
Multidimensional Observer Model and Perceptual Dimensions of Human Image Quality Assessment

Sheng Zhao

Department of Computer Science
University of Rochester
szhao32@ur.rochester.edu

Weikai Lin

Department of Computer Science
University of Rochester
wlin33@ur.rochester.edu

Yuhao Zhu

Department of Computer Science
University of Rochester
yzhu@rochester.edu

Abstract

Judging image quality is not only ecologically relevant to everyday humans tasks, but also underpins many machine vision tasks such as image generation. This paper proposes a framework to understand the inherent perceptual space underlying image quality judgment in humans. We propose a multi-dimensional observer model that represents images as distributions in a latent perceptual space and that models human judgment as comparing noisy samples. Being constrained by neural representations in the primate ventral stream and fit to large-scale behavioral data, the model enables analysis of perceptual structure while matching the predictive power of existing metrics. Using this model, we find that the perceptual spaces needed to account for image quality judgment in humans are extremely low-dimensional compared to the image space even when considering its sparsity. The exact structure of the space (e.g., dimensionalities, information encoded) varies between low-level and high-level quality judgments, suggesting that, despite a shared retinal encoding in the beginning, humans selectively construct task-dependent perceptual spaces in visual decision making.

1 Introduction

The quality of the visual input is ecologically relevant to everyday human tasks [1]. Judging the quality of an image underpins many machine vision tasks (e.g., image generation, denoising, and rendering), and has inspired a long line of perceptual metrics, from PSNR and SSIM [2] to deep-feature methods such as LPIPS [3], DreamSim [4], PieAPP [5], and DISTS [6]. However, these metrics are primarily optimized for predictive accuracy rather than for probing human perception. Fundamental questions therefore remain: what perceptual space organizes diverse images, how quality judgments are formed within this space, how many dimensions are required, and whether the structure of the space differs between low- and high-level judgments.

This paper proposes a framework to answer these questions. The key to our approach is a multi-dimensional observer model that formalizes two key assumptions underlying many IQA datasets. First, the representation of an image in the perceptual space is noisy, both across observers and within an observer. Second, in a pairwise comparison, the image that is more distant from the clean, reference image is seen as having a lower quality. Our observer model, thus, maps each image to a distribution in a multi-dimensional, latent metric space, where distance is proportional to perceptual

difference. The quality of an image is then modeled as the *expected* distance between its latent distribution and that of the reference image.

We enforce that this latent space be a simple affine transformation away from the neural representations found in the primate retinal and ventral stream (in area V1, V2, V4, and Inferior Temporal (IT) [7–9]). The physiologically meaningful representations permit model interpretation using known properties of visual processing. We fit the observer model using maximum likelihood, based on behavioral data obtained from standard triplet two-alternative forced choice (2AFC) experiments. On a diverse set of IQA datasets, including BAPPS [3], PieAPP [5], and NIGHTS [4], our observer model approaches the level of human performance and is on par or better than existing perceptual metrics.

Using this framework, we quantitatively probe the perceptual space of human IQA. We find that human IQA could be accounted for by extremely low-dimensional perceptual spaces: the dimensionality at which the model performance saturates is much lower than that of the image space, even when considering the sparsity present in natural images. Interestingly, the exact saturating dimensionality varies based on the task complexity. On low-level tasks where image quality is dictated by distortion such as blur and compression artifacts [3, 5], the model performance generally saturates within 10 dimensions, whereas for high-level tasks that require semantic-level similarity processing [4], the model saturates at a dimensionality an order of magnitude higher.

For low-level tasks, incorporating information from higher-order cortical regions, from V1 and V2 up to V4, produces only a slight performance improvement; including signals from area IT in fact degrades the model’s performance. In contrast, high-level tasks clearly benefit from information originating in high-level cortical areas, with activity from IT markedly surpassing that from V1.

In summary, this paper makes the following contributions:

- We introduce a multi-dimensional observer model that represents images as noisy samples in a latent space, permitting characterization of the perceptual structure underlying human IQA.
- By constraining the space to align with representations in the primate ventral stream and fitting the model to large-scale behavioral data, we achieve 2AFC prediction performance on par with existing perceptual metrics while offering interpretability.
- We show that only a small number of perceptual dimensions, far fewer than those of the image space, can account for human quality judgments, and that humans construct task-dependent perceptual spaces despite sharing common retinal representations across tasks.

2 Related work

Perceptual Metrics. Numerous perceptual metrics and loss functions exist, ranging from classic metrics such as PSNR and SSIM [2, 10, 11] to learning-based metrics such as LPIPS [3], DreamSim [4], DISTS [6], MILO [12], TOPIQ-FR [13], and PieAPP [5], which improve prediction accuracy through better feature encoders, multiscale aggregation, or task-specific calibration. Complementarily, the Visible Difference Predictor family [14–16] grounds its predictions in psychophysical models of the human visual system, incorporating contrast-sensitivity functions, chromatic and temporal channels, and visual masking. All of these methods, are designed to maximize prediction accuracy; none treats the dimensionality and structure of the underlying perceptual space as quantities to be recovered, which is the subject of this work.

Psychometric Scaling and Dimensions of Perception. There is a long history of work on psychometric scaling that relates perceptual experience to (differences in) physical stimuli, from classical formulations such as the Weber–Fechner law [17] and Stevens’ power law [18], to probabilistic models such as Bradley–Terry (BT) [19] and Thurstone [20], and more recent developments [21–23]. However, psychometric scaling typically maps stimuli onto a one-dimensional scale, limiting its ability to capture the structure of the underlying perceptual space, which we focus on.

A few recent studies have examined the dimensionality of the perceptual space underlying object similarity [24] and material similarity judgments [25]. Beyond focusing on different tasks (IQA in this work), these approaches build on the Bradley–Terry model; we discuss its connection to our observer model in Section 6.

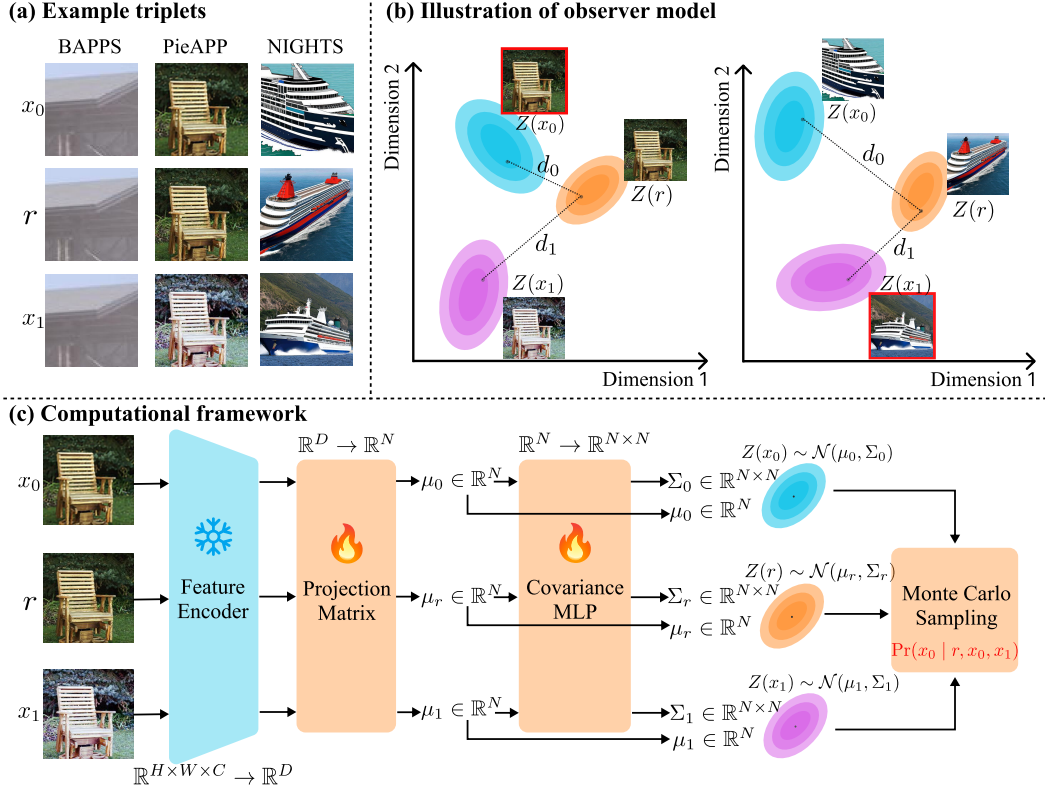


Figure 1: **Overview of the observer model.** (a): Example triplets (reference r and the two alternatives x_0 and x_1) from the three 2AFC datasets we evaluate. (b): Illustration of the observer model. Each image is modeled as an N -dimensional Gaussian random variable $Z(x) \sim \mathcal{N}(\mu(x), \Sigma(x))$ in an internal perceptual space. Each observation amounts of drawing a sample from the three distributions, and the candidate image that is closer to the reference is picked in that trial (visualized with a red bounding box). (c): The computational framework for the observer model.

3 Method

3.1 Observer Model

To probe the perceptual space underlying human IQA from behavioral data, we need to make assumptions about how observers respond to images. These assumptions are formalized in our *observer model*. Concretely, an observer’s visual system responds to each image with a noisy pattern of activity across a population of neurons. The noise arises both externally (photon shot noise) and internally (neural noise) [26, 27]. Observers judge image quality by internally comparing these noisy responses across images [28]. Because the internal responses are noisy, the same observer can return different answers on different trials. Different observers can also disagree on the same trial due to individual differences. What is reproducible about human behavior is not the individual answer, but its *distribution across trials*.

To formalize these intuitions, we encode each image x in an internal perceptual space as a random variable $Z(x)$, which we assume follows a N -variate Gaussian distribution:

$$Z(x) \sim \mathcal{N}(\mu(x), \Sigma(x)), \quad x \in \mathcal{X}, \quad (1)$$

where $\mathcal{X}$ represents the entire image space, and the functions $\mu : \mathcal{X} \mapsto \mathbb{R}^N$ and $\Sigma : \mathcal{X} \mapsto \mathbb{R}^{N \times N}$, respectively, map the stimulus x to the mean and covariance matrix of the corresponding distribution in the N -dimensional representational space. Importantly, the covariance matrix depends on x , so different images are permitted to carry different amounts of perceptual ambiguity. We will describe the parametric form of $\mu(x)$ and $\Sigma(x)$ in Section 3.2.

How does a human then make a decision? We use the triplet 2AFC task commonly seen in IQA datasets as an example. Each trial presents a reference r and two alternatives, x_0 and x_1 . On any given trial, an observer draws a sample each from the three distributions and obtains $z(r)$, $z(x_0)$, and $z(x_1)$. We posit that the observer, in that particular trial, regards x_0 as having a high quality than x_1 if $z(x_0)$ is closer to $z(r)$ than does $z(x_1)$.

We model the distance between $z(x_i)$ and the reference $z(r)$ in the perceptual space using the weighted Minkowski distance, i.e., the p -norm:

$$d_i = \|z(x_i) - z(r)\|_p = \left(\sum_{m=1}^n \alpha_m |z(x_i)_m - z(r)_m|^p \right)^{1/p}, \quad i \in \{0, 1\}, \quad \alpha_m \geq 0, \quad p \geq 1. \quad (2)$$

Both the weights α_m and the exponent p are free parameters. The weights α_m compensate for the different natural magnitudes of the perceptual channels [29]. The exponent p determines how independently the dimensions can influence the distance, ranging from separable dimensions when $p = 1$ to integral dimensions when $p = 2$ [30, 31]. Allowing p to vary let the model adapt to the underlying representational structure (see Appendix H for experimental results).

