things invariant (e.g., facing camera, background).

Training Prompts: Faces

id photo of a person from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of a man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a girl from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a boy from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage White European man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage White European woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult White European man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult White European woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged White European man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged White European woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly White European man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly White European woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Black African man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Black African woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Black African man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Black African woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Black African man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Black African woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Black African man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Black African woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of a teenage East Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage East Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult East Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult East Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged East Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged East Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly East Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly East Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage South Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage South Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult South Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult South Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged South Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged South Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly South Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly South Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Southeast Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Southeast Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Southeast Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of a young adult Southeast Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Southeast Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Southeast Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Southeast Asian man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Southeast Asian woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Middle Eastern man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Middle Eastern woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Middle Eastern man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Middle Eastern woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Middle Eastern man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Middle Eastern woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Middle Eastern man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Middle Eastern woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Hispanic Latino man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Hispanic Latino woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Hispanic Latino man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Hispanic Latino woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Hispanic Latino man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Hispanic Latino woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of an elderly Hispanic Latino man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Hispanic Latino woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Indigenous Native American man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a teenage Indigenous Native American woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Indigenous Native American man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a young adult Indigenous Native American woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Indigenous Native American man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a middle-aged Indigenous Native American woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Indigenous Native American man from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of an elderly Indigenous Native American woman from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a happy expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a sad expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with an angry expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a surprised expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a neutral expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a disgusted expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a fearful expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a contemptuous expression from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person laughing from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of a person crying from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with very light skin from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with light skin from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with medium skin from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with dark skin from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with very dark skin from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with short straight hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with long straight hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with short curly hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with long curly hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with coily hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with wavy hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with black hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with brown hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with blonde hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with red hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with grey hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with white hair from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a beard from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a mustache from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with stubble from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a bald person from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with an oval face from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a round face from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a square face from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

id photo of a person with a narrow face from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing glasses from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing sunglasses from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing a hijab from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing a turban from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing a hat from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with light makeup from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with heavy makeup from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with earrings from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with facial piercings from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a subtle neck tattoo from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with prominent facial tattoos from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with freckles from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with vitiligo from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with acne from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a slim person from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a heavyset person from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person wearing hearing aids from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with a facial difference from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face
id photo of a person with Down syndrome from the shoulders up, plain white background, looking straight at camera, even lighting, centered, eye level camera angle, unobstructed face

Training Prompts: Affordances

a photograph of a small shiny metal object, plain white background
a photograph of a rough wooden tool, plain white background

