JLD: PERCEPTUAL DISTANCE THROUGH A JACOBIAN LENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image compression, restoration, and generation all require a way to measure how different two images look to a person. Pixel error ignores how people see, while the most accurate perceptual distances are typically fitted to human judgments, tying them to a fixed data and resolution. For example, when image resolution is doubled, the correlation of DISTS with human scores on TID2013 drops from 0.815 to 0.717. We introduce the Jacobian Lens Distance (JLD), which derives its perceptual geometry from a frozen vision encoder rather than from human labels. JLD combines the locality of early patch features with the perceptual sensitivity captured by later encoder representations. Specifically, we use the encoder Jacobian to identify directions in the early feature space that most strongly affect the encoder output, producing a fixed metric tensor, $\mathbb{E}[J^\top J]$, which we call the Jacobian lens. The lens is fitted only once from 100 unlabeled images, taking about 35 seconds. Locally, this construction defines a pullback metric in pixel space, giving JLD a clear geometric interpretation that can be directly analyzed on real images. Across four standard perceptual databases, JLD achieves state-of-the-art performance and consistently outperforms LPIPS, DISTS, PieAPP, and DreamSim. JLD is also robust to changes in image resolution, on TID2013, its lens-term correlation remains nearly unchanged when the resolution is doubled, decreasing only from 0.850 to 0.845. We further introduce JLD-fast, which is $4\times$ faster than LPIPS-VGG while achieving a mean correlation of 0.911. Finally, JLD naturally extends to video, reaching a correlation of 0.786 on Waterloo IVC 4K compared with 0.611 for VMAF. Code.

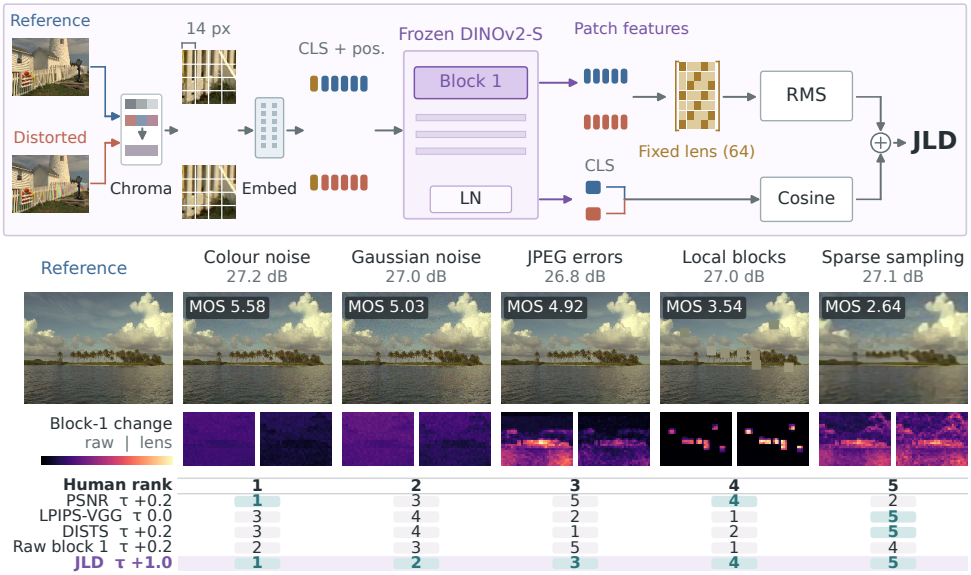

Figure 1: Top: block-1 patch features of the images are projected onto a fixed lens (JLD), fitted from the output's sensitivity without labels; Bottom: five TID2013 distortions at equal PSNR (26.8 to 27.2 dB), ordered by human score, with the per-patch change, and rank and Kendall τ against the human ranks.

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

1 INTRODUCTION

Measuring how different two images look to a person is a basic problem in imaging. Lossy codecs are tuned using perceptual measures, while restoration and generative models are trained and evaluated with them (Wang et al., 2004; Zhang et al., 2018). Pixel error, measured using mean squared error or PSNR, is simple and differentiable, but it treats every pixel change equally, whereas human vision does not (Wang et al., 2004; Ponomarenko et al., 2015). Figure 1 shows five distortions from TID2013 (Ponomarenko et al., 2015) whose PSNR values differ by less than 0.4 dB. Despite their similar pixel error, their perceptual severity is very different. Across all TID2013 pairs with equal PSNR, it selects the image preferred by human observers in only 51.1% of cases, whereas JLD does so in 79.9%. A useful perceptual distance should therefore assign small distances to changes that people barely notice and large distances to changes they find visually important. At the same time, if it is to be used as a loss or an evaluation measure, its geometry should remain well defined and interpretable.

Existing approaches largely fall into two families. Classical full-reference measures explicitly model properties associated with early vision. SSIM and MS-SSIM compare local luminance, contrast, and structure (Wang et al., 2004; 2003); FSIM, GMSD, and VSI compare phase-congruency, gradient, and saliency maps (Zhang et al., 2011; Xue et al., 2014; Zhang et al., 2014). Deep feature distances instead compare representations from pretrained networks. LPIPS, DISTS, PieAPP, and DreamSim, etc., are additionally fitted to human judgements (Zhang et al., 2018; Ding et al., 2022; Prashnani et al., 2018; Fu et al., 2023). This fitting makes such metrics poor at generalizability. When TID2013 is rescored at twice the image resolution, from 256 to 512 pixels, the Spearman correlation (SRCC) of DISTS with ground truth drops from 0.815 to 0.717, while LPIPS-VGG (Zhang et al., 2018; Simonyan & Zisserman, 2015) drops from 0.757 to 0.661. A separate line of work derives perceptual geometry from models of unlabelled data. IEM(Ohayon et al., 2026) integrates denoising errors under a learned density, while LASI(Severo et al., 2024) fits linear embeddings at inference time. These methods require no human labels and provide a principled geometry, but they are considerably more expensive to evaluate. For example, IEM requires 725.7 ms per image pair at 256×256 pixels, compared with 12.2 ms for JLD at 512×384 (Table 1).

In this work, we ask a simple question: *does a network that has only learned to see already contain a useful perceptual distance?* We find that it does, but the relevant information does not appear at any single network depth (Appendix Figure A2). In a frozen DINOv2-S encoder (Oquab et al., 2024), early patch tokens preserve local image structure. However, treating all 384 feature directions equally gives too much weight to directions that have little effect on the network output, reaching a SRCC of only 0.784 on all full TID2013 dataset. Late representations capture information that is more relevant to the encoder output, but much of the spatial detail has already been pooled away; the output tokens reach a correlation of only 0.742. The key observation is that these two properties can be combined. We retain the spatial detail of early features, but weight their directions according to how strongly they affect the network output. This raises the correlation to 0.850 and, unlike the unweighted alternatives, remains stable as image resolution increases. The same effect is visible in Figure 1, raw features rank the five distorted images almost at random relative to human observers ($\tau = 0.2$), whereas the weighted features reproduce the human ordering exactly.

The weights come from the Jacobian of the output with respect to an early patch token. This Jacobian describes how a small change in the token affects the encoder output. Averaging its squared response over natural images gives a single fixed metric tensor in feature space, $\mathbb{E}\left[J^\top J\right]$. For block 1 of DINOv2-S, this metric is strongly anisotropic, only 64 of its 384 eigendirections account for 76.5% of its trace. Based on this observation, we introduce the Jacobian Lens Distance (JLD). JLD retains the 64 leading eigendirections of $\mathbf{M}$ as a fixed projection, which we call the Jacobian lens. It never uses human judgements (Appendix Figure A1). Because the lens is a fixed linear map, the geometry induced by JLD can be computed and visualised directly on real images, independently of any benchmark. To our knowledge, JLD is the first perceptual distance to transfer a network's output sensitivity back to its early features through a single fixed metric. Contributions:

- We introduce a novel perceptual distance derived entirely from the output sensitivity of a frozen vision encoder, without human labels. Its lens term defines a pseudometric whose local form is a pullback metric in pixel space (Section 3).

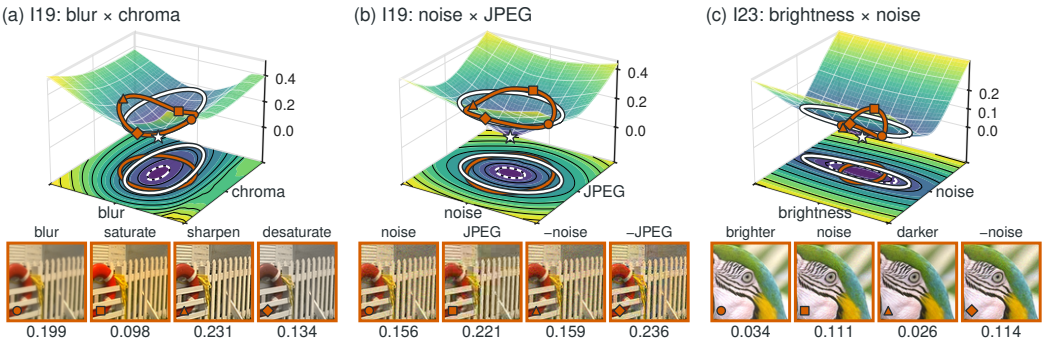

Figure 2: JLD on three planes spanned by two distortion directions of equal pixel energy around a reference (star), with the ellipse predicted by the pullback G (dashed). White ring: constant JLD; orange ring: constant MSE (25.1, 24.0 and 29.7 dB), whose four marked images are shown below with their JLD.

- We analyse the induced geometry directly on real images. The iso-distance contours of JLD agree with the ellipses predicted by the encoder, its sensitivity decreases with local contrast (Section 4).
- We evaluate JLD against 15 full-reference distances. JLD achieves the SOTA performance, and beat every learned deep distance, and remains stable as image resolution doubles. JLD-fast is four times faster than LPIPS-VGG, and a single fitted lens transfers across 14 encoders and extends to video (Section 5).

2 Background

A perceptual distance should assign small distances to images that look the same. A feature distance $\|\Phi(x)-\Phi(y)\|_{\mathbf{Q}}$, for a feature map Φ and positive semidefinite matrix $\mathbf{Q}$, satisfies the metric axioms by construction, while its local geometry around an image is determined by the pullback $A^\top \mathbf{Q} A$, where $A = \partial\Phi/\partial x$. Existing feature distances differ mainly in how $\mathbf{Q}$ is chosen, raw feature distance, DeepWSD and DeepDC weight all channels equally (Liao et al., 2022; Zhu et al., 2025), LPIPS learns a diagonal $\mathbf{Q}$ from human judgements (Zhang et al., 2018), DISTS fits weights on feature statistics (Ding et al., 2022), and DreamSim fine-tunes the encoder on human triplets (Fu et al., 2023). Classical metrics derive the tensor from models of early vision, either through the perceptual metric of a divisive-normalisation transform (Epifanio et al., 2003; Malo et al., 2006; Laparra et al., 2010; 2016) or through the Fisher information of a response model, with eigendistortions predicting the most and least visible directions (Berardino et al., 2017; Feather et al., 2025; Vacher & Mamassian, 2024); both produce an image-dependent matrix in pixel space. IEM integrates differences between denoising errors under a learned density (Ohayon et al., 2026), while LASI fits linear embeddings at inference time (Severo et al., 2024).

Pullback metrics through neural networks are established in other settings, active subspaces based on $\mathbb{E}[\nabla f \nabla f^\top]$ (Constantine et al., 2014), decoder-induced metrics on latent spaces (Arvanitidis et al., 2018), the pullback Fisher metric for steering language models (Wang et al., 2026), and the Jacobian lens of Gurnee et al. (2026), which decodes averaged intermediate-to-final Jacobians through the vocabulary. JLD differs from these approaches by deriving the tensor from a frozen encoder's own output, rather than from labels, a hand-designed vision model, or a learned density, and by fixing that tensor on early features so that it is shared across images; to our knowledge, no perceptual distance has been constructed this way.

3 The Jacobian lens distance

3.1 Encoder and features

JLD compares two aligned images using a single frozen encoder, DINOv2-S (ViT-S/14), pretrained without human judgements (Oquab et al., 2024). An RGB image $x \in [0,1]^{3\times H\times W}$ with sides divisible by $p = 14$ produces $N = HW/p^2$ patch positions. At position t, we take the patch token $\mathbf{h}_t(x) \in \mathbb{R}^d$, with $d = 384$, after block 1 of 12. The final layer-normalised class token $\mathbf{z}(x) \in \mathbb{R}^m$ is the output whose sensitivity defines the metric. Images are center cropped to a multiple of 14 pixels,

so the same lens applies at every image size. A lens can be fitted after any block; on the TID2013 pairs, it improves agreement with human scores at every depth, with the largest gain at block 1: 0.850, compared with 0.803 on the patch embedding, 0.816 after block 2, and at most 0.780 from block 3 onward (Appendix Figure A2). After one attention block, each patch has interacted once with its neighbours, so the tokens retain local structure while already incorporating context.

3.2 Sensitivity as a metric tensor

Raw feature distance weights all d directions equally, even though many have little effect on what the encoder computes. The Jacobian $J_t(x) = \partial \mathbf{z}(x)/\partial \mathbf{h}_t(x) \in \mathbb{R}^{m\times d}$ maps a small change δ at patch t to the first-order output change $J_t(x)\delta$. Averaging its squared response over images and positions gives a single fixed matrix,

$$\mathbf{M} = \mathbb{E}_x\Big[\frac{1}{N}\sum_{t=1}^{N} J_t(x)^\top J_t(x)\Big] = \sum_{i=1}^{d} \mu_i u_i u_i^\top, \qquad \mathbf{U}_k = [u_1 \cdots u_k], \tag{1}$$

where $\mu_1 \geq \cdots \geq \mu_d \geq 0$. The matrix $\mathbf{M}$ is positive semidefinite, and $\delta^\top \mathbf{M}\delta$ gives the expected squared output change caused by moving one patch by δ. We call the orthonormal matrix $\mathbf{U}_k$, with $k = 64$, the Jacobian lens. Among all k-dimensional subspaces, it retains the most sensitivity, $\mathrm{tr}(\mathbf{U}_k^\top \mathbf{M}\mathbf{U}_k)$, while any displacement orthogonal to it changes the output by at most $\sqrt{\mu_{k+1}}\,\|\delta\|$ in RMS (Proposition 1). The 64 retained directions account for 76.5% of the trace (Appendix A).

Fitting the lens requires no explicit Jacobian matrices (Appendix Figure A1; Algorithm 1). A Gaussian probe $v \sim \mathcal{N}(0, I_m)$ on the output, back-propagated once, gives $g_t = \nabla_{\mathbf{h}_t}(v^\top \mathbf{z}) = J_t^\top v$ at every patch, and $\mathbb{E}_v[g_t g_t^\top] = J_t^\top \mathbb{E}_v[vv^\top] J_t = J_t^\top J_t$; therefore, averaging $N^{-1}\sum_t g_t g_t^\top$ over crops and probes gives an unbiased estimate of $\mathbf{M}$ (Proposition 2). Using four 224×224 crops from each of 100 unlabelled DIV2K images (Agustsson & Timofte, 2017) and eight probes per crop, the fit takes 35 s on one A100 GPU and uses no image pairs, distortions, or human labels (Appendix A.2). Human data are used only to select among 245 configurations on CSIQ and KADID-10k development splits (Appendix H). The resulting lens displacement also tracks the encoder output far better than raw block-1 features (Appendix Figure A1).