On any single trial the observer picks x_0 if $d_0 < d_1$, i.e., $\mathbf{1}\{d_0 < d_1\}$, where $\mathbf{1}(\cdot)$ is the indicator function. Clearly, d_i is a sample from a random variable $D_i = \|Z(x_i) - Z(r)\|_p$, where $Z(x_i) - Z(r) \sim \mathcal{N}(\boldsymbol{\mu}(x_i) - \boldsymbol{\mu}(r), \boldsymbol{\Sigma}(x_i) + \boldsymbol{\Sigma}(r))$. Therefore, $\mathbf{1}\{d_0 < d_1\}$ itself is a sample from a Bernoulli random variable, whose expected value is the probability that x_0 is picked over x_1 :

$$\Pr(x_0 \mid r, x_0, x_1) = \mathbb{E}[\mathbf{1}\{D_0 < D_1\}]. \quad (3)$$

While the discussion so far focuses on triplet 2AFC tasks where the reference and both alternatives are simultaneously presented by an observer, our observer model also readily extends to other psychophysical paradigms (e.g., 2AFC tasks where the reference image is not shown [22] or subjective rating tasks). We discuss these extensions in Appendix E.

3.2 Fitting the Observer Model Using Behavioral Data

Parameterizing the Mean and Covariance Matrix Field. The internal representation of an image x on a given trial depends on its distribution, characterized by its mean $\boldsymbol{\mu}(x)$ and covariance matrix $\boldsymbol{\Sigma}(x)$ (Eq. 1). Parameterizing these two functions must meet two requirements. First, x lies in the image space $\mathcal{X}$, which is high dimensional, but the internal representation used for assessing quality may be of a much lower, but unknown, dimensionality. Therefore, both functions must permit flexible dimensionality reduction without committing to a fixed intrinsic dimensionality. Second, perceptual responses vary smoothly with stimulus changes, so both $\boldsymbol{\mu}(x)$ and $\boldsymbol{\Sigma}(x)$ must be smooth over x .

A direct projection from the pixel space to $\boldsymbol{\mu}(x)$, while meeting the two requirements, is undesirable, as quality assessment is unlikely determined by a weighted sum of pixel values. Instead, we first encode x as $B(x) \in \mathbb{R}^D$ in a space that is aligned with human visual processing. For instance, $B(\cdot)$ can be instantiated using pretrained convolutional neural networks such as CORnet-S [9] or VGG [32]. Features from these networks have been shown to predict neural responses across multiple areas of the visual cortex (e.g., V1, V2, V4, IT) in a layer-wise manner, providing a biologically grounded embedding of images [33, 34, 33, 35]. Alternatively, one can also represent $B(x)$ using the contrast pyramid of x [36], which approximates the retinal representation of an image [16].

We then obtain $\boldsymbol{\mu}(x)$ via a linear projection:

$$\boldsymbol{\mu}(x) = \mathbf{W} B(x) + \mathbf{b}, \quad \mathbf{W} \in \mathbb{R}^{N \times D}, \quad \mathbf{b} \in \mathbb{R}^N, \quad (4)$$

where N is a hyper-parameter that can be tuned. A linear projection is the minimum mechanism that enables flexibly control of the dimensionality N and preserves the dimensions in $B(x)$ up to an affine transformation, avoiding distortions that would make the latent dimensions harder to interpret. Both $B(\cdot)$ and the linear projection are smooth over x , so $\boldsymbol{\mu}(x)$ inherits the property. We additionally explore using PCA instead of a learned projection matrix for dimensionality reduction; see Appendix C for further details.

The covariance matrix field had an additional constraint that it must be a positive semi-definite matrix for every x . To that end, we rewrite $\boldsymbol{\Sigma}(x)$ through the Cholesky decomposition,

$$\boldsymbol{\Sigma}(x) = \mathbf{L}(x) \mathbf{L}(x)^\top, \quad (5)$$

and enforce that $\mathbf{L}(x) \in \mathbb{R}^{N \times N}$ is a lower-triangular matrix with non-zero diagonal entries; this ensures that $\Sigma(x)$ is positive semi-definite by construction. $\mathbf{L}(x)$ is generated by:

$$\mathbf{L}(x) = s(G_\psi(\mu(x))), \quad (6)$$

where G_ψ is a small MLP and $s(\cdot)$ applies a softplus function [37] to the diagonal elements in the MLP output to ensure positivity. $\Sigma(x)$ is smooth over x because the MLP output is naturally smooth over the input $\mu(x)$, which is smooth over x . Appendix A gives the MLP architecture.

Likelihood. Given the forms of the mean field $\mu(\cdot)$ and the covariance matrix field $\Sigma(\cdot)$, we can now describe the (log) likelihood function:

$$\log p(\mathcal{O} \mid \Theta) = \sum_{i=1}^T J^{(i)} \left[y^{(i)} \log P^{(i)} + (1 - y^{(i)}) \log (1 - P^{(i)}) \right], \quad (7)$$

where $\mathcal{O}$ represents observations collected over all the trails (each of which provides whether x_0 or x_1 is picked), $P^{(i)} = \Pr(x_0^{(i)} \mid r^{(i)}, x_0^{(i)}, x_1^{(i)})$ is the probability of choosing x_0 on triplet i predicted by our observer model (Eq. 3), $y^{(i)}$ is the fraction of actual human observations where x_0 is picked on triplet i , and $J^{(i)}$ is the total number of observations made on triplet i .

The trainable parameter set is $\Theta = \{\mathbf{W}, \mathbf{b}, G_\psi, \alpha, p\}$, where α and exponent p are the per-dimension weights and exponent from Eq. 2, $\mathbf{W}$ and $\mathbf{b}$ are used to generate the mean field in Eq. 4, and G_ψ is the MLP that generates the covariance matrix field in Eq. 5.

Model Fitting. We estimate the parameters Θ by maximizing the likelihood on human data provided in 2AFC datasets (Section 4). To use gradient descent for optimization, the likelihood function must be differentiable. The expectation in Eq. 3, however, generally has no closed form [38]. We resort to a numerical approximation using Monte Carlo (MC) simulations, where we draw K joint samples from the three distributions and averaging the distance comparisons over all K simulation runs:

$$\Pr(x_0 \mid r, x_0, x_1) = \mathbb{E}[\mathbf{1}\{D_0 < D_1\}] \approx \frac{1}{K} \sum_{k=1}^K (\mathbf{1}\{d_0^{(k)} < d_1^{(k)}\}), \quad (8)$$

where $d_i^{(k)}$ is a sample from D_i at MC simulation run k . The indicator function $\mathbf{1}(\cdot)$ is not differentiable. We use the standard Straight-Through Estimator (STE) [39] to allow gradient propagation. For Eq. 8 to be differentiable, the samples themselves must be differentiable. We achieve this using the standard reparameterization trick [40]. In Appendix A, we discuss the implementation details of the MC sampling and reparameterization as well as other design alternatives such as replacing the indicator function with a smooth function (e.g., sigmoid). We also discuss the implications of applying a prior in Appendix G.

4 Experiment Setup

Feature Encoder. Our primary choice of encoder $B(\cdot)$ in Eq. 4 is CORnet-S [9], a convolutional neural network with four major computational blocks. The output of each block has been shown to align with neural recordings of one of the four areas of the primate ventral stream: V1, V2, V4, and IT. We spatially pool each channel of each area’s output, obtaining one vector per area of dimension $D = 64, 128, 256, 512$ at V1, V2, V4, and IT, respectively. We additionally concatenate these four vectors into a composite vector of dimension $D = 960$; we call this encoder “CORnet-S Multi”.

While CORnet-S encodes images into the cortical space, we also experiment with an encoder that builds a contrast pyramid from an image [36]. The contrast pyramid, built from the Laplacian pyramid [41], represents the local contrast at different image regions at different spatial-frequency bands. It approximates the spatial processing by the retina and LGN [42], and is widely used in perceptual metrics [16, 43, 44] and perceptual loss functions [45, 46].

No encoder is fine-tuned and all features are precomputed. The implementation details are in Appendix A. We also evaluate VGG-16 [32] pretrained on ImageNet classification, which has long been used as perceptual features [47, 48, 3]. Results on it are shown in Appendix C.

Datasets. We use three widely used triplet 2AFC datasets spanning both low- and high-level IQA tasks: BAPPS [3] and PieAPP [5], where images within each triplet differ due to low-level distortions

Table 1: Performance of our observer model and comparison with other perceptual metrics. The baseline metrics are grouped as follows: methods that have the capability to predict choice probabilities in 2AFC tasks (Ours, CVVDP [16], and PieAPP [5]), methods that predict quality scores in the form of DMOS/MOS (DISTS [6], TOPIQ-FR [13], and MILO [12]) and in the form of distance in a learned embedding space (LPIPS [3], DreamSim [4], OpenCLIP [53], DINOv3 [54], and CLIP [55]), and classical, parameter-free metrics (SSIM [2] and PSNR). The KL divergence can only be calculated for models that could output choice probabilities. Ceilings are followed by their per-triplet 95% confidence interval (CI) averaged over all triplets. **Bold** = best, underline = second & third best.

Dataset	Ceiling $\pm$ 95% CI	CORner-S-Multi	CVVDP [16]	PieAPP [5]	DISTS [6]	TOPIQ-FR [13]	MILO [12]	DreamSim [4]	OpenCLIP [53]	DINOv3 [54]	CLIP [55]	LPIPS [3]	SSIM [2]	PSNR
BAPPS (Agr $\uparrow$)	.797 $\pm$.183	.688	.645	.629	<u>.678</u>	.667	.634	<u>.683</u>	.653	.640	.624	.662	.641	.633
cnn	.885 $\pm$.112	.831	.809	.765	<u>.813</u>	.808	.800	<u>.825</u>	.800	.796	.773	.814	.809	.801
color	.764 $\pm$.212	.634	.610	.612	<u>.615</u>	.616	.596	<u>.627</u>	.578	.572	.560	.588	<u>.624</u>	.624
deblur	.740 $\pm$.224	<u>.596</u>	.594	.547	.599	.599	.584	<u>.591</u>	.560	.572	.539	.573	<u>.594</u>	.590
f.int	.760 $\pm$.208	.621	.590	.606	<u>.626</u>	.610	.558	.632	.611	.613	.593	<u>.627</u>	.555	.543
s.res	.801 $\pm$.181	.697	.658	.619	<u>.692</u>	.686	.645	<u>.695</u>	.670	.624	.627	.690	.649	.642
trad.	.858 $\pm$.135	.790	.609	.703	<u>.749</u>	.695	.614	<u>.771</u>	.744	.734	.716	.714	.600	.573
BAPPS (KL $\downarrow$)	—	.170	.265	.236	—	—	—	—	—	—	—	—	—	—
PieAPP (Agr $\uparrow$)	.766 $\pm$.105	<u>.710</u>	.591	.717	.690	.651	.616	.719	.661	.651	.671	.640	.584	.584
PieAPP (KL $\downarrow$)	—	<u>.088</u>	.985	.083	—	—	—	—	—	—	—	—	—	—
NIGHTS (Acc $\uparrow$)	1.000 $\pm$ 0	<u>.859</u>	.703	.629	<u>.860</u>	.633	.621	.957	.848	.851	.837	.745	.621	.569

(e.g., blur, noise, compression), and NIGHTS [4], where variation arises from the stochasticity of text-to-image generation rather than explicit distortions.

We also evaluate on four other popular IQA datasets, which either do not use 2AFC tasks (LIVE [49], KADID-10k [50], and CSIQ [51]) or do not provide raw pairwise comparison data (TID2013 [52]). They do, however, provide a quality score/rating for each image in the form of (differential) Mean Opinion Score (MOS). We discuss how our method extends to these rating datasets in Appendix A.

Evaluation metrics. For BAPPS and PieAPP, which provide raw pairwise comparison data for each trial (whether x_0 or x_1 is picked), there are two common ways to evaluate the performance. First, we report the “agreement score” used by BAPPS [3]: $\frac{1}{T} \sum_i (p_i q_i + (1 - p_i)(1 - q_i))$, where p_i is the fraction of trials that, say, x_0 is picked in triplet i , q_i is the predicted probability that x_0 would be picked, given by Eq. 8, and T is the total number of triplets. This metric measures the probability that a model and an average human observer would agree on every trial. The theoretical ceiling is $\frac{1}{T} \sum_i \max\{p_i, 1 - p_i\}$, which is 79.7% and 76.5% for BAPPS and PieAPP, respectively. The inter-observer agreement, i.e., the expectation that two independent human observers agree on every trial, is $\frac{1}{T} \sum_i (p_i^2 + (1 - p_i)^2)$; this value is 73.1% on BAPPS and 68.7% on PieAPP.