a photograph of a transparent glass container, plain white background
a photograph of a colorful plastic object, plain white background
a photograph of a soft fabric pouch or bag, plain white background
a photograph of a rubber or silicone flexible object, plain white background
a photograph of a ceramic or porcelain container, plain white background
a photograph of a heavy stone or rock object, plain white background
a photograph of a paper or cardboard object, plain white background
a photograph of a leather object, plain white background
a photograph of a round spherical object, plain white background
a photograph of a flat thin disc-shaped object, plain white background
a photograph of a long cylindrical object, plain white background
a photograph of a sharp pointed object, plain white background
a photograph of a cubic box-shaped object, plain white background
a photograph of an irregular asymmetric object, plain white background
a photograph of a ring or hoop-shaped object, plain white background
a photograph of a curved arched object, plain white background
a photograph of a tiny miniature object that fits on a fingertip, plain white background
a photograph of a small pocket-sized object, plain white background
a photograph of a large heavy bulky object, plain white background
a photograph of an object with a long handle, plain white background
a photograph of an object with a looped handle, plain white background
a photograph of an object with a grip or knob, plain white background
a photograph of a smooth object with no handle or grip, plain white background
a photograph of a hollow container with a wide opening, plain white background
a photograph of a hollow container with a narrow spout, plain white background
a photograph of a solid object with no openings, plain white background
a photograph of an object with multiple holes or perforations, plain white background
a photograph of a stackable flat object, plain white background
a photograph of a collapsible or foldable object, plain white background
a photograph of a very smooth polished object, plain white background
a photograph of a rough heavily textured object, plain white background
a photograph of a soft padded object, plain white background
a photograph of a sharp-edged angular object, plain white background
a photograph of a hand tool like a wrench or hammer, plain white background
a photograph of a kitchen utensil, plain white background
a photograph of a container with a lid, plain white background
a photograph of a writing instrument, plain white background
a photograph of a fastener like a bolt, screw or clip, plain white background
a photograph of a natural object like a shell, seed or stone, plain white background
a photograph of an electronic device or gadget, plain white background
a photograph of a woven or braided object, plain white background
a photograph of a sharp cutting tool, plain white background
a photograph of a rope or cord-like object, plain white background
a photograph of a flat plate or tray, plain white background
a photograph of a measuring or gripping tool, plain white background
a photograph of a jointed or hinged object, plain white background
a photograph of a transparent object you can see through, plain white background
a photograph of a hollow object with a narrow neck, plain white background
a photograph of a small metal object with a long thin handle, plain white background
a photograph of a large wooden flat object with no openings, plain white background
a photograph of a transparent glass object with a narrow neck and wide base, plain white background
a photograph of a soft fabric object with a looped handle, plain white background
a photograph of a small ceramic container with a wide opening and no lid, plain white background
a photograph of a heavy metal object with multiple holes, plain white background
a photograph of a flexible rubber object with a cylindrical shape, plain white background

a photograph of a small smooth spherical metal object, plain white background
a photograph of a large hollow plastic container with a lid, plain white background
a photograph of a flat wooden object with a handle on one end, plain white background
a photograph of a sharp metal object with a pointed tip and a grip, plain white background
a photograph of a small plastic object with multiple buttons or controls, plain white background
a photograph of a woven object with an open basket shape, plain white background
a photograph of a long thin wooden cylindrical object, plain white background
a photograph of a heavy stone flat rectangular object, plain white background
a photograph of a metal coiled spiral object, plain white background
a photograph of a small leather flat foldable object, plain white background
a photograph of a ceramic object with a handle and a spout, plain white background
a photograph of a metal hinged object with two arms, plain white background
a photograph of a soft padded fabric object with a zipper, plain white background
a photograph of a transparent plastic object with a narrow tip, plain white background
a photograph of a rough stone object with an irregular pointed shape, plain white background
a photograph of a large cylindrical metal container with a lid, plain white background
a photograph of a small wooden object with an irregular carved shape, plain white background
a photograph of a flat perforated metal object with a long handle, plain white background
a photograph of a glass spherical object with a smooth polished surface, plain white background
a photograph of a rubber object with a bulbous hollow squeezable shape, plain white background
a photograph of a metal object with a threaded cylindrical shape, plain white background
a photograph of a large fabric object that collapses flat, plain white background
a photograph of a small electronic object with a flat rectangular shape and a screen, plain white background

Training Prompts: Aesthetics

Renaissance oil painting, classical composition, rich colors, detailed figures
Baroque oil painting, dramatic chiaroscuro lighting, dynamic composition
Romantic painting, dramatic landscape, emotional intensity, stormy atmosphere
Impressionist painting, loose brushstrokes, natural outdoor light, dappled colors
Post-Impressionist painting, bold expressive colors, thick brushwork
Pointillist painting, small dots of color, vibrant scene
Cubist painting, fragmented geometric forms, multiple perspectives simultaneously
Surrealist painting, dreamlike bizarre imagery, uncanny juxtapositions
Abstract Expressionist painting, gestural brushstrokes, raw emotion, large canvas
Minimalist artwork, simple geometric forms, very limited color palette
Pop Art, bold flat colors, commercial imagery, graphic design influence
Art Nouveau illustration, organic flowing lines, decorative natural motifs
Expressionist painting, distorted forms, intense emotional colors
Color field painting, large areas of flat uniform color, subtle gradients
Contemporary abstract painting, mixed media, conceptual layered composition
Contemporary street art, urban graffiti style, bold outlines, spray paint
Contemporary digital art, modern aesthetic, vibrant colors, crisp edges
Contemporary conceptual artwork, thought-provoking, unconventional materials
Medieval illuminated manuscript style, gold leaf, flat decorative figures
Ancient fresco painting, muted earth tones, classical figures on wall
Japanese woodblock print, flat areas of color, decorative patterns, nature scene
Watercolor painting, translucent washes, soft edges, delicate
Charcoal drawing, dramatic shadows, expressive gestural lines
Engraving or etching print, fine detailed lines, black and white