Algorithm 1: JLD

Fit (unlabelled crops $\mathcal{X}$, P probes, rank k)
$\hat{\mathbf{M}} \leftarrow 0$
for $x \in \mathcal{X}$, $v \sim \mathcal{N}(0, I_m)$, P times **do**
  $g_t \leftarrow \nabla_{\mathbf{h}_t} v^\top \mathbf{z}(x)$ for all t
  $\hat{\mathbf{M}} \leftarrow \hat{\mathbf{M}} + \frac{1}{N|\mathcal{X}|P}\sum_t g_t g_t^\top$
end for
$\mathbf{U}_k \leftarrow$ top-k eigenvectors of $\hat{\mathbf{M}}$
Score (images x, y)
$\tilde{x}, \tilde{y} \leftarrow Rx, Ry$ ▷ Eq. 2
$\Delta_t \leftarrow \mathbf{U}_k^\top\big(\mathbf{h}_t(\tilde{x}) - \mathbf{h}_t(\tilde{y})\big)$ for all t
$c \leftarrow 1 - \cos\big(\mathbf{z}(\tilde{x}), \mathbf{z}(\tilde{y})\big)$
return $\big(\frac{1}{N}\sum_t \|\Delta_t\|^2\big)^{1/2} + \frac{1}{2}c$

3.3 The distance

Human contrast sensitivity falls faster with spatial frequency for colour than for luminance (Campbell & Robson, 1968; Mullen, 1985). Following S-CIELAB (Zhang & Wandell, 1996), we use a fixed pre-filter that blurs chroma while preserving luminance,

$$Rx = T^{-1}\big(Y,\ G_\sigma * C_b,\ G_\sigma * C_r\big), \qquad (Y, C_b, C_r) = T(x), \qquad \sigma = 2 \text{ pixels}, \tag{2}$$

where T maps RGB to YCbCr and G_σ is a Gaussian. Both filtered images are passed through the encoder on the same patch grid, and

$$D(x,y) = \underbrace{\Big[\frac{1}{N}\sum_{t=1}^{N}\big\|\mathbf{U}_k^\top\big(\mathbf{h}_t(Rx) - \mathbf{h}_t(Ry)\big)\big\|^2\Big]^{1/2}}_{\text{lens term } D_{\text{lens}}(x,y)} + \underbrace{\tfrac{1}{2}\big[1 - \cos\big(\mathbf{z}(Rx), \mathbf{z}(Ry)\big)\big]}_{\text{output term}}. \tag{3}$$

The per-patch terms produce the lens maps in Figure 1, while the output term captures whole-image changes such as global colour shifts. Development data fixed its weight at $1/2$ (Appendix H). The lens term is euclidean distance after a fixed feature map, $D_{\text{lens}}(x,y) = \|\Phi(x) - \Phi(y)\|_2$, where Φ stacks the lensed tokens $N^{-1/2}\mathbf{U}_k^\top \mathbf{h}_t(Rx)$. We call it pseudometric: it is non-negative, symmetric,

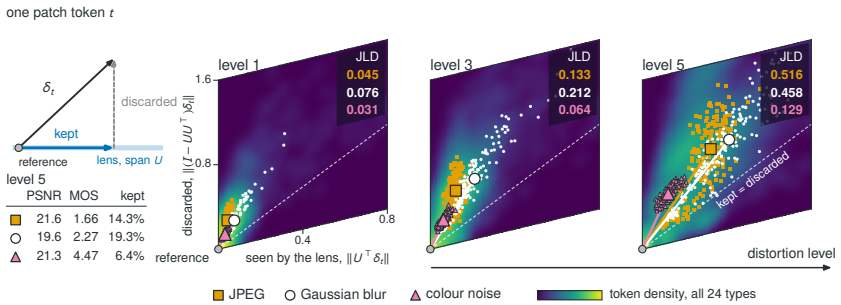

Figure 3: Three distortions of I19(TID2013) at levels 1, 3 and 5, per patch. Block-1 displacement the lens keeps (horizontal) against the part it discards (vertical). Markers sit at each family's RMS coordinates; their horizontal position is the lens term.

zero on identical images, and obeys the triangle inequality (Proposition 3). The output term need not satisfy the triangle inequality, and we empirically verify that violations by the full distance are rare (Section 4). JLD-fast keeps the pre-filter and lens term but stops the encoder after block 1, reducing the cost per pair from 118.5 to 10.7 GFLOPs; JLD-lin applies the lens to the linear patch embedding (Appendix A.4). Scoring requires two forward passes, and every operation is differentiable, so JLD can also be used as a loss function.

Near an image, the lens term becomes a quadratic form in pixel space: with $A_t(x) = \partial \mathbf{h}_t(Rx)/\partial x$,

$$D_{\text{lens}}(x, x+\varepsilon)^2 = \varepsilon^\top G(x)\,\varepsilon + o(\|\varepsilon\|^2), \qquad G(x) = \frac{1}{N}\sum_{t=1}^{N} A_t(x)^\top \mathbf{U}_k \mathbf{U}_k^\top A_t(x) \tag{4}$$

(Proposition 4). $G(x)$ is the pullback of the fixed lens through the encoder and therefore varies with the image. The output term contributes only $O(\|\varepsilon\|^2)$, so Equation 4 also gives the leading local behaviour of D.

4 Geometry of the Distance

In this section, we examine the geometry that the JLD induces on image space. Because JLD is an explicit function of a frozen encoder and a fixed projection, we can study its geometry directly on real images, before comparing it with human judgements. As guaranteed by Section 3.3, the lens term satisfies the triangle inequality on all 171.4 million same-reference triplets from four databases; the full distance violates it on at most 107 triplets per million, with a maximum violation of 3.6%. A single lens also generalises across images, a lens fitted to one image overlaps the shipped lens by a median 0.72, comparable to the 0.69 overlap between two fits of the same image using different methods (Appendix D.2).

4.1 The Local Shape of the Distance

Near an image x, Equation 4 gives $D(x, x+\varepsilon) \approx (\varepsilon^\top G(x)\,\varepsilon)^{1/2}$. With the eigen decomposition $G(x) = \sum_j \lambda_j e_j e_j^\top$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, perturbations at distance r form, to first order, $\{\varepsilon : \varepsilon^\top G(x)\varepsilon = r^2\}$, whose half-axis along e_j has length $r/\sqrt{\lambda_j}$. Directions with large λ_j are fixed, small pixel changes along them are registered strongly. Directions with small λ_j are soft and are tolerated more. For MSE, $G \propto I$, so the balls are spheres and all changes with equal energy are treated equally. Writing $\varepsilon^\top G(x)\varepsilon = \frac{1}{N}\sum_t \|\mathbf{U}_k^\top A_t(x)\varepsilon\|^2$ makes the two steps explicit, the lienarized first block $A_t(x)$ maps a pixel change to a token change that depends on local image content, and the fixed lens keeps only the component of that change that matters to the output. A pixel change can therefore be nearly invisible to JLD in two ways: it may barely move the tokens, or it may move them mostly along directions that the output ignores. Based on our observation, the second case is more common. Each term in $G(x)$ has rank at most k, so $\operatorname{rank} G(x) \leq kN = 64N$, while the image contains $3p^2N = 588N$ pixel values; therefore, at least 89% of pixel directions leave the lens term unchanged to first order at every image. Thus, JLD concentrates its sensitivity on an encoder-selected subspace that changes with the image.

Figure 2 restricts the distance to planes $\varepsilon = s\,a + t\,b$ spanned by two distortion directions a, b with equal pixel energy. On such a plane, G reduces to the 2×2 matrix $[a\ b]^\top G(x)[a\ b]$, whose contours

are ellipses with half-axis ratio $(\lambda_{\max}/\lambda_{\min})^{1/2}$, which measures their anisotropy. Around image I19 (TID2013) (Figure 2a), MSE produces nearly circular contours (radius ratio 1.06), whereas JLD produces ellipses 2.33 times longer along chroma than blur; LPIPS-VGG and DISTS give ratios of 1.61 and 1.64. The encoder predicts this shape well: the restricted pullback has anisotropy 2.07, its fixed axis lies 3° from blur, and its ellipse matches the innermost contour within 9.2% in radius. Blur is therefore a fixed direction, while an equal-energy chroma change is soft, consistent with the lower spatial acuity of human vision for colour (Mullen, 1985). The same prediction holds on the other planes, JPEG is stiff relative to noise (anisotropy 1.62, within 10.8%), while on I23 (TID2013) brightness is soft relative to noise (3.14). Across the three rings of equal MSE, JLD varies by factors of 2.4, 1.5, and 4.4, showing that equal pixel error can hide large differences in perceptual severity. When evaluated in small cells across a photograph, G becomes smaller in textured regions, where people also detect faint patterns less reliably (masking; SRCC -0.86 against local contrast, Appendix Figure A7).

4.2 Distortions through the lens

For a finite distortion, each block-1 token moves by $\Delta_t = \mathbf{h}_t(Ry) - \mathbf{h}_t(Rx)$, which splits orthogonally into a kept part $\mathbf{U}_k\mathbf{U}_k^\top\Delta_t$ and a discarded part $(I - \mathbf{U}_k\mathbf{U}_k^\top)\Delta_t$, with $\|\Delta_t\|^2 = \|\mathbf{U}_k^\top\Delta_t\|^2 + \|(I-\mathbf{U}_k\mathbf{U}_k^\top)\Delta_t\|^2$. The lens term measures the RMS of the kept component, whereas raw feature distance weighs both components equally. A displacement with no preferred direction would retain $k/d = 64/384$, or 16.7%, of its squared norm; a smaller fraction means that the distortion moves the tokens mainly along directions ignored by the output. Figure 3 shows the kept and discarded components for every patch. At level 5 on I19 (TID2013), JPEG and colour noise have nearly identical PSNR (21.6 and 21.3 dB), but very different MOS values, 1.66 and 4.47. The lens retains 14.3% of the JPEG displacement but only 6.4% of the noise displacement, making JLD four times larger for JPEG (0.516 versus 0.129), in agreement with human judgements. Colour noise moves the tokens substantially, but mostly along discarded directions; JPEG block edges move them along directions used by the output. Across all 120 distortions of I19 (TID2013), the lens retains 9.6% of the displacement energy; across all 24 distortion types, the median is 10.6% (Appendix D), well below the isotropic share. The lens therefore removes most block-1 feature motion; raw block-1 features count this discarded motion in full and rank the five images in Figure 1 almost at random.

5 Predicting human judgements

We compare JLD with 15 full-reference distances on human judgements. Because no human label enters the lens fit, these tests evaluate the geometry developed in Sections 3 and 4.

5.1 Image similarity benchmarks

We report Spearman rank correlation (SRCC) between each distance and human scores on TID2013, CSIQ, LIVE, KADID-10k and PIPAL (Ponomarenko et al., 2015; Larson & Chandler, 2010; Sheikh et al., 2006; Lin et al., 2019; Gu et al., 2020). LIVE, KADID-10k test references and PIPAL validation are held out; CSIQ and the KADID development references select the settings, while TID2013 informs the pre-filter. All baselines use their released preprocessing (Appendix H). Table 1 averages the first four databases. All costs are on A100 using 512×384 inputs, except IEM at 256×256. Using the published IEM and DeepDC settings leaves the mean ranking unchanged (Appendix H).

JLD achieves the highest mean SRCC over TID2013, CSIQ, LIVE and KADID-10k test among all distances in Table 1, 0.926, versus 0.915 for VSI, 0.913 for DeepWSD, 0.905 for DeepDC and 0.900

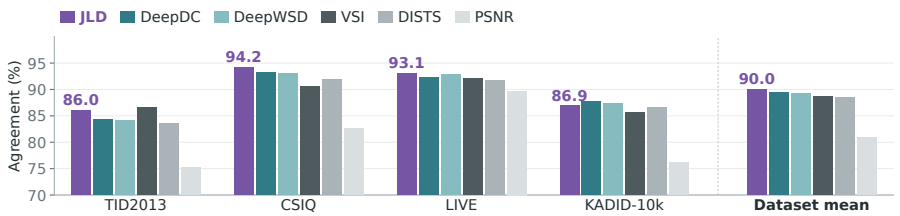

Figure 4: Agreement with 699,534 human same-reference choices, per dataset and as the mean over the four datasets.

Table 1: Image distances compared on agreement with human ratings and on cost per pair. JLD uses no human label, has the best mean over the datasets of every distance. **Bold** marks the best distance in a column; underlining marks the second. † marks IEM results from subsets of 2,032 KADID and 250 PIPAL pairs.

	Cost per pair ↓		Spearman correlation with human ratings ↑					
Method	GFLOPs	ms	TID2013 (diag.)	CSIQ (dev.)	LIVE test	KADID-10k test	**Mean** first four	PIPAL val
Learned deep distances, fitted to human judgements (similarity choices or quality scores)								
LPIPS-VGG	240.6	10.9	0.670	0.883	0.932	0.724	0.802	0.612
DISTS	240.7	9.9	0.818	0.943	0.954	0.885	0.900	0.704
PieAPP	385.7	16.5	0.844	0.897	0.918	0.864	0.881	0.706
DreamSim	212.2	51.3	0.812	0.911	0.910	0.851	0.871	**0.759**
Distances not fitted to human judgements								
PSNR	<**0.1**	**1.4**	0.687	0.809	0.873	0.673	0.760	0.255
SSIM	0.2	2.7	0.627	0.837	0.910	0.621	0.749	0.363
MS-SSIM	0.3	5.1	0.786	0.913	0.951	0.826	0.869	0.491
FSIM	<**0.1**	22.8	0.851	0.931	**0.965**	0.853	0.900	0.468
VSI	<**0.1**	9.9	**0.895**	0.940	0.949	0.876	0.915	0.450
GMSD	<**0.1**	2.6	0.804	0.957	0.960	0.849	0.893	0.585
NLPD	<**0.1**	15.6	0.799	0.937	0.937	0.811	0.871	0.370
DeepWSD	60.2	13.5	0.855	0.959	0.955	0.882	0.913	0.460
DeepDC	326.4	16.1	0.816	0.954	0.950	**0.900**	0.905	0.751
LASI	0.3	377.6	0.522	0.840	0.883	0.723	0.742	0.581
IEM†	4,999.1	725.7	0.810	0.917	0.922	0.821†	0.868†	0.182†
Ours: Label Free								
JLD-fast	10.7	2.5	0.875	0.956	0.945	0.870	0.911	0.578
JLD	118.5	12.2	0.877	**0.971**	0.964	0.892	**0.926**	0.624

for DISTS. Every gap has a 95% interval over resampled references that excludes zero (Appendix G). It exceeds every learned deep distance on all four datasets. On PIPAL's restoration and GAN outputs, where DreamSim (0.759) and DeepDC (0.751) lead, JLD reaches 0.624, above LPIPS-VGG and every classical distance. On BAPPS CNN distortions, it agrees with 0.821 of human choices, close to LPIPS-Alex (0.824) and DreamSim (0.826), note that both are trained on these data (Appendix F); it also beats DISTS on all distortion types (Appendix G).

Correlations summarise entire databases, so we also evaluate individual pairwise decisions. Across all 699,534 same-reference pairs with different human scores, JLD places the preferred image closer to the reference in 90.0% of cases, the highest among the 17 distances, ahead of DeepDC (89.4%) and DISTS (88.4%) (Figure 4); it ranks first on CSIQ, second on TID2013 and LIVE, and third on KADID-10k. Among the 24,623 pairs whose PSNR differs by less than 0.25 dB, where PSNR is at chance (48.7%), JLD agrees with people on 76.0%, compared with 74.4% for DISTS and 65.3% for LPIPS-VGG (mean over the four datasets). Figure A11 (Appendix G) shows four such pairs where JLD agrees with people while both DISTS and LPIPS-VGG do not. In a ten-person study on trials where the methods disagree, participants side with JLD on 24 of 28 decided trials and with DISTS on 14 of 26 (Appendix J).