Unlike our observer model, many existing perceptual metrics predict a quality score for an image (relative to a reference) instead of directly modeling the choice probabilities in a 2AFC task. When using these metrics to model the choice probabilities in a 2AFC task, a common method is to predict the image having a higher predicted score to always be picked [5]. For a fair comparison against these metrics, we also binarize the output of our observer model through majority voting of the predicted probabilities. Effectively, $q_i \in \{0, 1\}$. A sufficiently good model could out-perform the inter-observer agreement, but should always be worse than the theoretical ceiling.

There exist perceptual metrics that could model choice probabilities in a 2AFC task, as they produce perceptual scores via psychometric scaling that are consistent with the observed choice probabilities [16, 22, 5]. For comparison with these metrics, we compute and report the KL divergence between the choice distribution of human observers and that of the model predictions.

For NIGHTS whose labels are binary to begin with, we binarize the output of our observer model through majority voting; the agreement score then reduces to binary error rate (BER) [5]. We report 1 - BER as the prediction accuracy.

Protocol. For each feature encoder and dataset pair, we sweep the perceptual dimension N . As N increases, the prediction accuracy gradually saturates, at which point we have an empirical estimate of a task’s intrinsic perceptual dimensionality. We define the saturation point n^* as the smallest N at which the model is able to explain 95% of model’s own above-chance performance:

Table 2: Saturation point n^* of our observer model based on the agreement score. The results still hold true under KL divergence, whose results and full per-distortion breakdown are shown in Appendix C.

Feature Encoder	BAPPS		PieAPP		NIGHTS	
	Agr	n^*	Agr	n^*	Acc	n^*
Contrast Pyramid	.645	3	.647	2	.703	7
CORnet-S V1	.685	4	.701	6	.747	8
CORnet-S V2	.689	3	.712	5	.793	14
CORnet-S V4	.688	4	.714	8	.855	48
CORnet-S IT	.677	6	.702	14	.853	96
CORnet-S Multi	.688	3	.710	9	.859	96

Table 3: Agreement score comparison on BAPPS between frequency-space sparsity (dictated by top K dimensions/ch. retained after DCT) vs. perceptual-space sparsity (dictated by N). Shaded: a pair of comparison where the sparsity is the same ($K=N/3$).

Feature Encoder	Perceptual Space		Frequency Space (top- K /ch.)			
	$N=2$	$N=6$	$K=2$	$K=10$	$K=20$	$K=200$
Contrast Pyramid	.634	.637	.536	.574	.586	.627
CORnet-S V1	.666	.682	.552	.589	.615	.674
CORnet-S V2	.672	.685	.563	.608	.632	.686
CORnet-S V4	.667	.685	.564	.615	.636	.686
CORnet-S IT	.647	.669	.564	.609	.628	.674
CORnet-S Multi	.670	.686	.566	.612	.635	.684

$n^* = \min\{N : \frac{m(N)-0.5}{m^*-0.5} > 0.95\}$, where 0.5 is the chance level of 2AFC tasks, $m(N)$ is the model’s performance at dimension N , and m^* is the maximum performance when N is unconstrained. We use the following hyperparameters: Adam with learning rate 10^{-3} and cosine annealing, batch size 256, 50 epochs, and $K = 64$ Monte Carlo samples.

5 Results

5.1 Our Observer Model Predict Human 2AFC Tasks Accurately

Our goal is *not* to set a new state of the art in predicting human 2AFC choices, but to probe the human perceptual space of quality judgment. Nevertheless, to serve as a reliable tool for this purpose, our method must achieve competitive performance relative to existing metrics, as we demonstrate below.

Table 1 shows our accuracy on the three datasets and that of other perceptual metrics. See the caption for the descriptions of the baseline metrics. In this experiment we use the CORnet-S-Multi as the feature encoder. On the agreement score metric, our model reaches 0.688 agreement on BAPPS, outperforming all baselines; on PieAPP, our model narrowly trails the best method and is within 0.056 of the theoretical ceiling; on NIGHTS, our model essentially ties with DISTS and falls only behind DreamSim, which uses modern foundation models and has over 80 times more trainable parameters than ours. On KL divergence, our observer model performs the best compared to CVVDP and PieAPP on the BAPPS dataset, and narrowly trails as the second best on the PieAPP dataset.

In general, our model (and other metrics) performs worse on BAPPS than on PieAPP. One reason is that BAPPS collects only 5 observations per triplet, whereas PieAPP collects up to 40. A lower sample count means higher degrees of uncertainty and wider confidence intervals with respect to the underlying population distribution, as shown in Table 1.

5.2 Extremely Low-Dimensional Perceptual Spaces Account for Human Judgment

As the dimensionality of the perceptual space increases, the predictive performance gradually saturates. We show that quality judgments are well described by an extremely low-dimensional perceptual space, and the low dimensionality is not a trivial consequence of the sparsity of natural images.

Figure 2 shows the performance our observer model on the three datasets as a function of N . We show six variants of our model, each using a different feature encoder. Table 2 shows the saturating N and the corresponding performance. Overall, the saturating dimensionalities n^* are extremely low. On both low-level datasets, the saturating dimensionality is generally around 5, and on the high-level dataset, the dimensionality is about $10\times$ higher, but generally the saturating dimensionalities are much lower than the dimensionality of the image space (i.e., height $\times$ weight $\times$ color channels).

A low saturating dimensionality could potentially be trivially explained by the well-known sparsity (compressibility) of natural images in the frequency domain. To test this, we construct a frequency-space control experiment. Inspired by how JPEG exploits image sparsity for compression, for every image color channel (in the YUV color space) we apply the Discrete Cosine Transform (DCT) and retain only the top- K DCT coefficients. We then reconstruct the image through an inverse DCT, and apply our observer model to reconstructed images (see Appendix A for details). Since truncation is

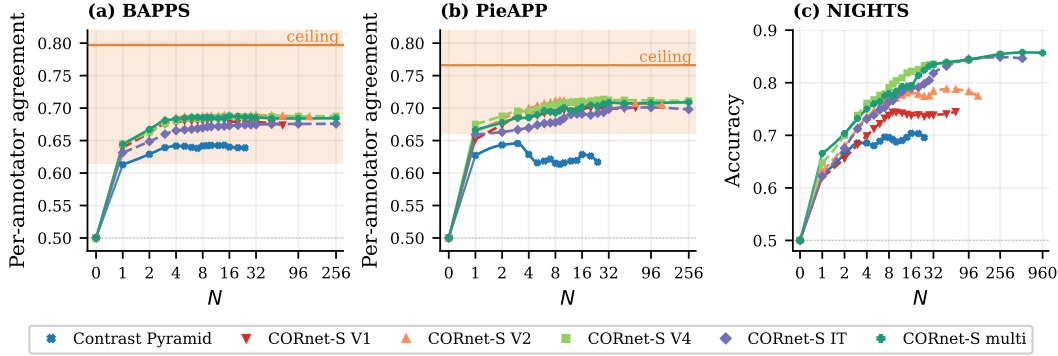


Figure. 2: Prediction performance of our observer model under different N . We show six variants of our model, each using a different feature encoder. Low-level IQA datasets (BAPPS, PieAPP) plateau within a handful of dimensions while the high-level IQA dataset NIGHTS requires more dimensionalities. For BAPPS and PieAPP, we show the 95% CI along with the theoretical ceiling derived the sample distribution; the y -axes are capped for clarity. NIGHTS ceiling is 1 and CI is 0.

done per channel, an image would have the same sparsity in the truncated frequency space and in our learned perceptual space when $K = N/3$, where N is the dimensionality of the perceptual space.

Table 3 shows the performance of our observer model and the DCT control under varying N s and K s. Matching the performance of our $N=6$ model requires roughly two orders of magnitude more dimensionalities in the DCT control. The shaded columns share the same sparsity ($N = 6, K = 2$), in which case the DCT control is about 0.1 lower in the agreement score. Thus, the low dimensionality we observe cannot simply be attributed to pixel-space sparsity.

5.3 Task Complexity Shapes Perceptual Dimensions and Representational Requirements

The dimensionality of perceptual judgment is strongly task-dependent, and the difference is consistent across encoder choices. As shown in Figure 2 and Table 2, for BAPPS and PieAPP, where judgments are driven by low-level image distortions [3, 5], saturation occurs at $n^* = 3-8$; for NIGHTS, where images differ in mid-level and semantic properties [4], n^* grows much larger, reaching 96 at IT. This is consistent with the intuition that tasks in NIGHTS involve judging higher-level, semantic differences, which in turn demand substantially higher-dimensional representations.

In both tasks across the three datasets, encoding images with the contrast pyramid gives the lowest performance; this indicates that retinal representation alone can not fully account for human judgments¹. Low-level IQA judgments depend on low-level, local image statistics, and are best supported by early-to-intermediate visual areas (V1 to V4), which all have a retinotopic map of the visual field [7]. Moving all the way to IT, which is primarily responsible for object recognition and does not have a clear retinotopic mapping, actually reduces alignment with human choices. In contrast, high-level judgments depend instead on structural and semantic organization, so they benefit from later visual representations such as IT, while performance using V1 and V2 representations are markedly lower.

We also find that a model trained on BAPPS incurs only a negligible drop in accuracy when applied directly to PieAPP, another low-level dataset, relative to a model trained on PieAPP itself. In contrast, performance degrades substantially when the same BAPPS-trained model is applied to NIGHTS, a high-level IQA dataset, compared to a model trained on NIGHTS. This corroborates that the perceptual space structure varies with task complexity. Appendix D provides the detailed results.

5.4 Interpreting the Perceptual Dimensions

If the human judgment can be captured in a relatively low-dimensional perceptual spaces, a natural question is: what does each dimension encode? Using the CORnet multi-layer model trained on BAPPS, Figure 3 shows the results at $N=2$. Figure 3(a) visualizes the covariance matrix field over the two learned dimensions, z_1 and z_2 , sampled on a 8×8 grid. Figure 3(b) offers a complementary

¹In fact, when we enforce the use of the contrast pyramid and increase the number of trainable parameters via N , the model starts fitting the noise, leading to a drop in prediction performance as seen in the results.

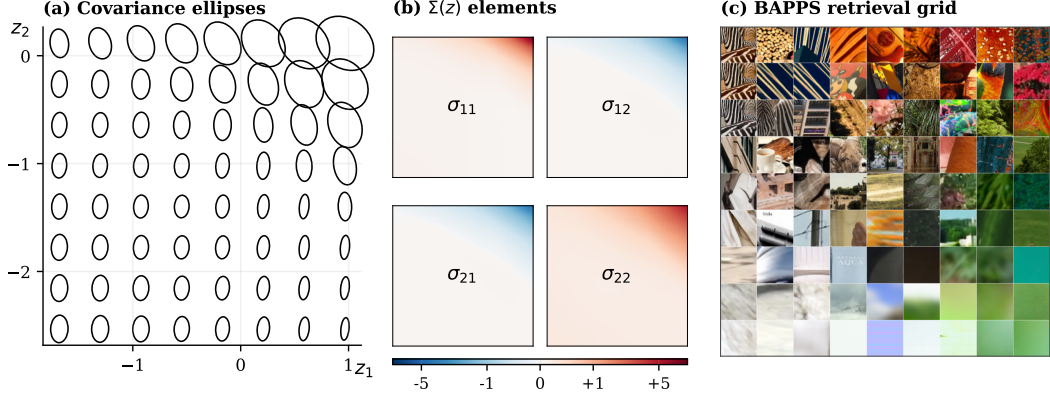


Figure. 3: Interpretation of the learned $N = 2$ model (CORnet-S-Multi, BAPPS). (a) The smoothly varying covariance matrix field over the two learned dimensions z_1 and z_2 sampled in a 8×8 grid; each sampled covariance contour is visualized as an ellipse (log-scale) bounded by two standard deviations. (b) A complementary visualization of the covariance matrix field showing how $\sigma_{z_1}^2$, $\sigma_{(z_1, z_2)}^2$, $\sigma_{(z_2, z_1)}^2$, and $\sigma_{z_2}^2$ vary individually in the learned 2D space. (c) 9×9 images that, when mapped to the perceptual space, vary vertically along the z_1 dimension from low to high and vary horizontally along the z_2 dimension from low to high. Reading along these two axes reveals that z_1 captures texture and z_2 captures color. More results are in Appendix B.

view, visualizing how each element of the covariance matrix — $\sigma_{z_1}^2$, $\sigma_{(z_1, z_2)}$, $\sigma_{(z_2, z_1)}$, and $\sigma_{z_2}^2$ — varies over the 2D perceptual space. The covariance matrix field varies smoothly, reflecting the fact that human responses change gradually with the input stimulus.