Dark melancholic painting, somber muted tones, moody oppressive atmosphere
Joyful colorful artwork, bright vibrant palette, uplifting cheerful scene
Serene peaceful landscape painting, calm soft atmosphere, gentle light
Dramatic epic historical painting, monumental composition, heroic scene
Unsettling eerie artwork, disturbing imagery, dark surreal atmosphere
Classical still life oil painting, detailed objects, dramatic lighting
fine art black and white photography, dramatic contrast, artistic composition
long exposure fine art photography, light trails, ethereal motion blur
classical marble sculpture, carved stone, ancient Greek or Roman style
folk art naive painting, flat simple figures, bright colors, outsider art
intricate geometric patterns artwork, tessellations, symmetry, ornate decoration
Gothic art, dark medieval religious imagery, pointed arches, somber figures
Bauhaus Constructivist design, functional geometric shapes, primary colors, bold layout
photorealistic painting indistinguishable from a photograph, hyper-detailed
Chinese ink wash painting, sparse brushstrokes, negative space, misty mountains
African tribal art, bold patterns, masks, ceremonial motifs, earthy colors

All authenticity prompts share a fixed scene placeholder, expanded to: *a narrow alley flanked by Victorian red brick buildings with arched windows, wet cobblestone ground, a green wooden door on the left wall, an iron streetlamp above, a wrought iron gate at the far end with a foggy street beyond.*

Training Prompts: Authenticity

CCTV security camera footage of [SCENE], grainy high-angle wide-angle shot, timestamp overlay, washed out colors
smartphone photo of [SCENE] taken at arm's length, slight wide-angle distortion, casual framing
smartphone snapshot of [SCENE], casual handheld shot, slightly off-center framing, mobile photography
disposable film camera photo of [SCENE], overexposed flash, light leaks, grainy, washed out colors
sharp professional photograph of [SCENE], shallow depth of field, blurred background bokeh, high resolution
dashcam footage of [SCENE], road perspective through windshield, wide angle, slightly fisheye
bodycam footage of [SCENE], chest-level perspective, slight motion blur
raw unedited photograph of [SCENE], natural colors, no post-processing, flat exposure
heavily retouched photograph of [SCENE], enhanced colors, airbrushed appearance, magazine style
cinematic color graded photograph of [SCENE], teal and orange tones, film look
vintage film filter photograph of [SCENE], faded colors, film grain, retro look
HDR photograph of [SCENE], high dynamic range, enhanced detail in highlights and shadows
photograph of [SCENE] saved at very low JPEG quality, blocky compression artifacts visible, degraded
analog film grain photograph of [SCENE], textured grain, warm tones, film photography
AI-generated image of [SCENE], uncanny realism, slight surreal quality, synthetic appearance
natural light photograph of [SCENE], soft outdoor lighting, no flash, available light
harsh artificial light photograph of [SCENE], strong shadows, fluorescent or neon lighting
commercial advertising photograph of [SCENE], polished and stylized, studio quality
photojournalism documentary photograph of [SCENE], candid moment, high contrast, reportage style
composited photograph of [SCENE] with a fake background, cutout edges visible, inconsistent lighting

Polaroid instant photo of [SCENE], square format, faded colors, white border, slightly overexposed
long exposure photograph of [SCENE], motion blur on moving elements, sharp background, slow shutter speed
black and white street photograph of [SCENE], high contrast, candid, silver gelatin film look