5.2 Seeing the distance

A useful perceptual distance should assign similar values to distortions of similar perceptual severity (Figure 5, held-out KADID-10k test images). For one reference, we choose from each of the 25 distortion types the level closest to a target JLD value of 0.233 and retain the 9 types within 7%; their PSNR ranges from 20.3 to 28.1 dB, yet their human scores remain close, with a standard deviation of 0.46 MOS points. Applying the same rule around a target PSNR of 21.5 dB (within 0.5 dB) retains 10 types whose JLD values vary ninefold, from 0.071 to 0.663, and whose human scores spread almost three times as much, with a standard deviation of 1.26. A brightening and a lens blur with equal PSNR receive human scores of 4.4 and 1.2, and JLD values of 0.07 and 0.66. The ruler in panel (b) starts at the 0.35 JLD quantile of another reference, making every step visible: ordered by JLD, the human score decreases at every step from 3.8 to 1.1, whereas PSNR is not even monotone, with a 14.6 dB saturation shift in the middle; across all 121 distortions of that

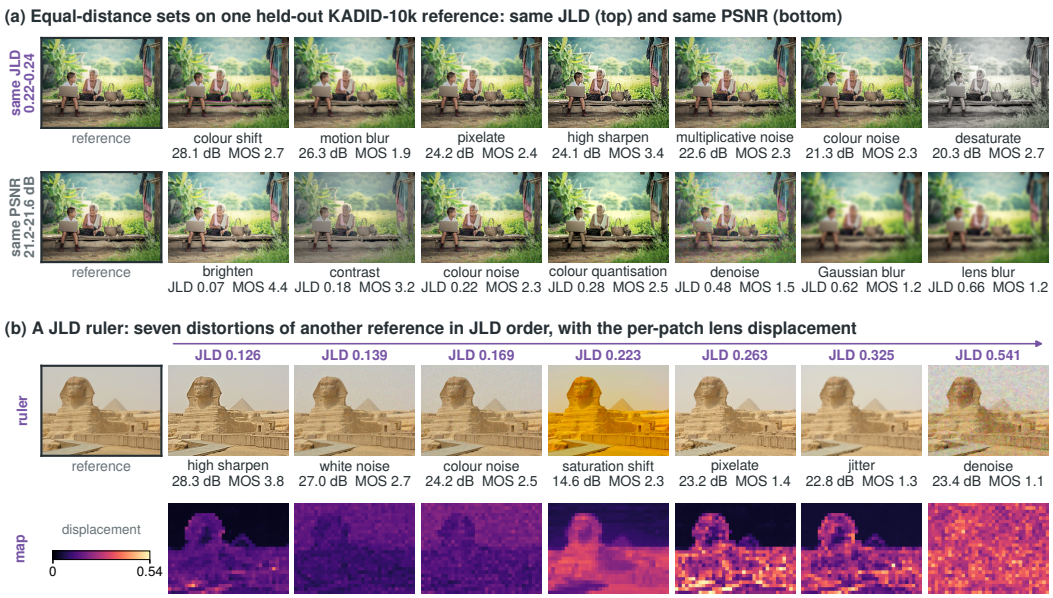

Figure 5: Equal JLD means equal damage. (a) Held-out KADID-10k distortions of one reference at equal JLD (top) and at equal PSNR (bottom), with human scores (MOS, 1 to 5). (b) Distortions of another reference in JLD order, with the per-patch JLD map of each. Selection rules in Section 5.2.

reference, JLD reaches 0.921 SRCC and PSNR 0.566. Each map gives an exact spatial account of its score because its root mean square equals the lens term: the saturation shift is distributed across the whole sphinx, whereas pixelation and jitter concentrate on its edges. Across complete datasets, neighbours in a distance's ordering of a reference's distortions should have similar human scores; on KADID-10k test, neighbouring distortions differ by 0.511 MOS points under JLD, 0.514 under DISTS, 0.797 under PSNR and 0.806 under LPIPS-VGG, and on TID2013 by 0.547 under JLD, 0.642 under DISTS and 0.972 under PSNR (MOS range 0 to 9).

5.3 Ablation

We ablate the components of JLD in Figure 6; the main gain comes from which directions the lens selects, not simply from how many are kept (Table A5). Using all 384 raw block-1 features gives a four-set mean of 0.772, and 64 random directions give the same value; 64 PCA directions reach 0.851, while the 64 lens directions reach 0.905 and perform better on every dataset. The chroma pre-filter and output term raise the mean further, to 0.911 and 0.926. PCA peaks at 16 directions and then declines, while the lens rises to 0.913 at $k = 64$ (Figure A12b); treating the selected directions equally also outperforms weighting them by the metric tensor, 0.916 versus 0.897 (Table A1). The construction is not specific to DINOv2: fitting a lens to each of 14 frozen encoders, including ViTs and CNNs from DINOv3 (Siméoni et al., 2025) to CLIP (Radford et al., 2021) and VGG16 (Simonyan & Zisserman, 2015), improves over raw features from the same block for all 14 (Figure A12d; Appendix K). It also converts the frozen no-reference TOPIQ and DOVER encoders into reference distances (Chen et al., 2024; Wu et al., 2023) (Appendix F).

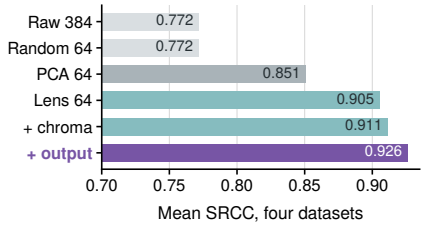

Figure 6: Building JLD step by step on DINOv2-S block 1.

5.4 Resolution, cost and video

To test stability across display resolutions, we rescore the labelled TID2013 and CSIQ pairs with 15 methods at short-side resolutions of 256, 384 and 512 pixels and at native size, using the lens term (Figure 7a,b; Table A6). From 256 to 512 pixels, the lens remains stable on TID2013 at 0.850, 0.860 and 0.845, while the unweighted block-1 features from which it is built fall from 0.784 to

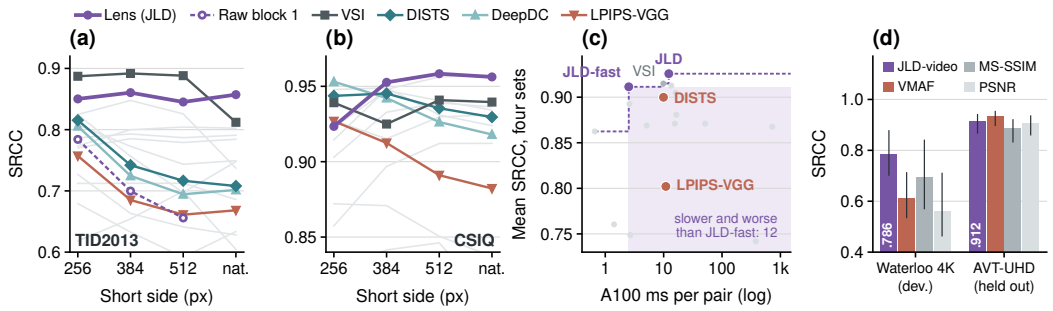

Figure 7: Resolution, cost and video. (a, b) SRCC at each short side and at native size; grey lines are the other ten methods, dashed is unweighted block-1 features. (c) Mean SRCC over the first four datasets of Table 1 against A100 time per pair, with the accuracy-cost frontier (dashed). (d) Video SRCC with 95% bootstrap intervals.

0.656, DISTS from 0.815 to 0.717 and LPIPS-VGG from 0.757 to 0.661. On CSIQ, the lens rises from 0.923 to 0.958, the highest of the 15 methods at 512 pixels. Its lowest TID2013 value, 0.845, exceeds the lowest value of every other method. JLD is also efficient to evaluate (Figure 7c), JLD-fast, which uses the lens term with the pre-filter, retains a mean of 0.911 at 2.5 ms per pair on an A100, 4.3 times faster than LPIPS-VGG (10.9 ms) and faster than DISTS (9.9 ms), while being both faster and more accurate than almost all metrics (Table **??**). The same construction extends to video by changing only the encoder, JLD-video uses a 16-direction lens on block 2 of a frozen VideoMAE-Base (Tong et al., 2022), fitted on 50 unlabelled clips; its space-time tokens capture temporal effects such as flicker that no single frame reveals (Appendix L). On the 240 Waterloo IVC 4K pairs (Li et al., 2019), which are used to choose the block and rank, it reaches 0.786 SRCC, compared with 0.694 for MS-SSIM, 0.611 for VMAF (Li et al., 2016) and 0.562 for PSNR, giving a 0.175 lead over VMAF (Figure 7d). On 120 held-out AVT-VQDB-UHD-1 pairs (Rao et al., 2019), it reaches 0.912 near SOTA VMAF's (0.932). Figure 8 revisits a held-out image pair with nearly equal PSNR (25.18 and 25.16 dB): viewers prefer A (MOS 3.67) to B (1.47); JLD agrees, whereas DISTS and LPIPS-VGG reverse the ordering (Appendix M).



Figure 8: Held-out KADID-10k pair. White boxes locate the enlarged boat detail.

6 Discussion and Conclusion

A frozen encoder's output Jacobian tells us how its early features should be weighted. JLD turns this sensitivity into one fixed metric tensor, fitted in 35 seconds from 100 unlabelled images. Among the distances evaluated here, it achieves the highest mean agreement with human judgements across all image databases, while remaining stable as image resolution doubles. The central principle is simple, measure with early features and weight them using later representations. This combination gives both accuracy and resolution stability (Appendix Figure A2). The lens improves over raw features for all 14 tested encoders, while its fixed geometry makes individual predictions directly inspectable through maps and contours. JLD-fast requires only 2.5 ms per pair. The seconds-long fit can also transfer the construction to a new encoder or domain without human ratings. More broadly, any differentiable encoder, including those for audio, medical imaging or 3D, can in principle define such a lens; closing its null space with the full metric tensor could further strengthen its use as a training loss (Appendix N). Perceptual weights need not come from human judgements, they can emerge from the geometry of a network that has simply learned to see.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

During the preparation of this work, the authors used ChatGPT (OpenAI) to assist with text editing and formatting the paper. The ChatGPT was also used as a coding assistant to implement some of the result analysis and figure scripts in Python. All suggestions, text and code produced with its help were reviewed, edited and verified by the authors, who take full responsibility for the content of the paper. The ideas, method, experimental design and results are the original work of the authors.

REPRODUCIBILITY STATEMENT

We provide the full anonymised code repository for all variants of JLD, together with the pre-fitted DINOv2-S lens, so the distance can be used without refitting. Section 3 and Algorithm 1 define the fit, the configuration and the distance, and Appendix C gives a runnable example. Appendices H, A.2, L and J give the baseline, fitting, video and human-study protocols. All image and video benchmarks are publicly available, and the full code base, including the lens fitting, evaluation and figure scripts, will be released publicly upon publication.

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1122–1131, 2017. doi: 10.1109/CVPRW.2017.150.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Alexander Berardino, Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3530–3539, 2017. URL https://arxiv.org/abs/1710.02266.

Fergus W. Campbell and John G. Robson. Application of Fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551–566, 1968. doi: 10.1113/jphysiol.1968.sp008574.

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. doi: 10.1109/TIP.2024.3378466. URL https://arxiv.org/abs/2308.03060.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active subspace methods in theory and practice: Applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014. doi: 10.1137/130916138. URL https://epubs.siam.org/doi/10.1137/130916138.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129(4):1258–1281, 2021. doi: 10.1007/s11263-020-01419-7. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7817470/.

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, 2022. doi: 10.1109/TPAMI.2020.3045810.

Irene Epifanio, Juan Gutiérrez, and Jesús Malo. Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding. *Pattern Recognition*, 36(8):1799–1811, 2003.

Jenelle Feather, David Lipshutz, Sarah E. Harvey, Alex H. Williams, and Eero P. Simoncelli. Discriminating image representations with principal distortions. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.15433.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023. doi: 10.52202/075280-2208.

Jinjin Gu, Haoming Cai, Haoyu Chen, Xiaoxing Ye, Jimmy S. Ren, and Chao Dong. PIPAL: A large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision*, volume 12356 of *Lecture Notes in Computer Science*, pp. 633–651. Springer, 2020. doi: 10.1007/978-3-030-58621-8_37.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, July 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Valero Laparra, Jordi Muñoz-Marí, and Jesús Malo. Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A*, 27(4):852–864, 2010.

Valero Laparra, Johannes Ballé, Alexander Berardino, and Eero P. Simoncelli. Perceptual image quality assessment using a normalized Laplacian pyramid. *Electronic Imaging*, 28(16):1–6, 2016. doi: 10.2352/ISSN.2470-1173.2016.16.HVEI-103.

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, 2010. doi: 10.1117/1.3267105.

Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. Netflix Technology Blog, June 2016. URL https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

Zhuoran Li, Zhengfang Duanmu, Wentao Liu, and Zhou Wang. AVC, HEVC, VP9, AVS2 or AV1? a comparative study of state-of-the-art video encoders on 4K videos. In *International Conference on Image Analysis and Recognition*, volume 11662 of *Lecture Notes in Computer Science*, pp. 162–173. Springer, 2019. doi: 10.1007/978-3-030-27202-9_14.

Xingran Liao, Baoliang Chen, Hanwei Zhu, Shiqi Wang, Mingliang Zhou, and Sam Kwong. DeepWSD: Projecting degradations in perceptual space to Wasserstein distance in deep feature space. In *ACM International Conference on Multimedia*, pp. 970–978, 2022. doi: 10.1145/3503161.3548193.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. KADID-10k: A large-scale artificially distorted IQA database. In *Eleventh International Conference on Quality of Multimedia Experience*, pp. 1–3, 2019. doi: 10.1109/QoMEX.2019.8743252.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.

Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group MAD competition - a new methodology to compare objective image quality models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1664–1673, 2016. doi: 10.1109/CVPR.2016.184.

Jesús Malo, Irene Epifanio, Rafael Navarro, and Eero P. Simoncelli. Nonlinear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, 15(1):68–80, 2006.

Kathy T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology*, 359(1):381–400, 1985. doi: 10.1113/jphysiol.1985.sp015591.

Guy Ohayon, Pierre-Etienne H. Fiquet, Florentin Guth, Jona Ballé, and Eero P. Simoncelli. Learning a distance measure from the information-estimation geometry of data. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4uTZobABec.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015. doi: 10.1016/j.image.2014.10.009.

Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. PieAPP: Perceptual image-error assessment through pairwise preference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817, 2018. doi: 10.1109/CVPR.2018.00194.

R. F. Quick, Jr. A vector-magnitude model of contrast detection. *Kybernetik*, 16(2):65–67, 1974. doi: 10.1007/BF00271628.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rakesh Rao Ramachandra Rao, Steve Göring, Werner Robitza, Bernhard Feiten, and Alexander Raake. AVT-VQDB-UHD-1: A large scale video quality database for UHD-1. In *IEEE International Symposium on Multimedia*, pp. 1–8, 2019. URL https://github.com/Telecommunication-Telemedia-Assessment/AVT-VQDB-UHD-1.

Daniel Severo, Lucas Theis, and Johannes Ballé. The unreasonable effectiveness of linear prediction as a perceptual metric. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=e4FG5PJ9uC.

Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. doi: 10.1109/TIP.2006.881959.

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. URL https://arxiv.org/abs/2508.10104.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1409.1556.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Jonathan Vacher and Pascal Mamassian. Perceptual scales predicted by Fisher information metrics. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=z7K2faBrDG.

Sihan Wang, Jiayi Zhao, Qingyan Cao, Hongbo Yao, and Lin Shu. FishBack: Pullback Fisher geometry for optimal activation steering in transformers. arXiv:2605.17231, 2026.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

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. doi: 10.1109/TIP.2003.819861.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wu_Exploring_Video_Quality_Assessment_on_User_Generated_Contents_from_Aesthetic_ICCV_2023_paper.html.

Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014. doi: 10.1109/TIP.2013.2293423.

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. doi: 10.1109/TIP.2011.2109730.

Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014. doi: 10.1109/TIP.2014.2346028.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Xuemei Zhang and Brian A. Wandell. A spatial extension to CIELAB for digital color image reproduction. In *Society for Information Display Symposium Digest of Technical Papers*, volume 27, pp. 731–734, 1996. URL https://wandell.vista.su.domains/blog/publications/.

Hanwei Zhu, Baoliang Chen, Lingyu Zhu, Shiqi Wang, and Weisi Lin. DeepDC: Deep distance correlation as a perceptual image quality evaluator. *IEEE Transactions on Image Processing*, 34: 7859–7873, 2025. doi: 10.1109/TIP.2025.3635025.

APPENDIX

A Method details . . . 14
B Proofs . . . 18
C Usage and reproducibility . . . 22
D Geometry of the distance . . . 24
E Reading the distance on real images . . . 25
F Full image results . . . 27
G Additional results . . . 29
H Development choices . . . 30
I What the lens sees in gratings . . . 31
J Validation and human study . . . 32
K Encoders . . . 33
L Video . . . 33
M Matched pixel error . . . 34
N Limitations . . . 34

A METHOD DETAILS

This section gives the implementation details of JLD. The encoder and features, lens calibration, rank selection, image-space corrections, variants, fit stability, and computational cost.