Figure 3(c) visualizes the visual information encoded by each dimension using the BAPPS dataset. We do that by comparing images that share a similar value in one dimension but differ significantly along the other dimension: images in the same row (column) share a similar z_1 (z_2) value but differ in z_2 (z_1). Ostensibly, z_1 encodes the texture/spatial frequency information while z_2 encodes the color information. These directions closely match classical distinctions in early vision, although they are learned here purely from behavioral data without any prior on their semantics. Results on the other two datasets are in Appendix B, and the general conclusions hold.

As N increases, the semantic meaning of each dimension becomes more fine-grained and harder to describe in natural language. We show in Appendix B that: 1) perceptual dimensions are still orthogonal to each other in high-dimensional perceptual spaces, and 2) perceptual dimensions for high-level IQA tasks encode more object semantics while those for low-level IQA tasks encode more color and texture information.

6 Discussion

Connection to Bradley-Terry Model. The Bradley-Terry (BT) model is a classical psychometric scaling method commonly used to predict choice distribution pairwise comparisons [19], and has been applied to triplet 2AFC tasks [5, 24, 25, 22]. The BT model assumes that the perceived quality score of an image follows the Gumbel distribution [56]. In Appendix F, we show the BT model and our method give similar predictive power, and our method can be thought of as giving a low-level interpretation of the phenomenological BT model.

Limitations. Since we jointly learn the mean and the covariance matrix field, the choice probability of a triplet remains the same if one scales the mean and covariance of each image proportionally. Thus, it is difficult to compare distributions between images not in the same triplet. In the future, we can gather more direct observations on the mean field by, for instance, collecting neural recordings during IQA tasks or integrating cross-content judgments.

7 Conclusion

This paper proposes a framework to model image quality judgment in humans. We show that the perceptual space in which humans assess image quality is extremely low dimensional, much lower than that of the image space even when considering the sparsity of the image space. We also show that humans construct task-dependent perceptual spaces from the shared retinal representations. By providing a principled bridge between behavioral data and the (learned) perceptual space, we open a new window into probing human perceptual processing of images.

References

- [1] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978.
- [2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- [4] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36:50742–50768, 2023.
- [5] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018.
- [6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.
- [7] Daniel J. Felleman and David C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1 1:1–47, 1991.
- [8] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.
- [9] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.
- [10] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1398–1402. IEEE, 2003.
- [11] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.
- [12] Ugur Cogalan, Mojtaba Bermana, Karol Myszkowski, Hans-Peter Seidel, and Colin Groth. Milo: A lightweight perceptual quality metric for image and latent-space optimization. *ACM Transactions on Graphics (TOG)*, 44(6):1–11, 2025.
- [13] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.
- [14] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.
- [15] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021.

- [16] Rafal K Mantiuk, Param Hanji, Maliha Ashraf, Yuta Asano, and Alexandre Chapiro. Colorvideodp: A visual difference predictor for image, video and display distortions. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.
- [17] Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860.
- [18] Stanley S Stevens. On the psychophysical law. *Psychological review*, 64(3):153, 1957.
- [19] 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.
- [20] Louis Leon Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.
- [21] Jonathan Vacher and Pascal Mamassian. Perceptual scales predicted by fisher information metrics. In *The Twelfth International Conference on Learning Representations*, 2024.
- [22] Maria Perez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2019.
- [23] Laurence T Maloney and Joong Nam Yang. Maximum likelihood difference scaling. *Journal of Vision*, 3(8):5–5, 2003.
- [24] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.
- [25] Filipp Schmidt, Martin N Hebart, Alexandra C Schmid, and Roland W Fleming. Core dimensions of human material perception. *Proceedings of the National Academy of Sciences*, 122(10):e2417202122, 2025.
- [26] Eero P Simoncelli, Liam Paninski, and Jonathan Pillow. Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, page 327, 2004.
- [27] William Bialek. Physical limits to sensation and perception. *Annual review of biophysics and biophysical chemistry*, 16(1):455–478, 1987.
- [28] David M Green and John A Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.
- [29] Robert M Nosofsky. Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39, 1986.
- [30] Roger N Shepard. Attention and the metric structure of the stimulus space. *Journal of mathematical psychology*, 1(1):54–87, 1964.
- [31] Wendell R Garner. *The Processing of Information and Structure*. Lawrence Erlbaum Associates, Potomac, MD, 1974.
- [32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [33] Wonbum Sohn, Xin Di, Zhen Liang, Zhiguo Zhang, and Bharat B Biswal. Explorations of using a convolutional neural network to understand brain activations during movie watching. *Psychoradiology*, 4: kkae021, 2024.
- [34] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.
- [35] Hui-Yuan Miao and Frank Tong. Convolutional neural network models applied to neuronal responses in macaque v1 reveal limited nonlinear processing. *Journal of Vision*, 24(6):1–1, 2024.
- [36] Eli Peli. Contrast in complex images. *Journal of the optical society of America A*, 7(10):2032–2040, 1990.
- [37] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

- [38] Daniel M Ennis and Kenneth Mullen. A general probabilistic model for triad discrimination, preferential choice, and two-alternative identification. In *Multidimensional models of perception and cognition*, pages 115–122. Psychology Press, 2014.
- [39] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.
- [40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [41] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pages 671–679. Elsevier, 1987.
- [42] Brian A Wandell. *Foundations of vision*. sinauer Associates, 1995.
- [43] Jeffrey Lubin. A visual discrimination model for imaging system design and evaluation. In *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, pages 245–283. World Scientific, 1995.
- [44] Mahesh Ramasubramanian, Sumanta N Pattanaik, and Donald P Greenberg. A perceptually based physical error metric for realistic image synthesis. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 73–82, 1999.
- [45] Taimoor Tariq, Nathan Matsuda, Eric Penner, Jerry Jia, Douglas Lanman, Ajit Ninan, and Alexandre Chapiro. Perceptually adaptive real-time tone mapping. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023.
- [46] Okan Tarhan Tursun, Elena Arabadzhyska-Koleva, Marek Wernikowski, Radosław Mantiuk, Hans-Peter Seidel, Karol Myszkowski, and Piotr Didyk. Luminance-contrast-aware foveated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
- [47] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.
- [48] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [49] Hamid R Sheikh, Muhammad F Sabir, Alan C Bovik, et al. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3440–3451, 2006.
- [50] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019.
- [51] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006–011006, 2010.
- [52] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015.
- [53] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.
- [54] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [56] Kristi Tsukida and Maya R Gupta. How to analyze paired comparison data. Technical report, 2011.

- [57] Aliaksei Mikhailiuk, María Pérez-Ortiz, and Rafal Mantiuk. Psychometric scaling of tid2013 dataset. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.
- [58] Maria Perez-Ortiz and Rafal K Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017.
- [59] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.
- [60] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pages 219–224, 2015.
- [61] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016.
- [62] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.
- [63] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.
- [64] Ann Marie Rohaly, Philip J Corriveau, John M Libert, Arthur A Webster, Vittorio Baroncini, John Beerends, Jean-Louis Blin, Laura Contin, Takahiro Hamada, David Harrison, et al. Video quality experts group: Current results and future directions. In *Visual Communications and Image Processing 2000*, volume 4067, pages 742–753. SPIE, 2000.
- [65] Fangfang Hong, Ruby Bouhassira, Jason Chow, Craig Sanders, Michael Shvartsman, Phillip Guan, Alex H Williams, and David H Brainard. Comprehensive characterization of human color discrimination thresholds. *eLife*, 14, 2025.
- [66] Amin Nejatbakhsh, Isabel Garon, and Alex Williams. Estimating noise correlations across continuous conditions with wishart processes. *Advances in Neural Information Processing Systems*, 36:54032–54045, 2023.

A Architecture and Methodology Details

Feature Encoders. We evaluate three frozen encoder families, each spatially pooled to a manageable feature dimension D before entering the observer head. **CORnet-S** [9] provides four ventral-stream-aligned areas (V1, V2, V4, IT) with channel counts 64, 128, 256, 512; each area is global-average-pooled (GAP) to a vector, and CORnet-S-Multi concatenates the four GAP vectors ($D = 960$). **VGG-16** [32] (ImageNet-pretrained) supplies post-block-3 and post-block-5 features, each GAP-pooled (**VGG-conv3**, $D = 256$; **VGG-conv5**, $D = 512$); their LPIPS-style five-block GAP concatenation [3] forms **VGG-multi** ($D = 1,472$); we additionally retain 4×4 and 2×2 spatial grids on conv3 and conv5 to test whether pooling is the bottleneck (**VGG-conv3** (4×4), $D = 4,096$; **VGG-conv5** (2×2), $D = 2,048$). **Contrast Pyramid** [36, 41] calculates the contrast pyramid of an image using Laplacian pyramid. We use the implementation in ColorVideoVDP [16] and stops before the CSF weighting step; we GAP-pool within each (channel, band) combination ($D = 24$). No encoder is fine-tuned; per-dataset features are computed once and cached on disk for reuse across the full N -sweep.

Covariance Matrix Field MLP. We parameterize the per-image covariance $\Sigma(x)$ of Eq. (1) via its Cholesky factor to guarantee positive definiteness:

$$\Sigma(x) = \mathbf{L}(x) \mathbf{L}(x)^\top \in \mathbb{S}_+^n, \quad \mathbf{L}(x) \in \mathbb{R}_{\text{lower-tri}}^{N \times N}.$$

The $N(N+1)/2$ entries of $\mathbf{L}(x)$ are produced by a small MLP G_ψ ,

$$G_\psi : \mathbb{R}^N \xrightarrow{\text{Linear}(N, 64)} \text{ReLU} \xrightarrow{\text{Linear}(64, 64)} \text{ReLU} \xrightarrow{\text{Linear}(64, N(N+1)/2)} \mathbb{R}^{N(N+1)/2},$$

with $\text{softplus}(\cdot) + \varepsilon$ ($\varepsilon = 10^{-4}$) applied to the diagonal entries of $\mathbf{L}(x)$ so that each diagonal is strictly positive and $\Sigma(x)$ is non-degenerate. The input to G_ψ is the projected mean $\boldsymbol{\mu}(x) \in \mathbb{R}^N$.

MC Sampling and Reparameterization. For Eq. 8 to be differentiable, the samples themselves must be differentiable. We achieve this using the standard reparameterization trick [40] instead of sampling $z(x)$ directly, we draw standard normal noise ε and push it through $\mathbf{L}(x)$.

In particular, on every trial, the model estimates $\Pr(x_0 \mid r, x_0, x_1)$ of Eq. 8 by the following sequence. We first compute the three means $\boldsymbol{\mu}(r)$, $\boldsymbol{\mu}(x_0)$, $\boldsymbol{\mu}(x_1)$ and the three Cholesky factors $\mathbf{L}(r)$, $\mathbf{L}(x_0)$, $\mathbf{L}(x_1)$ once per triplet, then sample neural responses K times through MC simulation:

$$z(x)^{(k)} = \boldsymbol{\mu}(x) + \mathbf{L}(x) \boldsymbol{\varepsilon}^{(k)}, \quad \boldsymbol{\varepsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n). \quad (9)$$

For each MC index $k = 1, \dots, K$, we draw three standard Gaussians $\varepsilon_r^{(k)}, \varepsilon_0^{(k)}, \varepsilon_1^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, form the two candidate-reference differences $z(x_0)^{(k)} - z(r)^{(k)}$ and $z(x_1)^{(k)} - z(r)^{(k)}$ via Eq. 9, compute the weighted p -norms $d_0^{(k)}, d_1^{(k)}$ from Eq. 2, and record the indicator $\mathbf{1}\{d_0^{(k)} < d_1^{(k)}\}$. Averaging these K simulations yields an estimate of $\Pr(x_0 \mid r, x_0, x_1)$.

STE. The indicator $\mathbf{1}\{d_0^{(k)} < d_1^{(k)}\}$ in Eq. 8 is not differentiable, so optimizing Θ by gradient descent requires a gradient estimator. We use the popular STE method to allow gradient flow [39]. An alternative is to replace the hard indicator function with a smooth proxy, the sigmoid function being a popular choice in the literature: $\sigma(\tau(D_1 - D_0))$, where $\sigma(\cdot)$ is the sigmoid function and τ is a learnable temperature controlling how close the smooth function is to the indicator function.

Empirically, we find that STE and the sigmoid method give similar results. We choose the STE method since with the sigmoid, the covariance matrix field becomes harder to interpret. This is because the sigmoid function, being smooth, directly gives a probability measure even with just a single sample each from the three distributions ($Z(x_0)$, $Z(x_1)$, and $Z(r)$). Therefore, the network would have no incentive to learn the covariance matrix field and rely on learning only the mean field (with a very narrow, spatially uniform covariance matrix field).

Parameter Count. Our pipeline factorizes into a *frozen* feature encoder and a small *trainable* head; only the head is updated during training, while the backbone is loaded from public pretrained weights and never modified. Table 4 reports trainable / frozen / total parameter counts for our model and the baseline methods used in the the main text. At its earliest saturation point ($N=3$, beyond which validation accuracy on BAPPS no longer improves) and at the BAPPS peak of the Multi-layer backbone reported in Table 2 ($N=12$), our trainable head is between one and four orders of magnitude smaller than every other learnable metric in the table, and never larger than 25K parameters.

Table 4: Parameter counts comparison.

Method	Trainable	Frozen	Total
DISTS	2,950	14.7M (VGG-16)	14.7M
MILO	45K	—	45K
TOPIQ-FR (CFANet, ResNet-50 end-to-end)	36.0M	53K (Normalization)	36.0M
DreamSim (LoRA $r=16$ on $3 \times \text{ViT-B/16}$)	1.77M	264.2M	266.0M
PieAPP	68.4M	—	68.4M
CORnet-S-Multi ($N=3$)	7.7K	52.9M	52.9M
CORnet-S-Multi ($N=12$)	21.6K	52.9M	52.9M

DCT Sparsity Control. The control reported in Section 5.2 asks whether the small saturated dimensionality n^* of our observer simply reflects the well-known compressibility of natural images: if a top- K -coefficient reconstruction in some natural image basis already matches the observer’s agreement at the matched budget $K = N/3$, then the low n^* is a property of the image sparsity, not of perceptual processing.

Concretely, for each input image, we first transform it to the YUV color space, apply the DCT, and then keep only the top K coefficients in each color channel. We then apply the inverse DCT to reconstruct the compressed image. The reconstructed image is then fed to the same feature encoder as in the main observer model, and the resulting features are passed to a learnable but square $D \times D$ projection (i.e., no dimensionality reduction) followed by the same covariance predictor as in the main model. We sweep $K \in \{2, 10, 20, 200\}$ per color channel. For a 256×256 image, each channel has $H \cdot W = 65,536$ real DCT coefficients, so the swept budgets correspond to retention ratios from 0.003% to 0.3%. The results are shown in Table 3.

Datasets. We mainly evaluate on three triplet 2AFC datasets (BAPPS, NIGHTS, PieAPP) (Section 4). We also evaluate on four other popular low-level IQA datasets, which either do not use 2AFC tasks (LIVE Release 2 [49], KADID-10k [50], and CSIQ [51]) or do not provide raw pairwise comparison data (TID2013 [52])². They do, however, provide a quality score for each image in the form of Mean Opinion Score (MOS) or Differential MOS (DMOS). We will show in Appendix E how our observer model can be extended to modeling rating-based datasets. Here, we describe the details of these datasets pertaining to our study. All datasets are used as released without filtering or relabeling.

BAPPS [3] is a low-level 2AFC triplet dataset of 64×64 image patches. Each trial shows a reference (sampled from MIT-Adobe 5K [59] for training and RAISE1K [60] for validation) and two corrupted candidates, and the annotator picks the one perceptually closer to the reference. Distortions span two families: traditional image-processing artifacts (photometric shifts, additive noise, Gaussian and bilateral blur, spatial warps, JPEG) and CNN artifacts produced by randomly parameterized autoencoders trained on denoising / colorization / super-resolution. Labels were collected with 2 annotators per triplet on the 151K-triplet training split and 5 annotators per triplet on the 36K-triplet validation split. The validation set is further stratified into six distortion subsets (traditional, CNN, superresolution, video deblurring, frame interpolation, colorization), which we report separately. The raw choice distribution is reported for each triplet.

NIGHTS [4] is a high-level 2AFC triplet dataset of 768×768 images synthesized by Stable Diffusion v1.4 from prompts of the form “an image of a $\langle \text{category} \rangle$ ”. Within a triplet the reference and two candidates share the same prompt but use different random seeds, so variation arises from the stochasticity of the diffusion sampling process rather than any explicit low-level distortion. Up to ten annotators rated each triplet, yielding a single binary label per triplet through majority vote; the released split contains 15,941 training, 1,958 validation, and 2,120 test triplets ($\approx 20\text{K}$ total).

PieAPP [5] is a low-level 2AFC triplet dataset built from 200 Waterloo Exploration reference images [61] at 256×256 , each corrupted by a diverse set of image-processing operations and algorithmic artifacts. For each pair of distorted variants of a given reference, approximately 40 annotators judge which is perceptually closer to the reference; the raw choice distribution for each triplet is reported. The official release provides 77,280 train+validation pairs and a 4,200-pair test.

²Mikhailiuk et al. [57] extended TID2013 with additional 2AFC experiments and provide raw pairwise comparison results [22, 58], but the reference images were not shown during the 2AFC tasks.

TID2013 [52] is an FR-IQA dataset with 25 reference images at 512×384 and 3,000 distorted images obtained by applying 24 distortion types at 5 intensity levels per reference. On each trial an observer sees the reference together with two distorted versions and selects the one that differs less from the reference; 971 observers contributed 524,340 such selections. The average votes that an image receives is used as its perceptual score.

KADID-10k [50] is an FR-IQA dataset with 81 references at 512×384 and 10,125 distorted images obtained from 25 distortion types at 5 intensity levels per reference. For each distorted image, the reference and the distorted image are shown side by side and an observer rates the distorted image on a five-point scale from 1 (“very annoying”) to 5 (“imperceptible”). The points are averaged across 30 ratings per image to yield a score for the image.

CSIQ [51] is an FR-IQA dataset with 30 references at 512×512 and 866 distorted images covering six distortion types (JPEG, JPEG-2000, global contrast decrement, additive pink Gaussian noise, additive white Gaussian noise, Gaussian blur) at 4–5 levels each. 35 observers arranged images along a horizontal axis so that physical distances reflect perceived quality differences, producing roughly 5,000 ratings that were converted into perceptual scores.

LIVE Release 2 [49] is an FR-IQA dataset with 29 references (typically 768×512) and 779 distorted images covering five distortion types: JPEG2000 compression, JPEG compression, additive white noise on RGB, Gaussian blur, and JPEG2000 transmission errors over a fast-fading Rayleigh channel. Each distorted image is rated on a continuous scale (with each session run by 20–29 subjects and approximately 23 ratings per image on average) and averaged over subjects to a perceptual score.

B Perceptual Space Interpretation and Visualization

B.1 Perceptual Space Interpretation at $N=2$ on PieAPP and NIGHTS

Figure 4 shows nearest-neighbour retrieval grids at $N = 2$ for PieAPP and NIGHTS. Both datasets recover dimensions similar in spirit to those in BAPPS (Figure 3(c)). On PieAPP, the vertical axis ranges from large-scale repetitive structure (architectural corridors, doorways) at the top to fine cluttered texture (gravel, mechanical parts) at the bottom, and the horizontal axis ranges from warm, colorful outdoor scenes to predominantly muted indoor objects, similar in spirit to BAPPS’s texture and color axes. On NIGHTS, one vertical axis tracks scene multiplicity, with single foreground subjects (a fox, a hamster, a harp) at the top and multiple repeating elements (fences, radiator stripes) at the bottom; the horizontal axis tracks color, with more colored images on the left and near-monochromatic images on the right.

B.2 Perceptual Space Interpretation at High Dimensionalities

When the dimensionality increases, interpreting the dimensions becomes more difficult, since each dimension carries more fine-grained meaning that might not even be easily described in natural language. We use a different method. Since our perceptual space $\mathbb{R}^N$ is an affine transformation from the space given by the feature encoder $\mathbb{R}^D$, each perceptual dimension is a weighted sum of the feature dimensions. We first characterize the concept extracted by each feature dimension using the classic work BRODEN by Bau et al. [62]. From the weights in the learned projection matrix, we can then aggregate the meaning associated with each dimension in the perceptual space.

Interpreting Each CORnet-S Channel. BRODEN classifies the visual concept extracted by each feature into five main categories each with a few detailed sub-categories associated with it: color, material, texture, part, and object. We find the top-3 sub-categories associated with each CORnet-S channel. Each subplot in Figure 5 shows the number of CORnet-S channels associated with each sub-category (see implementation details later). We see that V1 channels predominantly select for color and low-level texture; V2 is dominated by texture; V4 begins producing parts and object fragments (wheel, head, ear); IT is dominated by whole objects.

Perceptual Dimension Interpretation. The learned projection matrix $\mathbf{W} \in \mathbb{R}^{D \times N}$ identifies which of these feature channels each perceptual dimension reads from. Figure 6 shows the projection matrix for the three datasets, BAPPS at $N = 4$, PieAPP at $N = 4$, and NIGHTS at $N = 10$, using CORnet-S Multi. The project matrices are row normalized and visualized as heatmaps. BAPPS and PieAPP models do not rely on area IT, whose contribution becomes more significant for NIGHTS.



Figure 4: Visualizing the learned $N=2$ perceptual space under CORnet-S Multi across datasets. Each cell of the 9×9 grid is an image from the dataset situated near that point in the learned space, so traversing horizontally and vertically shows how the model’s representation morphs as one of the two latent dimensions varies low to high. Across the three datasets, the two axes consistently expose a texture/structure direction crossed with a color (PieAPP) or scene-content (NIGHTS) direction. See Figure 3(c) for results on BAPPS.

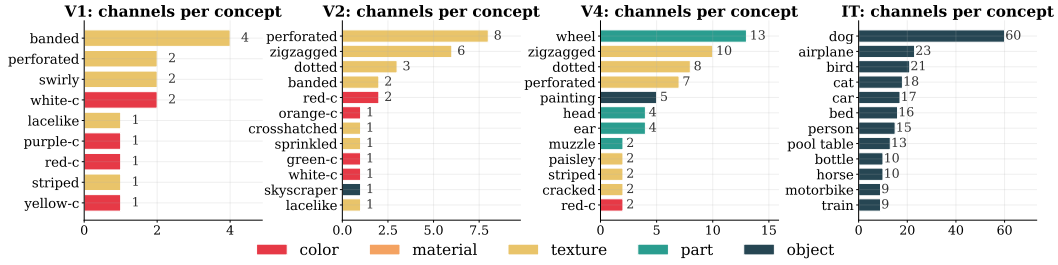


Figure 5: BRODEN [62] concept distribution per CORnet-S layer at the ratio $\geq 5\times$ alignment threshold. Top: fraction of channels whose top concept lies in each pixel-level category. Bottom: per-layer top concepts ranked by the number of channels claiming them.