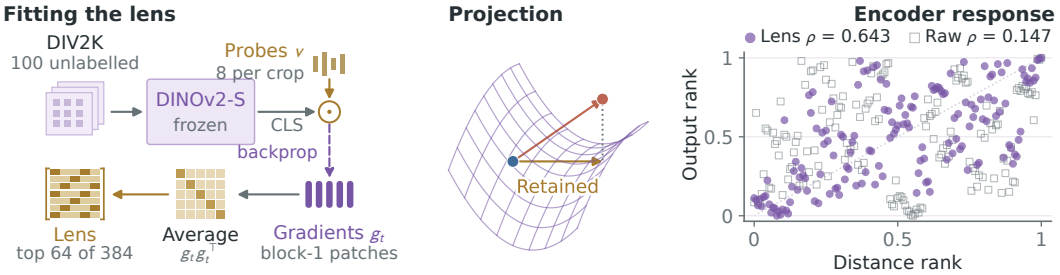

Figure A1: The Jacobian lens. Left: output probes are backpropagated to the block-1 patch tokens of 100 unlabelled images; the top 64 eigenvectors of the averaged gradient outer products form the lens. Middle: the lens keeps the part of a feature change that moves the output. Right: over 153 perturbations at 40 dB PSNR, the lens displacement ranks the encoder output change far better than the raw block-1 displacement.

A.1 ENCODER AND PATCH FEATURES

JLD compares two aligned images using one frozen encoder and one fixed linear projection. We use DINOv2-S (ViT-S/14), pretrained without quality judgements (Oquab et al., 2024). For an RGB image whose sides are multiples of the patch size $p = 14$, the encoder produces $N = HW/196$ patch positions. At each position t, we read the patch vector $\mathbf{h}_t \in \mathbb{R}^{384}$ from the residual stream after block 1; block 0 denotes the linear patch embedding. The final layer-normalised class token $\mathbf{z} \in \mathbb{R}^{384}$ defines the sensitivity used to fit the lens and supplies the whole-image output term. The distance has no fixed input size. Calibration uses 224×224 crops, corresponding to a 16×16 patch grid. At evaluation, images are centre-cropped to the largest multiple of 14 pixels in each dimension. For example, a 512×384 image becomes 504×378, or 36×27 patches. DINOv2 interpolates its positional embeddings to each grid, allowing the same fitted lens to operate across image sizes (Section 5.4).

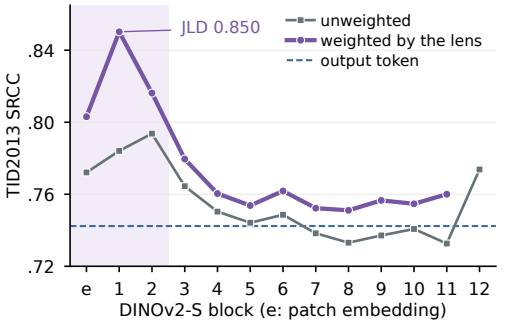

Figure A2: TID2013 SRCC of patch-token distances at each depth of DINOv2-S, raw or through the lens fitted there.

Figure A2 repeats the fit at every encoder depth. The lens-weighted distance peaks at block 1 with correlation 0.850, compared with 0.803 at the patch embedding, 0.816 after block 2, and at most 0.780 from block 3 onward. The same choice also gives the strongest resolution stability, from a

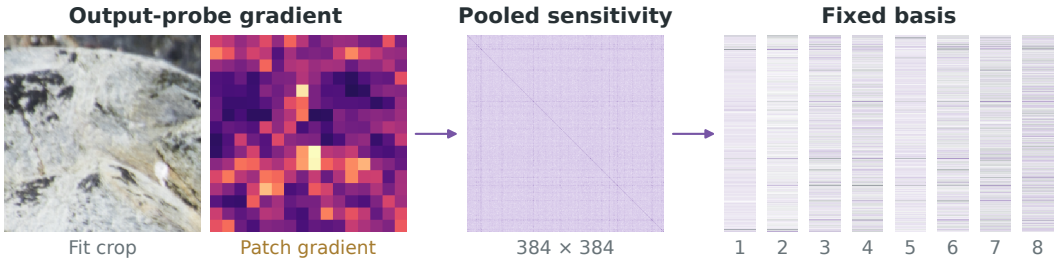

Figure A3: One calibration step. Probe gradients at every patch (left) are pooled into the 384×384 sensitivity matrix (middle), whose leading eigenvectors form the lens (right, first eight of 64).

256 to 512 pixel short side, the block-1 lens changes only from 0.850 to 0.845, whereas unweighted block-1 features fall from 0.784 to 0.656 and final patch features from 0.762 to 0.722.

A.2 Calibrating the Lens

The lens should preserve feature directions that affect the encoder output rather than directions that merely vary across images. For patch t, the Jacobian $J_t(x) = \partial \mathbf{z}(x)/\partial \mathbf{h}_t(x)$ maps a small token displacement δ to the first-order output change $J_t(x)\delta$. Averaging its squared response over images and positions gives the sensitivity matrix of Equation 1,

$$\mathbf{M} = \mathbb{E}_{x,t}\big[J_t(x)^\top J_t(x)\big], \qquad \delta^\top \mathbf{M}\delta = \mathbb{E}_{x,t}\|J_t(x)\delta\|^2.$$

Squaring before averaging prevents responses of opposite sign from cancelling. We omit cross-patch products, so $\mathbf{M}$ describes the sensitivity of a single patch intervention. Explicitly forming J_t would require one backward pass per output coordinate. Instead, we draw a Gaussian probe $v \sim \mathcal{N}(0, I)$ at the output. One backward pass of $v^\top \mathbf{z}$ gives $g_t = J_t^\top v$ at every patch, with $\mathbb{E}_v[g_t g_t^\top] = J_t^\top J_t$. For n images, C crops, and P probes, we estimate

$$\widehat{\mathbf{M}} = \frac{1}{nCP} \sum_{i,c,p} \frac{1}{N} \sum_{t} g_{icpt} g_{icpt}^\top, \qquad g_{icpt} = J_t(x_{ic})^\top v_{icp}, \quad v_{icp} \sim \mathcal{N}(0, I), \tag{5}$$

which is unbiased, with error decreasing as $(nCP)^{-1/2}$ (Proposition 2). We make $\widehat{\mathbf{M}}$ symmetrical and retain its 64 leading eigenvectors as the orthonormal lens $\mathbf{U}_k$, without eigenvalue weighting or whitening (Algorithm 1). Calibration uses 100 unlabelled DIV2K validation images (Agustsson & Timofte, 2017), disjoint from all quality datasets used for evaluation. The calibration crops are not chroma-filtered, and only activation gradients are computed; the encoder remains frozen. The complete fit takes 34.53 s on one A100 GPU with 327.8 MB peak GPU memory. It uses no image pairs, distortions, or human labels. Once fitted, the lens is reused for every comparison with no gradients. Figure A3 visualizes one calibration crop, a single output produces gradients at all patches, whose outer products are accumulated into the 384×384 sensitivity matrix. Its leading eigenvectors mix many feature coordinates rather than selecting individual channels.

A.3 Feature Rank

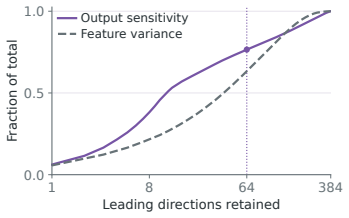

Figure A4: Cumulative share of $\mathrm{tr}\,\mathbf{M}$ (violet) and of feature variance (grey).

Figure A4 compares the cumulative spectrum of $\mathbf{M}$ with that of the feature covariance. Sensitivity is substantially more concentrated, the leading 64 of 384 directions contain 76.5% of $\mathrm{tr}\,\mathbf{M}$, so the lens discards rest of feature space while retaining over three quarters of the encoder's squared sensitivity. The sensitivity and covariance spectra also select different subspaces. For a diagnostic using the class token and mean patch token as output, the block-1 lens retains 78.7% of sensitivity but only 33.7% of feature variance. Its overlap with the 64 leading PCA directions is 0.358 at block 1 and 0.060 at block 12, compared with 0.167 in expectation for a random 64-dimensional subspace.

Table A1: Lens form. SRCC with MOS for the lens term under each choice of $\mathbf{Q}$ in $\|\mathbf{Q}^{1/2}\delta\|$, together with the development mean (CSIQ, KADID dev.) after adding the output term.

$\mathbf{Q}$ in the lens term	CSIQ	KADID dev.	TID2013	KADID test	Dev. mean, with output
$\mathbf{U}_k\mathbf{U}_k^\top$ (flat, released)	**0.956**	**0.876**	**0.874**	**0.870**	**0.932**
$\mathbf{M}$ (Mahalanobis)	0.946	0.848	0.859	0.845	0.926
$\Pi_k\mathbf{M}\Pi_k$ ($\sqrt{\mu}$-weighted lens)	0.934	0.836	0.859	0.830	0.926
I (raw block-1 features)	0.836	0.710	0.670	0.718	0.859

Development correlation peaks at 0.913 for $k = 64$, falls to 0.782 when all 384 directions are retained, and reaches at most 0.883 for PCA subspaces at any rank (Figure A5). Thus, the gain comes primarily from which directions are retained rather than dimensionality reduction alone. The released lens is flat, all retained directions receive equal weight rather than being weighted by μ_i. Table A1 compares this choice with eigenvalue-weighted and full Mahalanobis alternatives. Although the flat and eigenvalue-weighted lens terms rank pairs similarly, the flat projection agrees better with human scores on all four datasets. The full Mahalanobis metric $\mathbf{M}$ also performs worse, while raw block-1 features are substantially weaker.

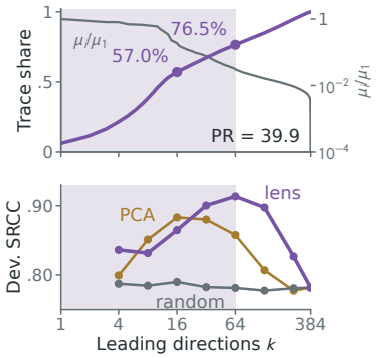

Figure A5: Spectrum of $\mathbf{M}$ (top; 64 kept directions carry 76.5% of the trace; PR: participation ratio) and development SRCC against rank k.

A.4 Lens Sensitivity

Output sensitivity is a substantially better ruler than raw feature energy. Across 153 perturbations of three reference images, RMS lens displacement has Spearman correlation 0.643 with the change of the final class token, compared with 0.147 for raw block-1 displacement. The lens therefore retains the component of an early feature change that propagates to the encoder output. The lens also has a direct pixel interpretation at block 0. There the patch features are a linear map W of a $3 \times 14 \times 14$ RGB patch, so each lens direction u acts as a matched filter $W^\top u$ on pixel changes; JLD-lin uses this construction directly. Biases and positional embeddings cancel when taking differences. Eigenvector signs are arbitrary, and nearly tied directions may rotate between fits, but sign changes and rotations within the retained subspace leave every distance unchanged.

A.5 Chroma pre-filter

Human sensitivity falls faster with spatial frequency for colour than for luminance (Campbell & Robson, 1968; Mullen, 1985). The unfiltered lens does not fully reproduce this behaviour, at the pixel Nyquist frequency, its red–green and blue–yellow responses retain 0.208 and 0.274 of their peaks, respectively, where human chromatic sensitivity is close to zero (Appendix I). Following S-CIELAB (Zhang & Wandell, 1996), we therefore apply a fixed pre-filter that blurs chroma while preserving luminance,

$$Rx = T^{-1}\big(Y,\ G_\sigma * C_b,\ G_\sigma * C_r\big), \qquad (Y, C_b, C_r) = T(x), \qquad \sigma = 2 \text{ pixels}, \tag{6}$$

where T maps RGB to YCbCr and G_σ is a Gaussian with reflect padding. The result is clipped to $[0, 1]$, and both filtered images are then evaluated on the same patch grid.

A.6 Pooling and Output term

Per-patch lens lengths are pooled quadratically. Under independent equal-variance Gaussian noise, an ideal observer combines per-patch d' in this form, corresponding to exponent-two probability summation (Quick, 1974). Quadratic pooling also makes the lens term a Euclidean distance after a fixed feature map and therefore a pseudometric (Proposition 3). The output term is one half of the cosine distance between the final layer-normalised class tokens. It complements the local lens term with a whole-image comparison, raising CSIQ correlation from 0.956 to 0.971. Its weight of $1/2$

Table A2: Fit stability. Subspace overlap with the 100-image lens and development mean SRCC (CSIQ, KADID dev.); Δ is relative to the 100-image fit.

Fit corpus	Overlap	SRCC	$\Delta \times 10^4$	Disjoint half	Overlap	SRCC	$\Delta \times 10^4$
10 images	0.877	0.913	-2.5	Half A, seed 0	0.963	0.913	-2.7
25 images	0.916	0.914	$+3.7$	Half A, seed 1	0.935	0.914	$+1.6$
50 images	0.963	0.913	-2.7	Half B, seed 0	0.943	0.914	$+4.0$
100 images	1.000	0.913	0.0	Half B, seed 1	0.951	0.913	-0.7

was selected on development data and is not sensitive, weights from 0.25 to 1 keep CSIQ between 0.968 and 0.971 and TID2013 between 0.872 and 0.877 (Appendix H).

A.7 Variants: JLD-fast, JLD-lin and JLD-video

JLD-fast retains the pre-filter and lens term, removes the output term, and stops the encoder after block 1. This reduces the cost per pair from 118.5 GFLOPs and 12.2 ms to 10.7 GFLOPs and 2.5 ms on an A100, while making the complete distance a pseudometric.

JLD-lin moves one step earlier and fits the lens on the block-0 patch embedding. The resulting distance is the RMS response of 64 fixed 14×14 RGB filters to the filtered difference image, and its behaviour under grid refinement can be analysed directly (Theorem 1).

JLD-video applies the same construction to a frozen VideoMAE-Base fine-tuned on Kinetics-400 (Tong et al., 2022). It uses a 16-dimensional lens at block 2, fitted to the class logits using 50 unlabelled clips and eight probes per clip. Squared projected differences between aligned space-time tokens are averaged over tokens and eight evenly spaced 16-frame clips before taking the square root. The video variant uses neither the chroma pre-filter nor an output term (Appendix L).

A.8 Fit stability

We test whether calibration depends strongly on the choice of unlabelled corpus. For two lenses $\mathbf{U}$ and $\widetilde{\mathbf{U}}$, we measure subspace overlap as $\|\mathbf{U}^\top \widetilde{\mathbf{U}}\|_F^2/k$, which equals 1 for identical subspaces and has expectation $k/384$ for random ones. Ten calibration images already give overlap 0.877 with the 100-image lens and change either development correlation by less than 0.002. Two disjoint 50-image halves, each fitted with two seeds, overlap the full fit by 0.935–0.963 and remain within 4×10^{-4} of its development correlation. Similarly, two probes per crop give overlap 0.967 and four give 0.990; using 16 probes instead of eight, or crop sizes of 112 or 448 instead of 224, changes either correlation by less than 0.002. The output target matters more than the calibration sample: targeting the next block rather than the final output gives overlap 0.510, CSIQ correlation 0.919, and KADID development correlation 0.842. These stability diagnostics used the preliminary target combining the class token and mean patch token; Appendix H reports the block, rank, and output-weight sweeps for the final configuration. To refit the lens for a new domain, collect roughly 100 representative unlabelled images that are disjoint from evaluation data, while keeping the encoder, block, output target, and normalisation fixed. Save the fitted basis together with the corpus manifest and random seed, and compare independent fits through their subspace overlap. Changing the encoder or output target requires a refit; selecting the rank or output weight from human ratings requires separate development and test data (Appendix C).

A.9 Implementation

Scoring one pair requires two forward passes, one 384×64 projection per patch, and one cosine computation, with no gradients. The implementation center crops each image to the patch grid, applies the chroma pre-filter, and returns the distance. The released lens is a 384×64 array stored together with its sensitivity matrix, feature covariance, and calibration metadata; inference requires only the lens and encoder. As shown in Figure 7c, JLD and JLD-fast lie on the measured speed–accuracy Pareto frontier. JLD-fast is 4.3 times faster than LPIPS-VGG on an A100 and 14 times faster on twelve CPU threads because it stops after the first of twelve encoder blocks. Full JLD evaluates the remaining blocks only to obtain the final class token. JLD timings exclude the pre-filter and output cosine; gives the complete timing protocol and all measured values.