The level of the ventral stream recruited by the perceptual space therefore tracks the semantic level of the dataset.

Each row of the project matrix represents a perceptual dimension. We pick the top-20 feature channels of each dimension, find the top-3 concept associated with each channel, and aggregate across all the channels to plot a concept distribution of each perceptual dimension. The distribution plots are visualized alongside the projection matrices.

From the distribution plots, we can see that at higher N the concept associated with of each dimension becomes more diverse and mixed: BAPPS and PieAPP concentrate weight on V1, V2, and V4, and their dimensions remain dominated by color and texture; NIGHTS develops a substantial dependence on IT, and its dimensions reach part- and object-level semantics. Higher dimensionality is therefore not redundant repetition but a finer partition of the same hierarchy into specialized subspaces.

Implementation Details. For each CORnet-S channel j , we binarize its activation at the top-0.5% threshold (estimated on a 5,000-image subsample, matching Bau et al. [62]) and upsample to BRODEN’s 112×112 resolution to obtain mask $M_j(x)$. For each concept c with ground-truth mask $L_c(x)$, we compute $\text{IoU}(j, c) = \sum_x |M_j \cap L_c| / \sum_x |M_j \cup L_c|$ and divide by a closed-form random-IoU baseline; this ratio puts color and texture on the same scale and is what we call a channel’s



Figure. 6: Projection structure of the learned perceptual dimensions. We visualize the projection matrix for BAPPS ($N = 4$), PieAPP ($N = 4$), and NIGHTS ($N = 10$). The heatmaps show row-normalized projection magnitudes $|W_{k,j}|$, where each row corresponds to one learned perceptual dimension and each column to one of the 960 CORnet-S-Multi channels. Dashed vertical lines mark the boundaries between V1, V2, V4, and IT channels. For each row, the bar on the right summarizes the BRODEN category composition of the top-20 weighted channels. The five segments, color-coded according to the bottom legend, correspond to color, material, texture, part, and object, and sum to one. Together, the heatmaps indicate where each dimension reads from in the visual hierarchy, while the side bars summarize what visual content those channels encode.

selectivity for the concept. A channel is considered aligned with a concept when its selectivity passes $5\times$. The dissection produces a top-3 concept list per channel ranked by raw IoU; we take the entry with the largest selectivity as the channel’s representative concept, since selectivity reflects discriminative power better than raw overlap. For each CORnet-S-Multi checkpoint with projection $\mathbf{W} \in \mathbb{R}^{n \times 960}$, we sort channels by $|W_{k,j}|$, take the top 20, and aggregate their representative concepts into the category-mix bars in Figure 6.

B.3 Dimensions are Independent

Finally, we empirically demonstrate that the learned perceptual dimensions are mutually orthogonal within the perceptual space, a property that is not guaranteed a priori. We use a method similar to that in Zbontar et al. [63]: each image in the dataset, assuming that are I of them, is mapped to a N -dimensional vector in the perceptual space $\mathbb{R}^N$, form a $N \times I$ matrix; in other words, each dimension is represented by a V -dimensional vector. We calculate the pairwise Pearson correlation between every dimension pair, and plot the results in Figure 7. Across all N s and all datasets, the

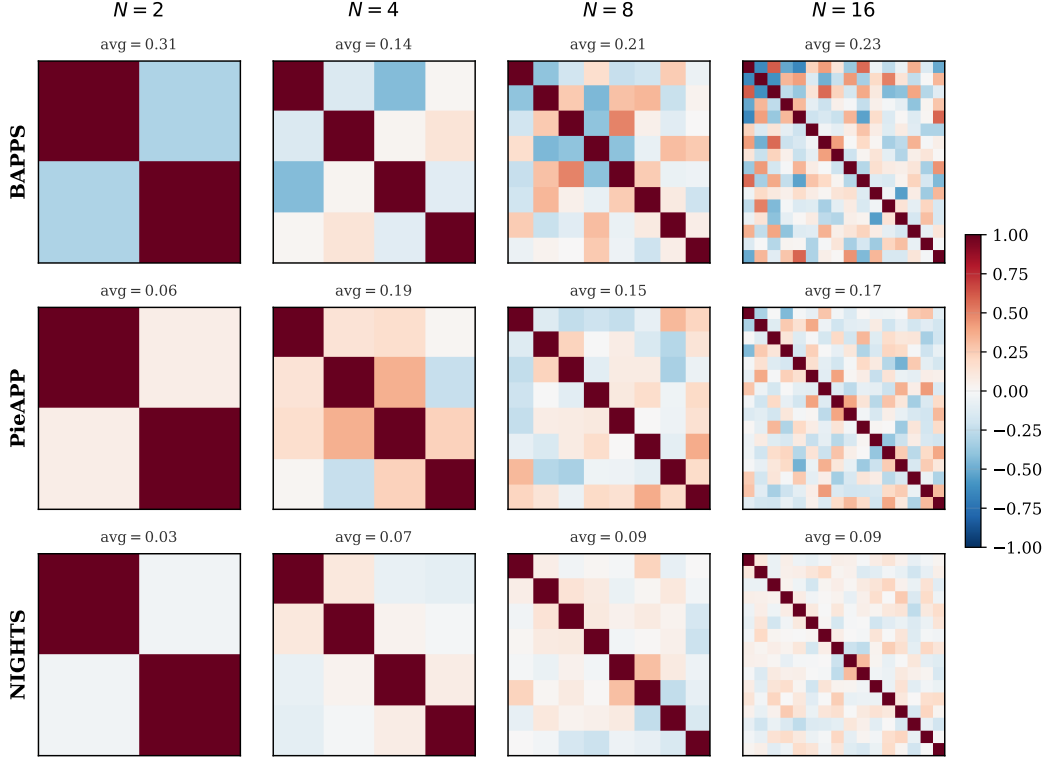


Figure. 7: Pairwise Pearson correlation between learned dimensions on the validation set. Rows are BAPPS, PieAPP and NIGHTS. Columns are $N = 2, 4, 8, 16$. The number above each cell is the average off-diagonal coefficient.

dimensions are largely independent of each other, with an average off-diagonal correlation coefficient smaller than 0.25.

C Results on Different Feature Encoders and Projection Choices

Table 5 reports the prediction performance under different feature encoders. We also evaluate two dimensionality methods: learned projection (L), which is used in the main text, and PCA (P). Shaded rows indicate the encoders used in the main text and all cells show peak performance and the corresponding saturating n^* .

Table 5: Full in-domain results across three benchmarks under two projection configurations (L = Learned, P = PCA). BAPPS: agreement/ n^* and KL/ n^* (0 = oracle). NIGHTS: accuracy/ n^* . PieAPP: agreement/ n^* . Theoretical ceilings reported in main-text Table 1.

Encoder	BAPPS Agr / n^*		BAPPS KL / n^*		NIGHTS Acc / n^*		PieAPP Agr / n^*	
	L	P	L	P	L	P	L	P
VGG-conv3 (4×4)	.694/6	.672/10	.167/6	.182/14	.804/48	.780/20	.711/8	.679/4
VGG-conv3	.695/4	.670/10	.161/4	.181/12	.808/16	.783/14	.715/6	.678/9
VGG-conv5 (2×2)	.672/4	.651/8	.182/4	.201/6	.799/96	.780/20	.690/10	.679/12
VGG-conv5	.674/4	.654/8	.178/5	.199/7	.830/96	.803/48	.687/4	.674/7
VGG-multi	.700/4	.674/9	.155/4	.179/10	.844/28	.821/48	.715/4	.685/7
CORnet-S V1	.685/4	.675/14	.175/6	.181/14	.747/8	.738/7	.701/6	.682/14
CORnet-S V2	.689/3	.678/14	.166/5	.174/16	.793/14	.781/12	.712/5	.682/20
CORnet-S V4	.688/4	.662/8	.162/8	.186/8	.855/48	.827/28	.714/8	.681/3
CORnet-S IT	.677/6	.665/14	.180/6	.188/10	.853/96	.848/96	.702/14	.683/14
CORnet-S Multi	.688/3	.664/10	.167/4	.188/9	.859/96	.848/96	.710/9	.685/20

The conclusion from Section 5.3—that the perceptual space underlying low-level quality judgments is much lower-dimensional than the perceptual space underlying high-level judgments—holds across both projection configurations and across every encoder we evaluate. Under the learned projection, n^* on BAPPS sits in the 3–6 range under both the agreement and KL divergence metrics, so the saturation diagnosis does not depend on a particular choice of metric. PCA’s performance is universally lower than that of the learned projection, showing the importance of the latter.

D Results on Cross-Dataset Prediction

Figure 8 quantifies the cross-dataset claim of Section 5. For each encoder we train the observer on BAPPS, freeze it, and evaluate it on a different dataset without any further fitting. The brown dashed line in each panel is the best in-domain peak on the BAPPS validation set. Transferring to PieAPP (Figure 8(a)), another low-level 2AFC dataset, costs the high-level CORnet-S encoder (V4, IT, multi-layer) almost nothing: their curves almost meet the in-domain ceiling. Transferring to NIGHTS (Figure 8(b)), in contrast, leaves every encoder several points below the in-domain ceiling, and the residual gap does not close as N grows or as we move up the visual hierarchy.

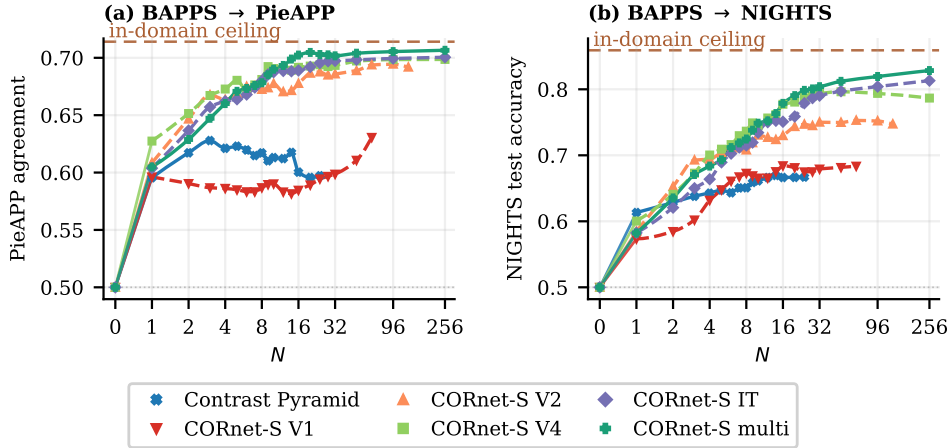


Figure. 8: Cross-dataset transfer of a BAPPS-trained observer across six feature encoders. (a): BAPPS $\rightarrow$ PieAPP using per-annotator agreement on PieAPP as the accuracy metric. (b): BAPPS $\rightarrow$ NIGHTS using 1-BER on the NIGHTS test set. The brown dashed line in each panel is the best in-domain ceiling on the test-side dataset.

Table 6 reports the corresponding peaks and saturation points across all encoders. The cross-domain pattern mirrors the in-domain pattern, with n^* on the target dataset often higher than its in-domain counterpart.

Table 6: Full cross-domain results under the default Learned projection with $\text{Cov} \leftarrow \mu(x)$. B $\rightarrow$ N: trained on BAPPS, tested on NIGHTS (accuracy/ n^*). N $\rightarrow$ B: trained on NIGHTS, tested on BAPPS (agreement/ n^*). B $\rightarrow$ P: trained on BAPPS, tested on PieAPP (agreement/ n^*). Theoretical ceilings reported in main-text Table 1. Shaded rows are the encoders used in the main text.