B PROOFS

We assume throughout that the encoder is frozen and that its patch features $\mathbf{h}_t$ and output $\mathbf{z}$ are differentiable in the input pixels, twice where required below. Images lie in the compact set $[0,1]^{3HW}$ for a fixed patch grid, and patch positions are sampled uniformly within each image. These assumptions keep the relevant Jacobians bounded and the Taylor remainders uniform; DINOv2, built from GELU activations, layer normalisation and softmax attention, satisfies them. The pre-filter R is affine away from saturated pixels (Appendix A.5), and statements involving Rx are understood on this region. As in Section 3,

$$J_t(x) = \frac{\partial \mathbf{z}(x)}{\partial \mathbf{h}_t(x)}, \qquad \mathbf{M} = \mathbb{E}_{x,t}[J_t^\top J_t] = \sum_{j=1}^{d} \mu_j u_j u_j^\top,$$

with $\mu_1 \geq \cdots \geq \mu_d \geq 0$. The lens is $\mathbf{U}_k = [u_1 \cdots u_k]$ and $\Pi_k = \mathbf{U}_k \mathbf{U}_k^\top$.

B.1 THE SENSITIVITY MATRIX AND THE LENS

The quadratic form

$$\delta^\top \mathbf{M} \delta = \mathbb{E}_{x,t} \|J_t \delta\|^2$$

is the Euclidean metric of output space pulled back through the map $\mathbf{h}_t \mapsto \mathbf{z}$, with the other intermediate tokens fixed, and then averaged over images and positions. If the output is observed through isotropic Gaussian noise with variance σ^2, then $J_t^\top J_t/\sigma^2$ is the Fisher information for $\mathbf{h}_t$. Thus, to first order, $\sqrt{\delta^\top \mathbf{M} \delta}/\sigma$ is the RMS detectability d' of the displacement δ.

Proposition 1 (Optimal subspace)**.** *Among all rank-k orthonormal bases V, the lens $\mathbf{U}_k$ maximises* $\operatorname{tr}(V^\top \mathbf{M} V)$. *For any displacement δ chosen independently of the image and position,*

$$\mu_k \|\mathbf{U}_k^\top \delta\|^2 \leq \mathbb{E}_{x,t} \|J_t(x)\delta\|^2 = \delta^\top \mathbf{M} \delta \leq \mu_1 \|\mathbf{U}_k^\top \delta\|^2 + \mu_{k+1} \|(I - \Pi_k)\delta\|^2. \tag{7}$$

Proof. The trace statement follows from Ky Fan's maximum principle. Expanding in the eigenbasis gives

$$\delta^\top \mathbf{M} \delta = \sum_j \mu_j \langle u_j, \delta \rangle^2.$$

For $j \leq k$, $\mu_j \in [\mu_k, \mu_1]$, while for $j > k$, $\mu_j \in [0, \mu_{k+1}]$, which gives Equation 7. □

For the released lens, $\mu_{65}/\mu_1 = 0.030$. Hence an equal-norm displacement lying entirely outside the lens has RMS sensitivity at most $\sqrt{\mu_{65}/\mu_1} = 0.173$ of the leading eigendirection. The worst-case variation within the retained subspace is much larger, since $\mu_1/\mu_{64} = 32.8$, but real distortions occupy a much narrower range. Across all reference–distortion pairs in CSIQ, KADID-10k and TID2013, the pooled ratio

$$\gamma(\delta) = \frac{(\delta^\top \mathbf{M} \delta)^{1/2}}{\|\mathbf{U}_k^\top \delta\|}$$

has a 90th-to-10th percentile ratio of only 1.26–1.50, compared with the worst-case factor $\sqrt{32.8} = 5.7$. This helps explain why the flat lens and the $\mathbf{M}$-weighted form rank image pairs similarly (Table A1).

Lemma 1 (Captured sensitivity is stable)**.** *Let A and B be symmetric matrices with top-k eigenbases $\mathbf{U}_A$ and $\mathbf{U}_B$, and let $\|\cdot\|_{(k)}$ denote the sum of the k largest singular values. Then*

$$0 \leq \operatorname{tr}(\mathbf{U}_A^\top A \mathbf{U}_A) - \operatorname{tr}(\mathbf{U}_B^\top A \mathbf{U}_B) \leq 2\|A - B\|_{(k)}.$$

Proof. The lower bound follows from Ky Fan's principle for A. For the upper bound, write the gap as

$$\operatorname{tr}(\mathbf{U}_A^\top (A - B) \mathbf{U}_A) + \Big[\operatorname{tr}(\mathbf{U}_A^\top B \mathbf{U}_A) - \operatorname{tr}(\mathbf{U}_B^\top B \mathbf{U}_B) \Big] + \operatorname{tr}(\mathbf{U}_B^\top (B - A) \mathbf{U}_B).$$

The middle term is non-positive by Ky Fan's principle for B, and each outer term is at most $\|A - B\|_{(k)}$. □

Importantly, this bound requires no eigengap. The sensitivity captured by the estimated subspace can therefore remain stable even when individual eigenvectors are not.

Proposition 2 (Random probes)**.** *Let $x_1, \dots, x_n$ denote independent calibration inputs, and draw P probes $v \sim \mathcal{N}(0, I_m)$ independently for each input. Let $\widehat{\mathbf{M}}$ average*

$$\frac{1}{N} \sum_t (J_t^\top v)(J_t^\top v)^\top$$

over the nP input–probe pairs, and let $\mathbf{M}_n$ be the corresponding average of the exact per-input sensitivity tensors.

Then:

1. *$\widehat{\mathbf{M}}$ is unbiased.*
2. *For fixed inputs, its relative RMS probe error is at most*

$$\beta \sqrt{\frac{2r}{nP}}, \qquad r = \frac{(\operatorname{tr} \mathbf{M}_n)^2}{\|\mathbf{M}_n\|_F^2} \le d,$$

   where $\beta \ge 1$ measures variation in the per-input traces.
3. *Over random inputs,*

$$\mathbb{E}\|\widehat{\mathbf{M}} - \mathbf{M}\|_F^2 = \frac{V_{\text{content}}}{n} + \frac{V_{\text{probe}}}{nP},$$

   with both terms finite.
4. *If $\widehat{\mathbf{U}}$ is the top-k eigenspace of $\widehat{\mathbf{M}}$, then*

$$\operatorname{tr}(\widehat{\mathbf{U}}^\top \mathbf{M} \widehat{\mathbf{U}}) \ge \sum_{j \le k} \mu_j - 2\|\widehat{\mathbf{M}} - \mathbf{M}\|_{(k)}.$$

Proof. For (i),

$$\mathbb{E}_v[(J_t^\top v)(J_t^\top v)^\top] = J_t^\top \mathbb{E}[vv^\top] J_t = J_t^\top J_t.$$

Moreover, one backward pass of $v^\top \mathbf{z}$ returns $J_t^\top v$ simultaneously at every patch.

For (ii), independent probes make the centred errors uncorrelated. Isserlis's theorem expresses their second moment through products $J_t^\top J_s$, and Cauchy–Schwarz bounds the resulting variance by $2(\operatorname{tr} \mathbf{M}_i)^2$. Averaging over nP samples and normalising by $\|\mathbf{M}_n\|_F^2$ gives the stated bound.

Part (iii) follows from the law of total variance. Part (iv) follows directly from Lemma 1 with $A = \mathbf{M}$ and $B = \widehat{\mathbf{M}}$. □

For the released calibration, the 100 source images provide $n = 400$ random crops, with $P = 8$ probes per crop. Here $r = 39.9$ and $\beta^2 = 1.58$, giving a probe-error bound of approximately $0.20\|\mathbf{M}\|_F$; image sampling contributes approximately $0.14\|\mathbf{M}\|_F$. The measured variation is smaller: two independent refits with fresh crops and probes differ from the released tensor by 8.4% of $\|\mathbf{M}\|_F$, their lenses have subspace overlap 0.962, and each loses at most 0.35% of the sensitivity captured by the other's fitted lens.

B.2 The lens term as a distance

Proposition 3 (Early-feature distance)**.** *The lens term D_{lens} of Equation 3 is a pseudometric on images evaluated on a common patch grid.*

Proof. With the pre-filter, lens, and patch grid fixed,

$$D_{\text{lens}}(x, y) = \|\Phi(x) - \Phi(y)\|_2,$$

where

$$\Phi(x) = N^{-1/2} \operatorname{vec}\big(\mathbf{U}_k^\top \mathbf{h}_1(Rx), \dots, \mathbf{U}_k^\top \mathbf{h}_N(Rx)\big).$$

Euclidean distance after a fixed map is non-negative, symmetric, satisfies the triangle inequality, and vanishes whenever the mapped representations agree. Hence D_{lens} is a pseudometric. □

Zero distance means that every projected patch feature agrees, giving a pair of lens metamers. The result also applies to JLD-fast and JLD-lin. A crop chosen independently for each pair would not define a single map Φ across a triangle and would therefore not inherit this guarantee.

Proposition 4 (Local quadratic form)**.** *Let*

$$A_t(x) = \frac{\partial \mathbf{h}_t(Rx)}{\partial x}.$$

For a small pixel perturbation ε,

$$D_{\text{lens}}(x, x+\varepsilon)^2 = \varepsilon^\top G(x)\varepsilon + o(\|\varepsilon\|^2),$$

where

$$G(x) = \frac{1}{N}\sum_t A_t(x)^\top \mathbf{U}_k \mathbf{U}_k^\top A_t(x).$$

The matrix $G(x)$ is positive semidefinite and has rank at most kN. If $\mathbf{z}(Rx) \neq 0$, the output term is $O(\|\varepsilon\|^2)$.

Proof. Each patch feature has the expansion

$$\mathbf{h}_t(R(x+\varepsilon)) - \mathbf{h}_t(Rx) = A_t(x)\varepsilon + o(\|\varepsilon\|).$$

Projecting, squaring, and averaging over patches gives the stated quadratic form. Each summand has rank at most k, hence $\operatorname{rank} G(x) \leq kN$.

Let $\hat{z} = \mathbf{z}/\|\mathbf{z}\|$. The output term can be written as

$$\frac{1}{4}\|\hat{z}(R(x+\varepsilon)) - \hat{z}(Rx)\|^2,$$

which is the square of a first-order change and is therefore $O(\|\varepsilon\|^2)$. □

Thus, $G(x)$ is the pullback of the fixed feature-space form $\mathbf{U}_k\mathbf{U}_k^\top$ into pixel space. The lens is fixed globally, but its local pixel geometry depends on the image. The lens term is the chordal distance induced by this feature map, not a geodesic length.

Moreover,

$$\varepsilon^\top G(x)\varepsilon = \frac{1}{N}\sum_t \|\mathbf{U}_k^\top A_t(x)\varepsilon\|^2,$$

so the first-order metamers are

$$\ker G(x) = \{\varepsilon : \mathbf{U}_k^\top A_t(x)\varepsilon = 0 \ \forall t\}.$$

Their dimension is at least $3HW - kN$. Likewise, in token space, any $\delta \perp \operatorname{span} \mathbf{U}_k$ is invisible to the lens and satisfies

$$\delta^\top \mathbf{M}\delta \leq \mu_{k+1}\|\delta\|^2$$

by Equation 7.

B.3 Refining the pixel grid

For JLD-lin, define the per-patch lens energy

$$\rho_N(t) = \|\mathbf{U}_k^\top(\mathbf{h}_t(Rx) - \mathbf{h}_t(Ry))\|^2,$$

so that D_{lens}^2 is its average over the patch grid. At block 0, the patch token is the affine map

$$E\,\text{vec}(P_t x) + e_t,$$

and the offset cancels between two images. Hence

$$\rho_N(t) = \|W\,\text{vec}(P_t(Rx - Ry))\|^2, \qquad W = \mathbf{U}_k^\top E,$$

which is the squared response of k fixed 14×14 RGB filters.

Theorem 1 (Grid consistency of JLD-lin). *Let two images on the unit square have an L-Lipschitz difference $\delta = x - y$, and assume that the pre-filter does not saturate. Sample them at $n = 14g$ pixels per unit length, giving $N = g^2$ patches. Let $W_{\mathrm{dc}} \in \mathbb{R}^{k\times 3}$ contain the DC gains of the filters and define*

$$\rho(u) = \|W_{\mathrm{dc}}\delta(u)\|^2.$$

Then

$$\left| D^2_{\mathrm{lens},N} - \int \rho(u)\, du \right| \le \frac{c_1 L}{n} + \frac{c_2 L^2}{n^2} = O(N^{-1/2}), \tag{8}$$

where c_1 and c_2 depend only on $\|\delta\|_\infty$, $\|W\|_2$, $\|W_{\mathrm{dc}}\|_2$, and the colour transform. For band-limited images with angular frequency at most B, $L \le \sqrt{6}B$.

Proof. Every pixel contributing to patch t, including those reached through the chroma filter, lies within $r_p = 13\sqrt{2}$ pixels of the patch centre c_t. Lipschitz continuity therefore gives

$$\rho_N(t) = \|W_{\mathrm{dc}}\delta(c_t) + b_t\|^2, \qquad \|b_t\| \le \frac{14\kappa L r_p \|W\|_2}{n},$$

where $\kappa = \sqrt{3}\|T\|_2\|T^{-1}\|_2$. Consequently,

$$|\rho_N(t) - \rho(c_t)| \le 2\|W_{\mathrm{dc}}\|_2\|\delta\|_\infty\|b_t\| + \|b_t\|^2.$$

The function ρ is $2\|W_{\mathrm{dc}}\|_2^2\|\delta\|_\infty L$-Lipschitz, so replacing its value at each patch centre by the corresponding cell average contributes another $O(n^{-1})$ term. Summing the two errors yields Equation 8. The band-limited statement follows from Bernstein's inequality applied channel-wise. □

The limiting form depends only on the filters' DC gains, which contain just 0.28% of their total energy. Thus, the learned filters are predominantly band-pass edge and texture detectors. The experimentally relevant resolutions lie well before this asymptotic DC limit; the resolution stability measured in Section 5.4 is therefore a property of this finite-resolution band-pass regime. For block-1 features, one attention step intervenes between pixels and tokens, so resolution stability is measured empirically rather than covered by Theorem 1.

B.4 One lens for every image

For one image, define

$$\mathbf{M}_x = \frac{1}{N}\sum_t J_t(x)^\top J_t(x), \qquad \mathbf{M} = \mathbb{E}_x \mathbf{M}_x,$$

and let $\mathbf{U}_x$ denote the top-k eigenspace of $\mathbf{M}_x$.

Proposition 5 (One lens for every image).

1. *The global lens $\mathbf{U}_k$ maximises*

$$\mathbb{E}_x \operatorname{tr}(V^\top \mathbf{M}_x V)$$

over all rank-k orthonormal bases V.

2. *For every image,*

$$\operatorname{tr}(\mathbf{U}_x^\top \mathbf{M}_x \mathbf{U}_x) - \operatorname{tr}(\mathbf{U}_k^\top \mathbf{M}_x \mathbf{U}_k) \le 2\|\mathbf{M}_x - \mathbf{M}\|_{(k)}.$$

Proof. For (i), linearity gives

$$\mathbb{E}_x \operatorname{tr}(V^\top \mathbf{M}_x V) = \operatorname{tr}(V^\top \mathbf{M} V),$$

which is maximised by $\mathbf{U}_k$ by Proposition 1. Part (ii) follows from Lemma 1 with $A = \mathbf{M}_x$ and $B = \mathbf{M}$. □

Thus, the global lens is the rank-k subspace that maximises sensitivity on average over images, while its loss relative to an image-specific optimum is controlled directly by the difference $\mathbf{M}_x - \mathbf{M}$. Empirically, the global lens gives a median image-wise sensitivity loss of only 5.0%, consistent with the strong fit stability reported in Appendix A.8.

B.5 THE OUTPUT TERM AND THE TRIANGLE INEQUALITY

For unit vectors a and b,

$$1 - \cos(a, b) = \frac{1}{2}\|a - b\|^2.$$

This is a squared Euclidean distance and does not generally satisfy the triangle inequality. For example, three unit vectors separated by angles 0, θ, and 2θ, with $0 < \theta < \pi/2$, satisfy

$$1 - \cos(2\theta) > 2(1 - \cos\theta).$$

Thus, the lens term is a pseudometric, but the full JLD distance is not guaranteed to be one. Section 4 measures how rarely the full distance violates the triangle inequality in practice. A single Euclidean norm over concatenated lens features and a suitably scaled unit output would restore the triangle inequality and gives similar rankings (Appendix H).

C USAGE AND REPRODUCIBILITY

The released package scores a pair of image paths, arrays, or tensors in three lines. It centre-crops each image to a multiple of 14 pixels, applies the chroma pre-filter, and evaluates Equation 3. A differentiable interface is also provided.

```
from jld import JLD

jld = JLD.pretrained("full")     # use "fast" for JLD-fast
d = jld(ref, dist)
```

The returned scalar is a distance, so smaller values indicate greater similarity. The supplied lens can be used directly and requires no refitting. For a new domain, it can instead be recalibrated from unlabelled images without human ratings (Appendix A.8). The following is the example code. Here we evaluate 224×224 centre crops; replacing `crop224` with `native` evaluates the full image after cropping it to the patch grid.

```
from jld.images import load_pair
from jld.metric import JLD, JLDConfig, WorkspaceStats
from jld.observers import make_observer

r = "/path/to/jlens-artifacts"
s = (
    "cache/workspace/dinov2_s14/"
    "iso_report_cls_seed0_n100_p8/L1.npz"
)

cfg = JLDConfig(
    layers=(1,),
    k=64,
    kind="jlens",
    form="euclid",
    pooling="mean_sq",
    chroma_sigma=2.0,
    report_weight=0.5,
)

enc = make_observer(
    "dinov2_s14",
    layers=[1],
    report="cls",
    device="cpu",
)

metric = JLD(
    enc,
    {1: WorkspaceStats.load(f"{r}/{s}")},
    cfg,
```

```
)

data = f"{r}/data/tid2013/tid2013"
x, y = load_pair(
    f"{data}/reference_images/I04.BMP",
    f"{data}/distorted_images/i04_10_4.bmp",
    14,
    "crop224",
)

print(f"{metric(x, y):.9f}")
```

REFITTING THE LENS

To fit a replacement lens, use representative unlabelled images that are disjoint from evaluation data. The released configuration uses 100 images, four 224×224 crops per image, eight Gaussian probes per crop, and seed 12.

```
from pathlib import Path
import numpy as np

from jld.lens import fit_broadcast, basis

paths = sorted(Path("calibration_images").glob("*.png"))[:100]
assert len(paths) == 100

enc = make_observer(
    "dinov2_s14",
    layers=[1],
    report="cls",
    device="cpu",
)

fit = fit_broadcast(
    enc,
    paths,
    layers=[1],
    n_crops=4,
    crop=224,
    n_probes=8,
    probe="iso",
    seed=0,
    batch=8,
)

fit[1].metadata["image_paths"] = [str(p) for p in paths]
fit[1].save("calibration_L1.npz")

U = basis(fit[1], k=64, kind="jlens")
np.save("lens64.npy", U)

metric = JLD(enc, {1: fit[1]}, cfg)
```

Fitting keeps all encoder parameters frozen but requires gradients with respect to intermediate activations, so it must not be run in inference mode. An accelerator reduces fitting time. For a new domain, only the unlabelled calibration corpus should normally change. The block, rank, and output weight should be changed only with separate development evidence. The corpus manifest, random seed, encoder checkpoint, configuration, and fitted statistics should be stored with the lens.

D GEOMETRY OF THE DISTANCE

This section reports additional measurements supporting Section 4, the feature energy retained by the lens, an additional brightness–noise plane, the stability of a single global lens across images, and the local geometry within a image.

D.1 ISO-DISTANCE CONTOURS: BRIGHTNESS AGAINST NOISE

The third plane of Figure 2 compares a uniform brightness offset with Gaussian noise on TID2013 image I23, with both directions normalised to equal RMS pixel energy. MSE is nearly isotropic on this plane (anisotropy 1.02), whereas JLD stretches its contours by a factor of 2.78 along the brightness direction. LPIPS-VGG and DISTS give similar anisotropies of 2.74 and 2.84, while the local pullback of JLD predicts 3.14. At fixed JLD distance 0.08, PSNR ranges from 19.6 dB for a brightness change to 32.6 dB for noise. Conversely, on a ring of equal MSE at 29.7 dB, JLD varies 4.4-fold, from 0.026 for darkening to 0.114 for noise. This provides another example in which equal pixel error corresponds to substantially different perceptual severity.

D.2 ONE LENS SERVES EVERY IMAGE

The released lens is one fixed matrix shared by every image. To test whether this global choice loses important image-specific structure, we fit a separate lens to each of the 100 DIV2K calibration images and 25 TID2013 references, using the calibration procedure of Appendix A.2 restricted to that image. Each image is fitted twice with independent randomness. We compare two lenses $\mathbf{U}$ and $\widetilde{\mathbf{U}}$ through their subspace overlap $\|\mathbf{U}^\top \widetilde{\mathbf{U}}\|_F^2/k$ (Appendix A.8). A one-image lens has median overlap 0.72 with the released lens, including 0.710 on the held-out TID2013 references. For comparison, independently refitting the same image gives median overlap 0.69, while two random 64-dimensional subspaces have expected overlap 0.167 (Figure A6, top).

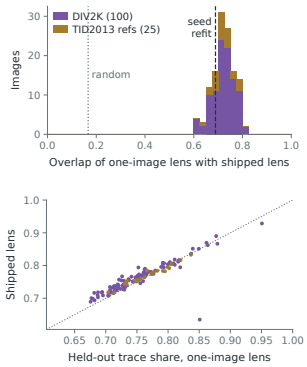

Figure A6: Top: subspace overlap with the released lens. Bottom: sensitivity captured on an independent refit.

We also evaluate sensitivity on the independent refit for each image. On this held-out estimate, the released lens captures more sensitivity than the first one-image fit for 103 of the 125 images (Figure A6, bottom). Pooling the 100 DIV2K one-image estimates also recovers the released lens with subspace overlap 0.954. These results indicate that much of the discrepancy between a one-image lens and the global lens comes from estimation noise rather than stable image-specific structure, supporting the use of one lens across images.

D.3 THE METRIC TENSOR FIELD

In this section, we draw the local form G over a image (Figure A7). Each glyph is G in one 28×28 pixel cell: its radius along a direction is the lens displacement per unit pixel energy of a faint grating whose wave vector points in that direction. A large glyph marks a cell where a small change already moves the distance, and an elongated glyph points along the grating wave direction to which it responds most. We measure G in each of the 234 cells of TID2013 I19 with windowed gratings along x, along y and their sum, in luminance and in the two opponent colour directions, by finite differences of the released lensed tokens (halving the step changes a radius by at most 0.65%). Where a pixel distance draws one circle everywhere, JLD draws a masking map: the glyph radius falls with the local contrast of the cell (Spearman -0.86; -0.95 against gradient energy), and the smoothest quarter of the image is 2.3 times as sensitive as the most textured quarter, the behaviour the grating study of Appendix I finds on flat backgrounds. The glyphs stretch along the local structure: in 91% of the 86 cells with a clear gradient direction the stiff axis of G lies more than 60° from the gradient, so a grating that cuts across the fence is seen and one aligned with it is masked. Colour depends on scale: colour sensitivity is 1.65 times luminance at 1/32 cycle per pixel and 0.44 times at 1/8, and red-green sensitivity falls 5.1-fold from 1/32 to 1/4 while luminance peaks at 1/8.

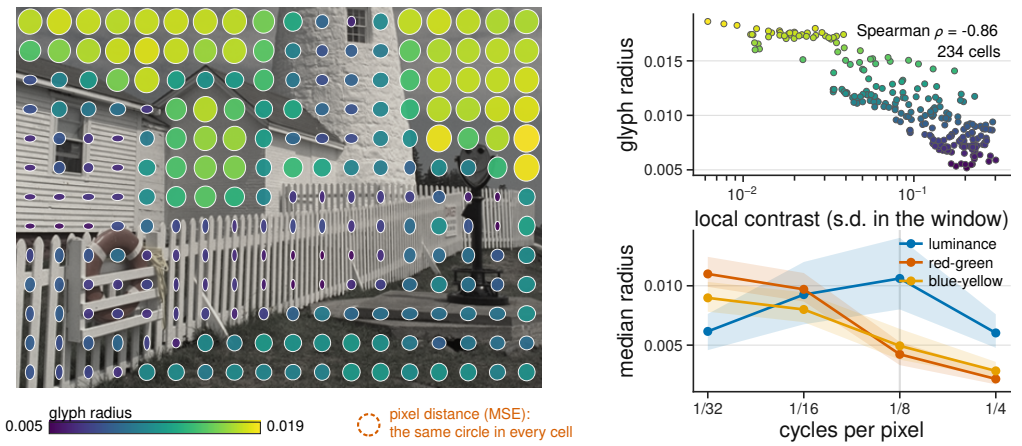

Figure A7: Local form G of JLD over TID2013 I19, one glyph per 28×28 cell (gratings of 1/8 cycle per pixel); the dashed circle is a pixel distance. Right: glyph radius against local contrast, and median radius per colour axis against spatial frequency (bands: interquartile range).

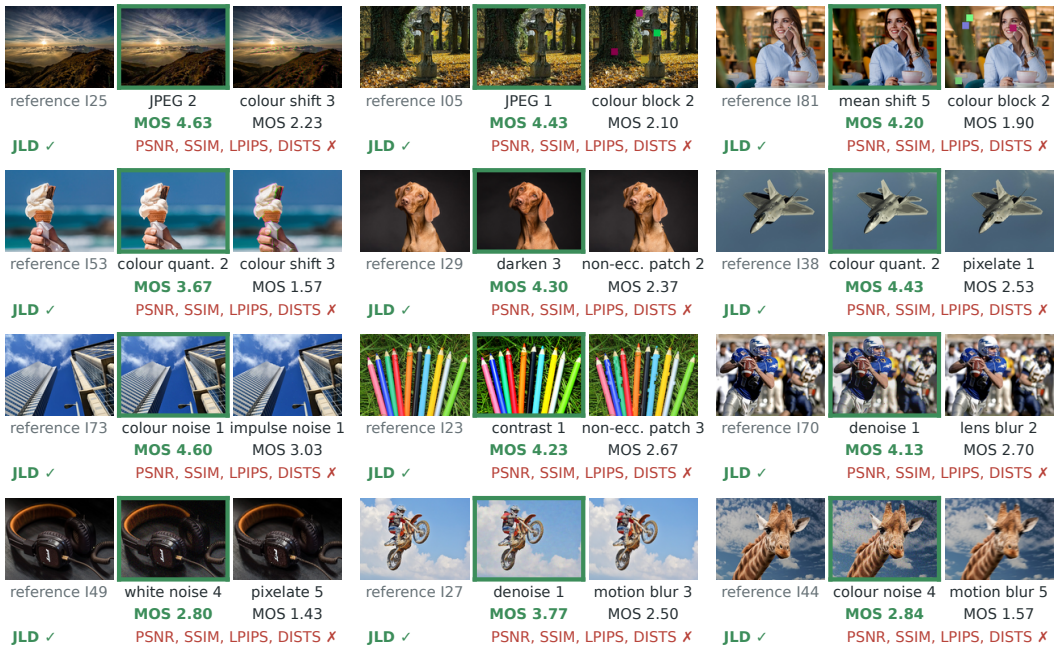

Figure A8: Selected 2AFC cases in which only JLD agrees with human scores. The green-framed image has higher MOS and receives the smaller JLD distance, while PSNR, SSIM, LPIPS-VGG, and DISTS prefer the other image.

E Reading the distance on real images

This section shows JLD on images that can be judged directly by the reader. Following prior perceptual-distance work, we use two-alternative forced-choice (2AFC) examples as in LPIPS (Zhang et al., 2018) and DreamSim (Fu et al., 2023), and group maximum differentiation (gMAD) (Ma et al., 2016) as in DISTS (Ding et al., 2022). All examples come from the held-out KADID-10k test split (Lin et al., 2019), containing 65 references, 25 distortion families, and five levels per family. The scores are produced by the released configuration used in Table 1, and each figure states its selection rule.

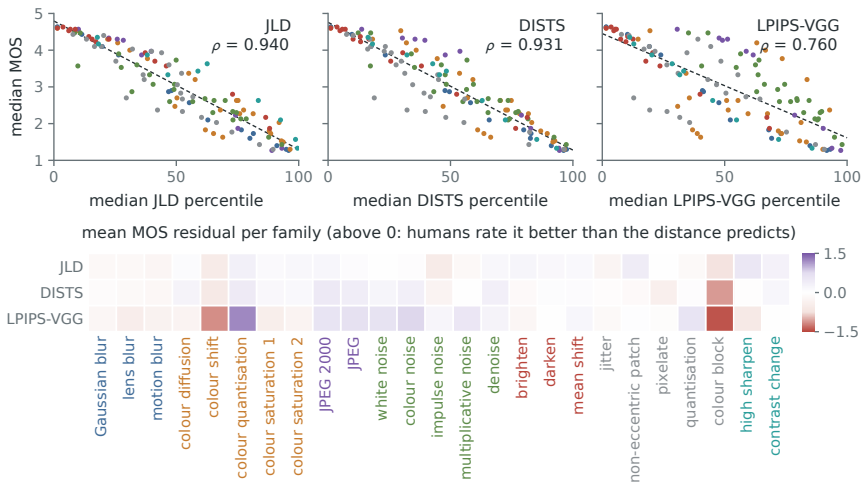

Figure A9: Distance against MOS across distortion families and levels on the KADID-10k test split. Each point is the median over 65 references; dashed lines are linear fits. Bottom: mean MOS residual for each family. Red denotes under-penalisation by the distance and violet over-penalisation.

E.1 2-AFC

For every same-reference pair, we define a 2AFC trial when the MOS difference is statistically significant,

$$|\Delta\text{MOS}| > 1.96\sqrt{\frac{v_A + v_B}{30}},$$

where v_A and v_B are the reported rating variances. This yields 387,603 trials out of 503,750 possible pairs.

JLD agrees with the human preference on 93.3% of these trials, compared with 93.0% for DISTS, 82.9% for LPIPS-VGG, 81.2% for PSNR, and 78.4% for SSIM. There are 1,954 pairs on which JLD alone agrees with humans while all four alternatives disagree, compared with 1,418 pairs showing the reverse pattern. Figure A8 shows examples from the former set, selected by decreasing MOS gap with at most one pair per reference. Many failures of the other distances involve a local defect or colour shift competing with a milder global change such as brightness, contrast, compression, or denoising. Figure 4 reports the corresponding aggregate pairwise analysis across all four image databases without this significance filter.

E.2 Consistency across distortion families

A perceptual distance used as a loss or quality threshold should assign similar values to distortions of similar perceptual severity. If one distance value represented one amount of perceptual damage, the 125 family–level cells in Figure A9 would lie near a common decreasing curve. This behaviour is strongest for JLD and DISTS. Across the 125 cells, JLD reaches SRCC 0.940 with a linear-fit error of 0.38 MOS, while DISTS reaches 0.931 and 0.40. LPIPS-VGG reaches only 0.760 with an error of 0.70 MOS, and its residual exceeds one MOS point for three colour distortion families. The remaining systematic errors of JLD are also visible: it under-penalises colour block distortion by -0.72 MOS and colour shift by -0.53, while over-penalising high sharpening by $+0.51$.