Encoder	B $\rightarrow$ N: NIGHTS Acc / n^*	N $\rightarrow$ B: BAPPS Agr / n^*	B $\rightarrow$ P: PieAPP Agr / n^*
VGG-conv3 (4×4)	.751/24	.671/20	.703/16
VGG-conv3	.772/20	.657/14	.694/6
VGG-conv5 (2×2)	.724/20	.647/48	.681/28
VGG-conv5	.749/16	.638/9	.683/20
VGG-multi	.808/96	.675/8	.693/7
CORnet-S V1	.690/16	.654/1	.627/3
CORnet-S V2	.748/12	.661/5	.699/16
CORnet-S V4	.798/20	.668/14	.702/28
CORnet-S IT	.827/256	.663/24	.703/20
CORnet-S Multi	.834/96	.670/24	.709/16

E Extension to Rating-Based IQA Datasets

Our observer model can be readily extended to rating-based IQA datasets, where each image is associated with a perceptual score. In our setting, the *expected* Minkowski distance between an image’s distribution and that of the reference image in the learned perceptual space can be treated as the perceptual score of the image.

We train the observer model (CORnet-S-Multi encoder, $N=24$) end-to-end on the full KADID-10k MOS labels with an $L_1 + (1-\rho)$ loss, where ρ is the Pearson correlation between predicted and ground-truth quality. A two-parameter sigmoid score head $q = \sigma(a \log(1+d) + b)$ with a, b learnable maps the weighted Minkowski distance d of Eq. 2 to a scalar predicted MOS-badness $q \in [0, 1]$; this affine head reshapes the heavy tail of d without breaking monotonicity. Per-pair distances are computed by Monte Carlo with $K=512$ samples; the Spearman rank-order correlation coefficients (SRCC), Pearson linear correlation coefficients (PLCC), and Kendall rank correlation coefficients (KRCC) are reported as absolute values per the convention, with PLCC computed after the 4-parameter logistic remapping recommended by the Video Quality Experts Group (VQEG) [64] to align predicted scores with the subjective MOS scale before linear comparison.

We evaluate the resulting model’s zero-shot on three standard rating-based IQA benchmarks: TID2013, CSIQ, and LIVE Release 2, all of which contain traditional algorithmic distortions (JPEG, JPEG2000, blur, noise, colour/contrast shifts) similar in kind to KADID-10k. KADID-10k itself is excluded from the comparison because our MOS-fit model is trained on its full set; for the same reason, TOPIQ-FR’s released checkpoint, which was trained by Chen et al. [13] on the full KADID-10k and is the public benchmark used here, would also produce training-set numbers on KADID-10k and is omitted on that dataset. Baseline numbers for PSNR, SSIM, LPIPS-VGG, DISTS, DreamSim, PieAPP, CVVDP, MILO, and TOPIQ-FR are computed with the same evaluation protocol.

Table 7: The observer model extends to rating-based IQA datasets with absolute (D)MOS labels. Each cell reports SRCC/PLCC/KRCC; higher is better for all three. **Bold** = best, underline = second & third best, ranked independently per (dataset, metric) column.

Method	TID2013	CSIQ	LIVE
CORnet-S Multi	<u>0.860 / 0.870 / 0.666</u>	<u>0.946 / 0.948 / 0.796</u>	0.937 / <u>0.927</u> / 0.781
MILO	0.876 / 0.876 / 0.692	0.959 / 0.942 / 0.817	0.953 / 0.913 / 0.817
TOPIQ-FR	0.917 / 0.921 / 0.744	0.967 / 0.964 / 0.838	0.898 / 0.901 / 0.717
CVVDP	0.853 / 0.864 / 0.672	0.896 / 0.883 / 0.728	0.917 / 0.902 / 0.766
DreamSim	0.809 / 0.830 / 0.612	0.932 / 0.941 / 0.768	0.899 / 0.902 / 0.729
PieAPP	0.850 / 0.846 / 0.652	0.845 / 0.829 / 0.666	0.925 / 0.915 / 0.765
DISTS	0.708 / 0.755 / 0.521	0.930 / 0.938 / 0.764	0.955 / 0.954 / 0.814
LPIPS-VGG	0.670 / 0.749 / 0.497	0.883 / 0.905 / 0.697	<u>0.943 / 0.945 / 0.791</u>
SSIM	0.710 / 0.702 / 0.511	0.813 / 0.804 / 0.613	0.927 / 0.910 / 0.762
PSNR	0.687 / 0.677 / 0.496	0.809 / 0.816 / 0.599	0.925 / 0.924 / 0.756

Results. Table 7 shows the results. Across all three datasets and all three metrics, our method often ranks among the top three compared the baselines. This shows that our observer model can well account for human data even on rating-based IQA datasets.

We summarize the results on both triplet 2AFC datasets (Table 1) and the rating-based IQA datasets (Table 7) into a single table Table 8, where we show the average ranking of a method across all six datasets. Our method maintains strong performance across both settings. It achieves the best average rank of 2.67 and ranks within the top three on five of the six datasets.

Saturation Dimensionality Persists. Figure 9 plots SRCC at every swept N for the CORnet-S Multi observer reported in Table 7, on KADID-10k val (20% by-reference held-out) and on the two zero-shot test sets, TID2013 and CSIQ. The low-dimensional saturation pattern observed in the main text holds. The curve of KADID-10k saturates at $n^* = 3$, while TID2013 and CSIQ saturate at $n^* = 2$; SRCC then fluctuates by less than ± 0.005 across $n \in \{24, 32, 48, 96, 256\}$. Because TID2013, CSIQ, and KADID-10k all consist of traditional low-level distortions (JPEG, JPEG2000, blur, noise, color/contrast shifts), this results here corroborate the findings on the triplet 2AFC tasks (Section 5.3).

Table 8: Per-dataset rank of each method across the six datasets covered by Tables 1 and 7, restricted to the nine baselines that appear in both. Score per dataset: BAPPS & PieAPP use per-annotator agreement; NIGHTS uses accuracy; TID2013 & CSIQ & LIVE use SRCC. Lower rank is better; the right-most column is the mean across the six datasets.

Method	BAPPS	NIGHTS	PieAPP	TID2013	CSIQ	LIVE	avg. rank
CORnet-S Multi	1	2	3	3	3	4	2.67
DreamSim	2	1	1	6	4	9	3.83
DISTS	3	3	4	8	5	1	4.00
TOPIQ-FR	4	7	5	1	1	10	4.67
MILO	8	9	7	2	2	2	5.00
LPIPS	5	4	6	10	7	3	5.83
CVVDP	6	5	8	4	6	8	6.17
PieAPP	10	6	2	5	8	6	6.17
SSIM	7	8	9	7	9	5	7.50
PSNR	9	10	10	9	10	7	9.17

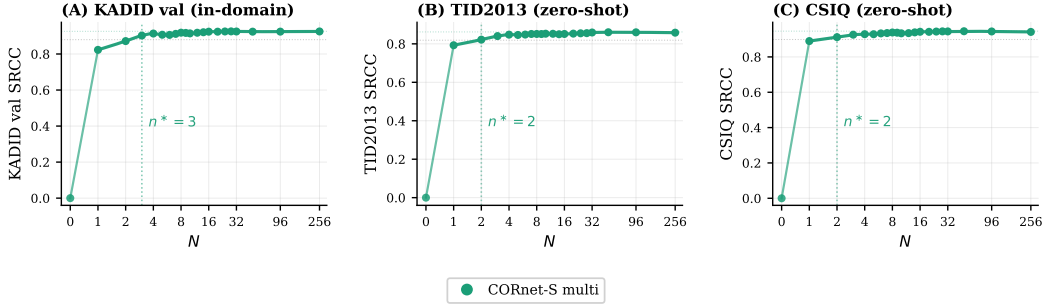


Figure. 9: Saturation of the KADID fitted CORnet-Multi observer model ($N=24$ in Table 7) along the sweep schedule. (A) KADID-10k val (20% by-reference held-out). (B) TID2013 (zero-shot). (C) CSIQ (zero-shot). The vertical band marks the 95%-of-peak saturation point n^* , computed against an SRCC chance baseline of 0.

F Comparison with Bradley–Terry Model

The BT model [19] is commonly used to relate quality scores of images/objects to choice probabilities in pairwise comparisons. Given two images, each with a quality score s_i and s_j , the probability of choosing i over j (under a reference r) is modeled to be: $\Pr(i|r, i, j) = \frac{1}{1+e^{s_j-s_i}}$.

One can derive this model by assuming that the quality score of each image is a random variable S_i , which has mean s_i and follows a certain distribution: $S_i \sim \mathcal{D}$. In a pairwise comparison, the probability that image i is picked over image j (with respect to the reference r) is then the expectation that $S_i - S_j > 0$. That is, $\Pr(i | r, i, j) = \mathbb{E}[\mathbf{1}\{S_j < S_i\}]$. One can show that when the distribution $\mathcal{D}$ is a Gumbel distribution with a constant standard deviation, this formulation is exactly the BT model [56]; when $\mathcal{D}$ is a Gaussian distribution with, again, a constant standard deviation, this formulation is equivalent to the Thurstone model Case V [20].

Note the similarity between this formulation and Eq. 3. The near-identical forms are not coincidental; it is because our observer model, while modeling the distribution of high-dimensional, internal responses of an image, can also be equivalently interpreted as modeling the one-dimensional, perceptual-score distribution of an image. The score can be given by the Minkowski distance between the image and the reference in the perceptual space (Eq. 2). The distribution of this perceptual score, however, is neither Gumbel nor Gaussian, even if the perceptual space itself is Gumbel or Gaussian, as it involves calculating the Minkowski distance. Alternatively, one can, like Schmidt et al. [25], map each image into an embedding and use the cosine similarity between the embedding of an image and that of a reference as the score, and simply assume that such a score is a drawn from a Gumbel random variable.

Our implementation of the BT model uses the latter approach. Specifically, we strip the multi-dimensional, image-dependent noise from the observer model while keeping the N -dimensional latent embedding $\mu(x)$ and the learned projection $\mathbf{W}$ intact, so that the only change from the full observer model is the noise parameterization. We then replace the weighted Minkowski distance of Eq. 2 with the cosine similarity between each candidate’s latent and the reference’s embedding. Writing $s_i = \tau \cos(\mu(r), \mu(x_i))$ for the score of candidate i scaled by a learnable temperature τ , the probability of choosing x_0 is simply $\frac{1}{1+e^{s_1-s_0}}$, as given by the BT model. The underlying assumption is that s_0 and s_1 both follow a Gumbel distribution that differ only in the mean.

Table 9 and 10 compare the two methods (our full observer model and the BT model) across 12 feature encoders, both in-domain and cross-domain. Experimentally, we find that the BT model and our method give similar predictive power. By using physiologically grounded feature encoders, however, our method permits examining (the dimensions of) the internal perceptual space, which the original BT model is incapable of³, and can be thought of as giving a low-level interpretation of the phenomenological BT model.

Table 9: BT vs. full observer model, in-domain. $\Delta = \text{BT} - \text{Obs}$. BAPPS: per-annotator agreement (Obs wins 9/12, avg $\Delta = -0.4\%$). NIGHTS: 2AFC accuracy (BT wins 4/12, avg $\Delta = +0.5\%$).

Backbone	D	BAPPS (Agr)				NIGHTS (Acc)					
		BT	Obs	Δ	BT n^*	Obs n^*	BT	Obs	Δ	BT n^*	Obs n^*
CORnet-S V1	64	.686	.685	+.001	4	4	.789	.747	+.042	14	8
CORnet-S V2	128	.691	.689	+.002	7	3	.807	.793	+.014	16	14
CORnet-S V4	256	.687	.688	-.001	6	4	.838	.855	-.017	96	48
CORnet-S IT	512	.677	.677	+.000	8	6	.848	.853	-.005	96	96
CORnet-S Multi	960	.685	.688	-.003	6	3	.857	.859	-.002	96	96
VGG-conv3	256	.688	.695	-.008	5	4	.822	.824	-.003	96	24
VGG-conv5	512	.671	.674	-.003	6	4	.850	.828	+.022	256	96
VGG-conv3 (4×4)	4096	.679	.694	-.015	5	6	.789	.794	-.005	12	28
VGG-conv5 (2×2)	2048	.664	.672	-.009	8	4	.842	.795	+.048	96	32
VGG-multi	1472	.686	.700	-.013	5	4	.841	.852	-.011	28	28
Average				-.004					+.005		

Table 10: BT vs. full observer model, cross-domain. $\Delta = \text{BT} - \text{Obs}$. B→N: 2AFC accuracy (BT wins 9/12, avg $\Delta = +2.0\%$). N→B: per-annotator agreement (BT wins 6/12, avg $\Delta = +0.3\%$).