E.3 gMAD stress tests

gMAD (Ma et al., 2016) searches specifically for cases where two distances disagree. For each defender level, we restrict its score to within one percentile point, giving about 162 candidate images, and let the attacker choose its closest and farthest images within that band. The attacker wins when humans significantly prefer its closer image ($z > 1.96$). We repeat this test at nine operating points, from the 10th to the 90th percentile. JLD and DISTS expose different weaknesses in one another: each wins as attacker at 6 of the 9 operating points (Figure A10). DISTS identifies cases where

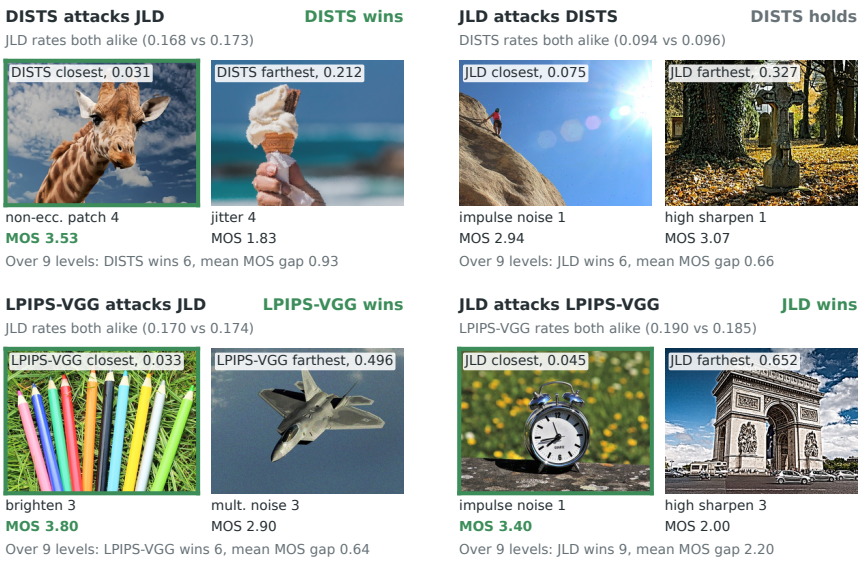

Figure A10: gMAD (Ma et al., 2016) comparisons on the KADID-10k test split. The defender is restricted to images with nearly equal distance, while the attacker selects the pair it considers most different. The attacker wins when human scores significantly prefer its closer image (green frame).

JLD assigns too much cost to a spatially displaced patch relative to jitter. Conversely, JLD identifies cases where DISTS treats high sharpening as similarly mild to light noise or colour diffusion. When attacking LPIPS-VGG, JLD wins at all 9 operating points with a mean human-score gap of 2.20 MOS. LPIPS-VGG wins in the reverse direction at 6 operating points, but with a substantially smaller mean gap of 0.64 MOS.

F Full image results

Tables A3 and A4 report correlations on all six image sets under published preprocessing (P) and identical cropping (C). Because PLCC uses a separate nonlinear mapping for each dataset, it measures within-dataset prediction rather than calibration transfer. JLD reaches SRCC 0.971 on CSIQ development and 0.964 on LIVE. On TID2013, which informed the chroma pre-filter diagnosis, it reaches 0.877. On the held-out sets, JLD reaches 0.892 on KADID-10k test, 0.603 on PIPAL public, and 0.624 on PIPAL validation. DeepDC remains stronger on KADID-10k test and PIPAL, while VSI remains stronger on TID2013. On BAPPS, JLD performs strongly on CNN-generated distortions but is weaker on traditional distortions, with agreement scores of 0.821 and 0.681, respectively. For comparison, LPIPS-VGG reaches 0.814 and 0.714, while DISTS reaches 0.813 and 0.749. Across the remaining BAPPS subsets, JLD reaches 0.625 on colour changes, 0.595 on deblurring, 0.608 on frame interpolation, and 0.694 on super-resolution. LPIPS was trained using BAPPS human judgements, while DISTS used KADID-10k during training (Zhang et al., 2018; Ding et al., 2022). The no-reference extension reaches 0.486 on held-out KonIQ images, while CLIP-IQA reaches 0.685 on the full dataset. Because these evaluations use different subsets, the two numbers should not be directly compared. Appendix **??** instead studies the complementary setting of converting quality-trained encoders into full-reference distances.

Table A3: SRCC and PLCC per dataset. Violet: JLD family; grey: fitted on human judgements. Higher is better; **bold** and underlining mark the best and second distinct displayed values in each column.

	TID2013		CSIQ		LIVE		KADID test		PIPAL pub.		PIPAL val.	
Metric	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
PSNR	0.687	0.679	0.809	0.828	0.873	0.868	0.673	0.678	0.407	0.415	0.255	0.292
SSIM	0.627	0.686	0.837	0.815	0.910	0.915	0.621	0.630	0.504	0.509	0.363	0.426
MS-SSIM	0.786	0.834	0.913	0.899	0.951	0.948	0.826	0.823	0.562	0.593	0.491	0.566
FSIM	0.851	0.877	0.931	0.919	**0.965**	**0.961**	0.853	0.850	0.589	0.615	0.468	0.570
VSI	**0.895**	**0.898**	0.940	0.925	0.949	0.945	0.876	0.876	0.539	0.560	0.450	0.524
GMSD	0.804	0.860	0.957	0.954	0.960	0.960	0.849	0.848	0.583	0.629	0.585	0.658
MAD	0.743	0.759	0.930	0.929	0.951	0.950	0.731	0.733	–	–	0.665	0.687
NLPD	0.799	0.844	0.937	0.927	0.937	0.934	0.811	0.809	0.482	0.509	0.370	0.452
VIF	0.677	0.774	0.919	0.928	0.964	0.960	0.685	0.695	–	–	0.503	0.581
DeepWSD	0.855	0.877	0.959	0.951	0.955	0.951	0.882	0.881	0.561	0.567	0.460	0.491
DeepDC	0.816	0.841	0.954	0.959	0.950	0.947	**0.900**	**0.898**	**0.654**	**0.669**	0.751	0.754
LASI	0.522	0.582	0.840	0.854	0.883	0.881	0.723	0.724	0.514	0.504	0.581	0.571
IEM	0.810	0.818	0.917	0.906	0.922	0.920	0.821	0.819	–	–	0.182	0.243
LPIPS-VGG	0.670	0.759	0.883	0.906	0.932	0.934	0.724	0.732	0.573	0.611	0.612	0.670
LPIPS-Alex	0.744	0.771	0.923	0.920	0.923	0.916	0.824	0.824	–	–	0.620	0.633
DISTS	0.818	0.847	0.943	0.946	0.954	0.953	0.885	0.885	0.615	0.632	0.704	0.722
PieAPP	0.844	0.833	0.897	0.881	0.918	0.910	0.864	0.866	–	–	0.706	0.698
DreamSim	0.812	0.840	0.911	0.928	0.910	0.918	0.851	0.848	–	–	**0.759**	**0.781**
Lens, no pre-filter	0.857	0.871	0.956	0.944	0.942	0.936	0.866	0.865	0.587	0.602	0.575	0.629
JLD-lin	0.826	0.828	0.905	0.885	0.910	0.907	0.809	0.807	0.468	0.473	0.344	0.402
JLD-fast	0.875	0.886	0.956	0.944	0.945	0.939	0.870	0.869	0.592	0.605	0.578	0.634
JLD	0.877	0.897	**0.971**	**0.974**	0.964	**0.961**	0.892	0.889	0.603	0.640	0.624	0.683

Table A4: Comparing KRCC and SRCC

	TID2013		CSIQ		LIVE		KADID test		PIPAL pub.		PIPAL val.	
Metric	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC	KRCC	SRCC
PSNR	0.496	0.686	0.599	0.806	0.680	0.870	0.485	0.675	0.276	–	0.174	–
SSIM	0.455	0.628	0.632	0.837	0.731	0.910	0.449	0.623	0.349	–	0.249	–
MS-SSIM	0.605	0.784	0.738	0.912	0.803	0.949	0.635	0.826	0.397	–	0.341	–
FSIM	0.666	0.737	0.768	0.931	**0.837**	**0.964**	0.664	0.768	0.416	–	0.323	–
VSI	**0.716**	0.812	0.780	0.940	0.799	0.949	0.688	0.837	0.375	–	0.309	–
GMSD	0.634	0.802	0.812	0.957	0.827	0.960	0.665	0.848	0.414	–	0.415	–
MAD	0.545	0.743	0.759	0.930	0.800	0.954	0.557	0.732	–	–	0.471	–
NLPD	0.624	0.791	0.772	0.936	0.777	0.930	0.621	0.812	0.335	–	0.254	–
VIF	0.515	0.678	0.753	0.920	0.828	**0.964**	0.512	0.686	–	–	0.350	–
DeepWSD	0.667	0.874	0.815	0.963	0.810	0.893	0.700	0.887	0.391	–	0.319	–
DeepDC	0.622	0.702	0.810	0.918	0.801	0.947	**0.727**	0.827	**0.469**	–	0.550	–
LASI	0.368	–	0.641	–	0.696	–	0.534	–	0.356	–	0.405	–
IEM	0.624	–	0.735	–	0.747	–	0.629	–	–	–	0.123	–
LPIPS-VGG	0.497	0.668	0.697	0.882	0.765	0.932	0.535	0.726	0.404	–	0.438	–
LPIPS-Alex	0.548	0.748	0.750	0.924	0.751	0.924	0.632	0.827	–	–	0.435	–
DISTS	0.626	0.708	0.788	0.930	0.809	0.947	0.708	0.818	0.437	–	0.510	–
PieAPP	0.657	0.846	0.711	0.896	0.749	0.918	0.674	0.864	–	–	0.512	–
DreamSim	0.614	0.810	0.738	0.915	0.745	0.904	0.657	0.853	–	–	**0.563**	–
Lens, no pre-filter	0.672	0.857	0.811	0.956	0.785	0.942	0.677	0.866	0.413	0.587	0.406	0.575
JLD-lin	0.648	0.826	0.723	0.905	0.734	0.910	0.617	0.809	0.322	0.468	0.237	0.344
JLD-fast	0.691	0.875	0.811	0.956	0.790	0.945	0.683	0.870	0.417	0.592	0.408	0.578
JLD	0.693	**0.877**	**0.849**	**0.971**	0.830	**0.964**	0.711	**0.892**	0.426	**0.603**	0.443	**0.624**

Table A5: Building JLD step by step on one frozen encoder (SRCC), and other distances on the same tokens. Random directions gain nothing; the directions of $\mathbf{M}$ give most of the gain. **Bold** marks the best and underlining the second distinct displayed value. † marks a score on 1,000 TID2013 pairs.

Variant	TID2013 (diag.)	CSIQ (dev.)	LIVE	KADID-10k test	Mean
Full block 1	0.660	0.841	0.859	0.727	0.772
Random 64	0.660	0.841	0.859	0.726	0.772
PCA 64	0.766	0.912	0.925	0.800	0.851
Lens 64	0.857	0.956	0.942	0.866	0.905
+ pre-filter	0.875	0.956	0.945	0.870	0.911
+ output term = JLD	**0.877**	**0.971**	**0.964**	**0.892**	**0.926**
Output only	0.748	0.840	0.910	0.808	0.826
Pointwise Jacobian	0.679†	–	–	–	–
Other distances on the same tokens					
Uniform LPIPS	0.661	0.867	0.912	0.740	0.795
Standardised Euclidean	0.611	0.802	0.832	0.684	0.732
DeepWSD-style	0.662	0.843	0.865	0.728	0.774
DeepDC-style	0.615	0.827	0.819	0.763	0.756

G Additional results

G.1 Ablations

Table A5 builds JLD one component at a time on DINOv2-S block 1 and compares alternative distances on the same token features. PCA 64 uses regularised Mahalanobis distance ($\rho = 10$), while lens 64 uses Euclidean distance; all other token baselines use the same unfiltered block-1 features. Using all 384 raw features gives a four-dataset mean SRCC of 0.772, and projecting onto 64 random directions leaves it unchanged. The 64 leading PCA directions improve the mean to 0.851, while the 64 leading eigenvectors of $\mathbf{M}$ reach 0.905 and improve every dataset. Thus, the main gain comes from ranking feature directions by the encoder output's sensitivity rather than by feature variance or dimensionality reduction alone. The output term alone reaches 0.826, showing that most of the signal comes from the lens term. The chroma pre-filter raises TID2013 from 0.857 to 0.875, while the output term raises LIVE from 0.945 to 0.964 and KADID-10k test from 0.870 to 0.892. On the same lensed tokens, a DeepWSD-style comparison matches the Euclidean lens at 0.905, whereas a DeepDC-style comparison falls to 0.744. The simple Euclidean distance therefore gives up no accuracy once the feature directions have been selected by the Jacobian lens.

G.2 Resolution results

Table A6 gives the TID2013 results underlying Section 5.4 for 11 of the 15 evaluated methods. The full database is rescored after resizing each image to short sides of 256 and 512 pixels, while the original human scores are kept fixed. This tests whether a distance preserves its agreement with human judgements when the same content is evaluated at a different image resolution. The lens remains nearly unchanged, from 0.850 at 256 pixels to 0.845 at 512 pixels. By comparison, the unweighted block-1 features fall from 0.784 to 0.656, DISTS from 0.815 to 0.717, and LPIPS-VGG from 0.757 to 0.661. These results support the central observation of Section 5.4: weighting early features by output sensitivity is substantially more stable to resolution changes than comparing the same early features directly.

Table A6: TID2013 SRCC at 256 and 512 px short side. The lens remains stable while several deep feature distances degrade.

Method	256	512	Change
Deep features			
Lens	0.850	0.845	**−0.005**
PieAPP	0.618	0.700	+0.081
LPIPS-VGG	0.757	0.661	−0.096
DISTS	0.815	0.717	−0.099
DeepDC	0.806	0.695	−0.111
LPIPS-Alex	0.824	0.697	−0.127
Classical			
NLPD	0.787	0.787	0.000
PSNR	0.712	0.713	0.000
VSI	0.887	0.888	+0.001
MS-SSIM	0.770	0.779	+0.009
SSIM	0.727	0.641	−0.086

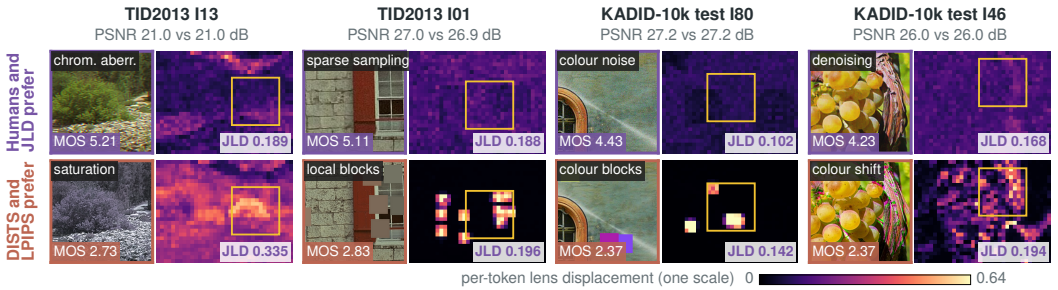

Figure A11: At equal PSNR, JLD ranks the pair as people do while DISTS and LPIPS-VGG reverse it. Each column shows a detail crop (left) and JLD's per-token lens displacement over the whole image (right; one colour scale; the box marks the crop).

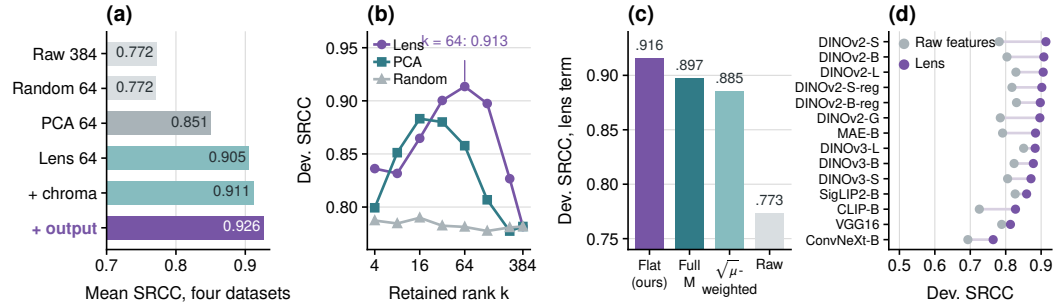

Figure A12: Which parts of JLD carry the agreement. (a) Building the distance on DINOv2-S block 1 (mean SRCC, four datasets). (b) Development SRCC against rank k. (c) Weightings of the retained directions. (d) Lens (violet) against raw features (grey) for 14 frozen encoders.