Backbone	D	B→N (Acc)			N→B (Agr)		
		BT	Obs	Δ	BT	Obs	Δ
CORnet-S V1	64	.722	.687	+ .035	.661	.654	+ .007
CORnet-S V2	128	.757	.759	− .002	.656	.661	− .005
CORnet-S V4	256	.830	.803	+ .026	.661	.668	− .007
CORnet-S IT	512	.837	.806	+ .031	.664	.663	+ .001
CORnet-S Multi	960	.836	.824	+ .012	.665	.670	− .005
VGG-conv3	256	.790	.766	+ .024	.646	.657	− .011
VGG-conv5	512	.809	.745	+ .064	.654	.638	+ .016
VGG-conv3 (4×4)	4096	.719	.733	− .014	.647	.671	− .024
VGG-conv5 (2×2)	2048	.762	.718	+ .044	.653	.647	+ .006
VGG-multi	1472	.820	.805	+ .016	.654	.675	− .021
Average				+ .020			+ .003

G MAP Using Wishart Prior on the Covariance

The main text fits the observer model by maximum likelihood. One can also apply a prior on the covariance matrix field. In particular, the Wishart distribution is widely used as a probabilistic model for covariance [65, 66]. We impose a weak prior that each covariance matrix is a sample from a Wishart distribution, and integrate the prior with the likelihood function using maximum a posteriori (MAP) estimation.

³We make a distinction between the original BT model [19], where each image/object is directly associated with a score that is directly estimated through maximum likelihood, and the BT model augmented by a multi-dimensional observer model, like the one we present here, where the score is derived from a high-dimensional perceptual space. In the latter case, one could in principle still study the structure of the perceptual space (e.g., the saturating dimensionality).

G.1 MAP with Wishart Distribution Prior

The Wishart distribution $\mathcal{W}_n(\mathbf{V}, \nu)$ is a distribution over $n \times n$ positive semi-definite matrices, parameterized by a scale matrix $\mathbf{V} \in \mathbb{S}_+^n$ and degrees of freedom $\nu \geq n$. Its probability density function is:

$$p(\Sigma) = \frac{|\Sigma|^{(\nu-n-1)/2} \exp(-\frac{1}{2}\text{tr}(\mathbf{V}^{-1}\Sigma))}{2^{\nu n/2} |\mathbf{V}|^{\nu/2} \Gamma_n(\nu/2)}, \quad (10)$$

where $\Gamma_n(\cdot)$ is the multivariate gamma function $\Gamma_n(a) = \pi^{n(n-1)/4} \prod_{j=1}^n \Gamma(a - (j-1)/2)$.

We fix $\nu = n + 2$ (the minimal value above $n + 1$ that gives a positive log-det coefficient). We choose $\mathbf{V} = \frac{1}{\nu} \mathbf{I}$, which yields an isotropic prior with mean $\mathbb{E}[\Sigma] = \mathbf{I}$. Both ν and $\mathbf{V}$ can in theory be learned, which we leave to future work.

With the prior, the optimization problem becomes MAP:

$$\hat{\Theta} = \arg \max_{\Theta} (\log p(\mathcal{D} | \Theta) + \lambda \log p(\Theta)), \quad (11)$$

with $\lambda \geq 0$ controlling how strongly the prior is enforced relative to the likelihood (if the Bayes rule is to be strictly followed, $\lambda = 1$). The first term is the log likelihood function in Eq. 7. The second term is the prior defined over the free parameters Θ , and can be trivially derived by plugging Eq. 5 and Eq. 6 into Eq. 10; we omit the exact express for simplicity.

G.2 Results

We compare plain maximum likelihood ($\lambda = 0$, baseline) against MAP with $\lambda = 0.05$ and $\lambda = 1.0$ (strict MAP) across 12 feature encoders. Table 11 shows the results on BAPPS and NIGHTS (in-domain tests), and Table 12 shows the results when a BAPPS-trained model is applied to NIGHTS and vice versa (cross-domain tests; similar to those in Section D). Δ shows the prediction performance gain from the using the prior.

There is a small net negative for in-domain tests (-0.4% BAPPS agreement, $\Delta = -0.0\%$ on NIGHTS for $\lambda = 0.05$ and -0.6% for $\lambda = 1.0$), as expected from a prior that trades training-set fit for less extreme covariances; the dimensionality finding itself is unchanged ($n^* = 3-4$ on BAPPS). The effects on cross-domain tests is roughly neutral on average. Overall, the Wishart prior delivers no consistent gain in either test. This is likely due to the fact that the baseline MLE method is jointly learning both the mean field and the covariance matrix, and any potential effects of the priors on the covariance matrix field could be learned through the basic MLE itself already.

Table 11: Wishart MAP vs. MLE in-domain. $\Delta = \text{MAP} - \text{MLE}$. BAPPS: per-annotator agreement ($\lambda=0.05$: avg $\Delta = -0.4\%$; $\lambda=1$: avg $\Delta = -0.4\%$). NIGHTS: 2AFC accuracy ($\lambda=0.05$: avg $\Delta = -0.0\%$; $\lambda=1$: avg $\Delta = -0.6\%$).

Backbone	BAPPS (Agr)						NIGHTS (Acc)				
	D	MLE	$\lambda=.05$	Δ	$\lambda=1$	Δ	MLE	$\lambda=.05$	Δ	$\lambda=1$	Δ
CORnet-S V1	64	.691	.677	-.014	.676	-.015	.773	.757	-.016	.751	-.022
CORnet-S V2	128	.696	.688	-.008	.688	-.008	.806	.811	+0.004	.799	-.006
CORnet-S V4	256	.695	.690	-.005	.691	-.004	.856	.852	-.004	.851	-.005
CORnet-S IT	512	.681	.681	+0.000	.681	-.000	.847	.847	+0.000	.838	-.008
CORnet-S Multi	960	.697	.694	-.003	.694	-.003	.857	.859	+0.001	.850	-.006
VGG-conv3	256	.695	.692	-.003	.694	-.001	.824	.815	-.008	.813	-.011
VGG-conv5	512	.674	.670	-.004	.674	-.000	.828	.831	+0.002	.822	-.006
VGG-conv3 (4×4)	4096	.694	.692	-.002	.692	-.001	.794	.787	-.006	.787	-.006
VGG-conv5 (2×2)	2048	.672	.671	-.002	.671	-.001	.794	.803	+0.008	.795	+0.000
VGG-multi	1472	.700	.698	-.002	.699	-.001	.851	.849	-.002	.848	-.002
Average				-.004		-.004			-.000		-.006

H Parameters of Weighted Minkowski Distance Across the Visual Hierarchy

Exponent. We use a weighted Minkowski distance (p-norm) to quantify the difference between two images (Eq. 2). In psychophysics, *separable* dimensions—those that observers can attend to and judge independently—yield similarity judgments best described by a city-block metric ($p = 1$), while

Table 12: Wishart MAP vs. MLE cross-domain. $\Delta = \text{MAP} - \text{MLE}$. B $\rightarrow$ N: 2AFC accuracy ($\lambda=0.05$: avg $\Delta = -0.1\%$, wins 6/12; $\lambda=1$: avg $\Delta = -0.0\%$, wins 7/12). N $\rightarrow$ B: per-annotator agreement ($\lambda=0.05$: avg $\Delta = 0.0\%$, wins 6/12; $\lambda=1$: avg $\Delta = +0.3\%$, wins 6/12).

Backbone	D	B $\rightarrow$ N (Acc)					N $\rightarrow$ B (Agr)				
		MLE	$\lambda=.05$	Δ	$\lambda=1$	Δ	MLE	$\lambda=.05$	Δ	$\lambda=1$	Δ
CORnet-S V1	64	.687	.700	+ .013	.687	+ .001	.653	.639	− .014	.641	− .012
CORnet-S V2	128	.759	.759	+ .001	.762	+ .003	.661	.653	− .007	.652	− .008
CORnet-S V4	256	.803	.805	+ .001	.812	+ .008	.667	.658	− .008	.664	− .003
CORnet-S IT	512	.806	.810	+ .005	.808	+ .002	.663	.664	+ .001	.664	+ .001
CORnet-S Multi	960	.824	.826	+ .002	.824	+ .000	.669	.669	− .000	.668	− .001
VGG-conv3	256	.766	.761	− .005	.773	+ .007	.656	.652	− .003	.662	+ .006
VGG-conv5	512	.745	.746	+ .001	.755	+ .010	.637	.637	+ .000	.645	+ .007
VGG-conv3 (4 $\times$ 4)	4096	.733	.715	− .018	.727	− .006	.671	.671	+ .000	.669	− .001
VGG-conv5 (2 $\times$ 2)	2048	.718	.727	+ .009	.722	+ .003	.646	.646	− .000	.646	− .000
VGG-multi	1472	.805	.802	− .003	.803	− .001	.674	.675	+ .000	.675	+ .001
Average				− .001		− .000			+ .000		+ .003

Table 13: Learned weighted Minkowski parameters p and per-channel weights α at the saturation point n^* (in parentheses after p). For α we report both the mean and standard deviation across the n^* learned channels. Underline indicates the best-performing model in each case.

Encoder	BAPPS		NIGHTS	
	$p(n^*)$	$\bar{\alpha} \pm \sigma_\alpha$	$p(n^*)$	$\bar{\alpha} \pm \sigma_\alpha$
CORnet-S V1	1.01 (4)	.08 $\pm$.03	1.66 (8)	.67 $\pm$.03
CORnet-S V2	1.09 (3)	.12 $\pm$.02	1.66 (14)	.72 $\pm$.04
CORnet-S V4	1.28 (4)	.18 $\pm$.03	1.34 (48)	.82 $\pm$.05
CORnet-S IT	2.29 (6)	.40 $\pm$.02	1.63 (96)	.71 $\pm$.01
CORnet-S Multi	1.36 (3)	.22 $\pm$.02	<u>1.51 (96)</u>	<u>.75 $\pm$.02</u>

integral dimensions—those perceived as a fused whole—yield Euclidean metrics ($p = 2$) [29–31]. Mathematically, at $p = 1$, the relative contribution of a dimension to the distance is independent of the magnitudes of the other dimensions, whereas at $p = 2$, the relative contribution of one dimension depends on the others⁴.

Table 13 shows the Minkowski exponent p that the observer model learns at the saturation point n^* for each encoder-dataset combination. On BAPPS, p climbs from values near 1 at V1 to values near 2 at IT. On NIGHTS the same observer settles in a narrower 1.3–1.7 range across the hierarchy. Looking at the best performing model in each dataset, in the low-level dataset BAPPS, the learn p is much lower than that on the high-level dataset NIGHTS, indicating that low-level IQA is better explained by separable dimensions, whereas high-level IQA is better explained by integral dimensions.

Weights. The per-channel weights α in Eq. 2 compensate for the differing natural magnitudes of perceptual channels. At the saturation point n^* , the weights are approximately uniform within each encoder: the coefficient of variation $\sigma_\alpha/\bar{\alpha}$ is at most 8% on BAPPS (V1 is the only exception, at 37% with $n^* = 4$) and at most 6% on NIGHTS, meaning no channel preferentially dominates the distance. This is expected, since any difference across dimensions could potentially be learned through the projection matrix itself.

⁴Consider the L2 distance $d = \sqrt{x^2 + y^2}$; $\partial d/\partial x = x/\sqrt{x^2 + y^2}$, which depends on y . In contrast, for the L1 distance $d = |x| + |y|$, we have $\partial d/\partial x = \text{sign}(x)$, which is independent of y .