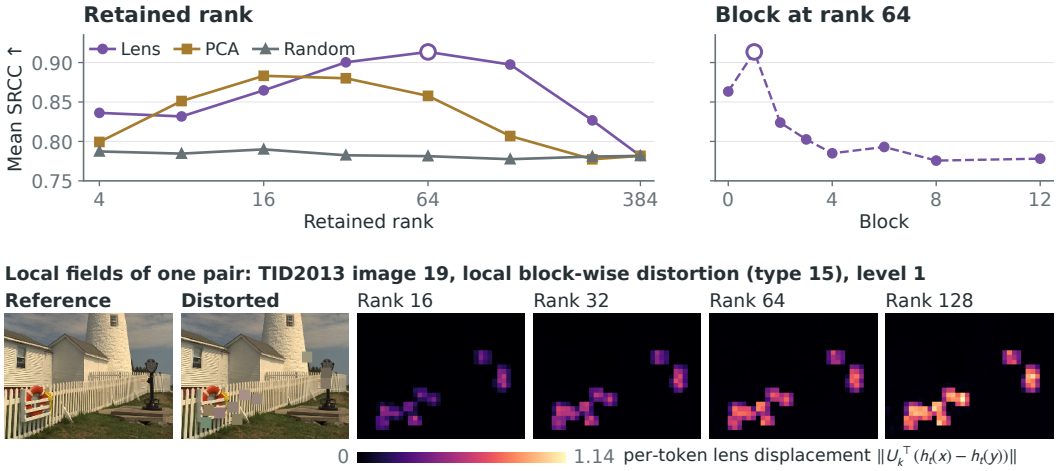

Figure A13: Mean development SRCC over CSIQ and KADID-10k development as a function of retained rank and encoder block. Bottom: block-1 lens maps at several ranks. Diagnostic fits use the preliminary combined output.

H Development choices

Labelled data are used only to select distance settings, never to fit the lens. We evaluate 245 configurations on CSIQ and KADID-10k development and choose the simplest configuration within 0.005 of the best equally weighted mean SRCC. The selected CLS-target configuration reaches 0.915, compared with 0.916 for the best tested configuration.

The sweep considers $k \in \{4, 8, 16, 32, 64, 128, 256, 384\}$. Figure A13 shows that the Jacobian basis consistently outperforms PCA and random projections, with block 1 and $k = 64$ selected for the

Table A7: Moderate output weights give similar four-dataset means. Violet: selected settings; white: controls. **Bold** and underlining mark the best and second displayed values.

Form	TID2013 (diag.)	CSIQ (dev.)	LIVE	KADID test	KADID dev	Mean
Lens alone ($\lambda = 0$)	0.875	0.956	0.945	0.870	0.876	0.911
Sum ($\lambda = 0.25$)	**0.877**	0.969	0.959	0.885	0.887	0.923
JLD ($\lambda = 0.5$)	**0.877**	**0.971**	0.964	0.892	0.892	0.926
Sum ($\lambda = 0.75$)	0.875	0.970	0.966	0.895	0.895	**0.927**
Sum ($\lambda = 1$)	0.872	0.968	**0.967**	**0.897**	**0.896**	0.926
Sum ($\lambda = 2$)	0.861	0.959	0.965	**0.897**	0.894	0.920
Output alone	0.748	0.840	0.910	0.808	0.812	0.826
Single norm ($w = 0.375$)	0.871	0.966	0.963	0.893	0.893	0.923

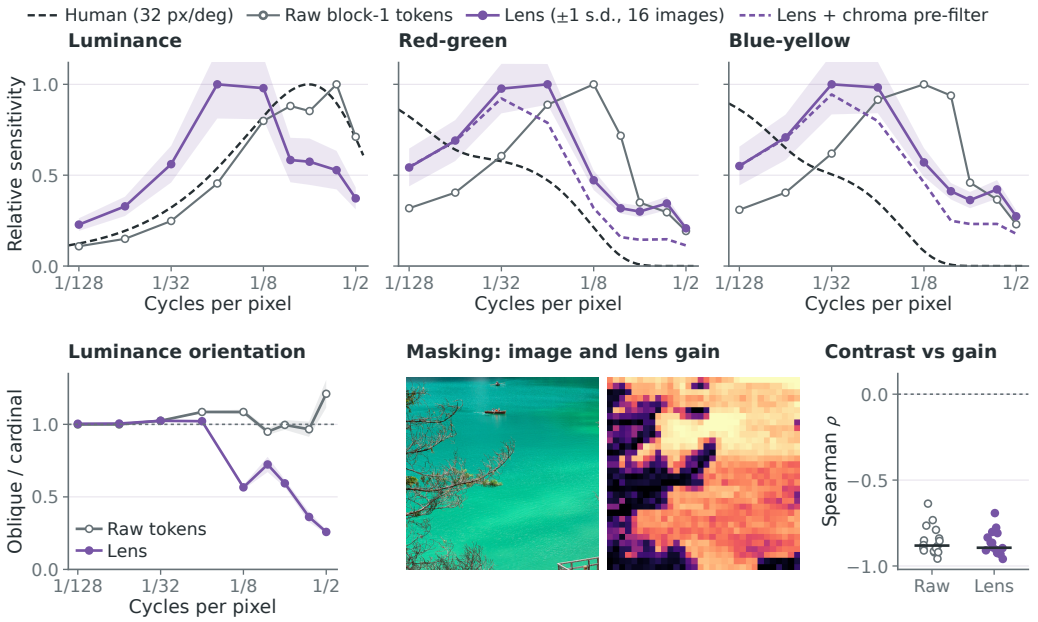

Figure A14: Grating probes of the lens on 16 DIV2K crops. Top: peak-normalised sensitivity across spatial frequency for luminance, red–green, and blue–yellow gratings, compared with human references. Bottom left: sensitivity to oblique relative to cardinal gratings. Bottom right: local lens gain for a luminance grating (bright denotes higher sensitivity).

final distance. As an additional control, a lens pooled over natural images reaches 0.861 SRCC on 1,000 TID2013 pairs, compared with 0.679 for sensitivity estimated separately from each reference image. The output-term weight is fixed from CSIQ and KADID-10k development. Matching the score spreads gives weights 0.312 and 0.696 on the two sets, whose mean 0.504 is rounded to $1/2$. Performance is insensitive to this choice: weights from 0.25 to 1 give mean SRCC between 0.923 and 0.927 across the four standard image sets. A single Euclidean norm over the concatenated lens and output features reaches 0.923, compared with 0.926 for the released sum; unlike the released sum, it satisfies the triangle inequality (Appendix B).

I What the lens sees in gratings

We probe JLD with classical psychophysical stimuli to ask whether a lens fitted without human judgements recovers known properties of visual sensitivity. Following contrast-sensitivity experiments (Campbell & Robson, 1968; Mullen, 1985), we add sinusoidal gratings that vary in spatial frequency, colour direction, orientation, and background texture, and compare the lens with raw block-1 features. Figure A14 shows three effects. First, the lens suppresses high-frequency chromatic changes substantially more than raw block-1 features; the fixed chroma pre-filter moves this

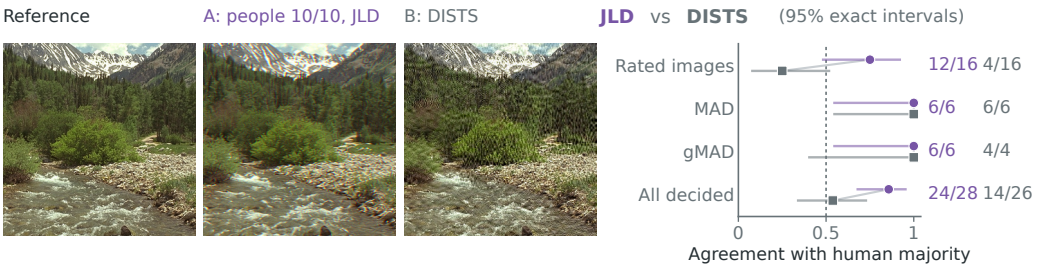

Figure A15: Human-study agreement. Left: one rated-image trial in which all ten participants and JLD chose A and DISTS chose B. Right: share of decided trials on which each metric agrees with the human majority (exact 95% binomial intervals; dotted: chance).

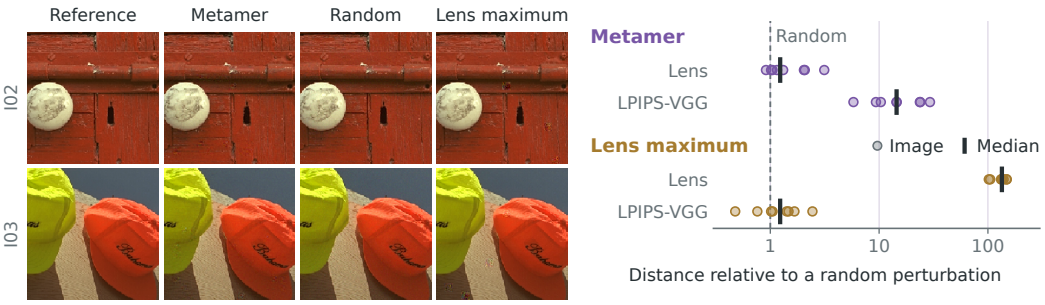

Figure A16: Approximate metamers, random perturbations, and lens-maximising perturbations at nominal 40 dB PSNR. Right: distances relative to the same-reference random perturbation; the dashed line denotes 1.

response further toward human chromatic sensitivity (Mullen, 1985; Zhang & Wandell, 1996). Second, the lens is less sensitive to fine oblique gratings than to horizontal and vertical ones, reproducing the qualitative oblique effect that is absent from the raw features. Third, its local gain decreases in textured regions: for a $1/8$ cycle-per-pixel luminance grating, gain has median Spearman correlation -0.89 with local luminance contrast across the 16 images, consistent with visual masking. The agreement is not complete. In particular, the lens does not reproduce the human luminance contrast-sensitivity function: its response peaks at a lower spatial frequency. The unfiltered lens also remains too sensitive to the finest chromatic patterns, motivating the fixed $\sigma = 2$ chroma pre-filter described in Appendix A.5. Thus, the frozen encoder already contains several human-like sensitivity patterns, while the pre-filter corrects its clearest systematic mismatch.

J Validation and Human Study

We provide additional checks of the geometry of JLD, a small human study on cases where perceptual metrics disagree, and targeted perturbations for directions retained or discarded by the lens.

J.1 Human Study

Ten participants completed 40 trials selected where perceptual metrics disagree, choosing which distorted image was closer to its reference, with a "cannot tell" option and randomised sides. Across trials with an untied majority, the lens agrees with the majority on $24/28$ decisions, compared with $14/26$ for DISTS. On the 16 trials drawn from rated image pairs, the majority agrees with MOS on all 16; JLD agrees on $12/16$ and DISTS on $4/16$.

J.2 Metamers and Targeted Perturbations

A displacement orthogonal to the lens has zero projected distance, providing a direct way to probe what the metric ignores. For eight references, we construct approximate metamers, lens-maximising perturbations, and random perturbations at the same nominal 40 dB PSNR. As shown in Figure A16, approximate metamers remain close under the unfiltered lens (median ratio 1.23 relative to random perturbations) while LPIPS-VGG rises to 14.5. Conversely, lens-maximising perturbations increase

Table A8: Development-selected projections across encoder families. Violet: lens; white: controls. **Bold** and underlining mark the best and second displayed values in each comparison.

			Selected block					Best block	
Encoder	Block	Rank	Lens	Raw	PCA 64	Δ raw	Δ PCA	Raw	PCA 64
DINOv2-S	1	64	**0.913**	0.782	0.858	**+0.132**	+0.056	0.807	0.858
DINOv2-B	1	64	0.908	0.804	0.859	+0.104	+0.049	0.804	0.859
DINOv2-L	4	32	0.906	0.829	**0.891**	+0.077	+0.015	0.829	**0.895**
DINOv2-G	4	16	0.896	0.785	0.875	+0.112	+0.022	0.795	0.875
DINOv2-S-reg	1	64	0.902	0.818	0.871	+0.084	+0.031	0.820	0.871
DINOv2-B-reg	1	32	0.898	0.831	0.873	+0.067	+0.024	0.831	0.873
DINOv3-S	1	64	0.871	0.805	0.825	+0.066	+0.047	0.816	0.828
DINOv3-B	2	16	0.878	0.824	0.825	+0.054	+0.053	0.842	0.846
DINOv3-L	3	32	0.883	**0.851**	0.845	+0.033	+0.038	**0.857**	0.859
MAE-B	2	64	0.884	0.791	0.850	+0.093	+0.034	0.833	0.868
CLIP-B	1	16	0.828	0.726	0.850	+0.102	-0.022	0.787	0.883
SigLIP2-B	4	64	0.859	0.827	0.803	+0.032	+0.056	0.827	0.803
VGG16	3	16	0.813	0.789	0.806	+0.025	+0.007	0.802	0.813
ConvNeXt-B	1	16	0.765	0.693	0.697	+0.071	**+0.068**	0.764	0.760

the lens distance by a factor of 134, while LPIPS-VGG changes by only 1.23. These results confirm that the lens is strongly directional, it suppresses feature changes outside the output-sensitive subspace while strongly responding to changes aligned with it.

K Encoders

We apply the same calibration procedure to 14 frozen encoders spanning DINOv2, DINOv3, MAE, CLIP, SigLIP 2, VGG16, and ConvNeXt (Oquab et al., 2024; Siméoni et al., 2025; He et al., 2022; Radford et al., 2021; Tschannen et al., 2025; Simonyan & Zisserman, 2015; Liu et al., 2022). Each encoder receives its own lens, fitted on the same unlabelled DIV2K corpus, with block and rank selected on CSIQ and KADID-10k development. As shown in Table A8 and Figure A12d, the lens improves over raw features for all 14 encoders and over PCA 64 for 13 of 14. DINOv2-S gives the strongest development result and is therefore used as the default encoder. The same trend extends to video. at block 2, VideoMAE K400 reaches SRCC 0.786 with lens 16, compared with 0.771 for PCA 64 and 0.710 for full features. A new encoder therefore requires only a frozen output and an intermediate feature representation; the estimator of Appendix A.2 can then be used to fit its lens.

L Video

JLD-video applies the same construction to a frozen VideoMAE-Base fine-tuned on Kinetics-400 (Tong et al., 2022). We use block 2 with a 16-dimensional lens, fitted to the class logits of 50 unlabelled clips using eight random probes. For each video pair, we sample eight aligned 16-frame clips, project the space-time token differences through the lens, average their squared lengths over tokens and clips, and take the square root. Video uses neither the chroma pre-filter nor an output term. On Waterloo IVC 4K (Li et al., 2019), used to select the block and rank, JLD-video reaches SRCC 0.786. For comparison, PCA 64 reaches 0.771, all 768 features 0.710, VMAF 0.611, and PSNR 0.562. The frame-wise image JLD reaches 0.710, showing that the space-time representation captures information unavailable from individual frames.

L.1 Held-out video

We evaluate the frozen Waterloo-selected configuration on all 120 released AVT-VQDB-UHD-1 test pairs (Rao et al., 2019); no AVT video is used for fitting or model selection. Decoded videos are scaled to a common display resolution of 1920×1080, and all methods are evaluated under the same display protocol. On AVT, JLD-video reaches SRCC 0.912, compared with 0.932 for VMAF, 0.908 for PSNR, and 0.884 for MS-SSIM. On Waterloo, JLD-video reaches 0.786 compared with 0.611 for VMAF. Thus, JLD-video is the only one of these four methods above 0.78 on both datasets,

Table A9: Video results on both datasets. Within: mean per-source SRCC. Lower: the smaller of the two dataset SRCCs. Bold marks the best value in each column.

	Fitted to	Waterloo 4K (dev., 240)		AVT-UHD (held out, 120)					AVT
Metric	video MOS	SRCC	Within	SRCC	KRCC	PLCC	Within	Lower	encoded
JLD-video	no	**0.786**	0.967	0.912	0.761	0.918	0.944	**0.786**	0.582
VMAF	yes	0.611	0.834	**0.932**	**0.790**	**0.939**	0.953	0.611	0.663
MS-SSIM	no	0.694	0.967	0.884	0.718	0.888	0.945	0.694	0.561
PSNR	no	0.562	**0.968**	0.908	0.758	0.915	**0.964**	0.562	**0.754**

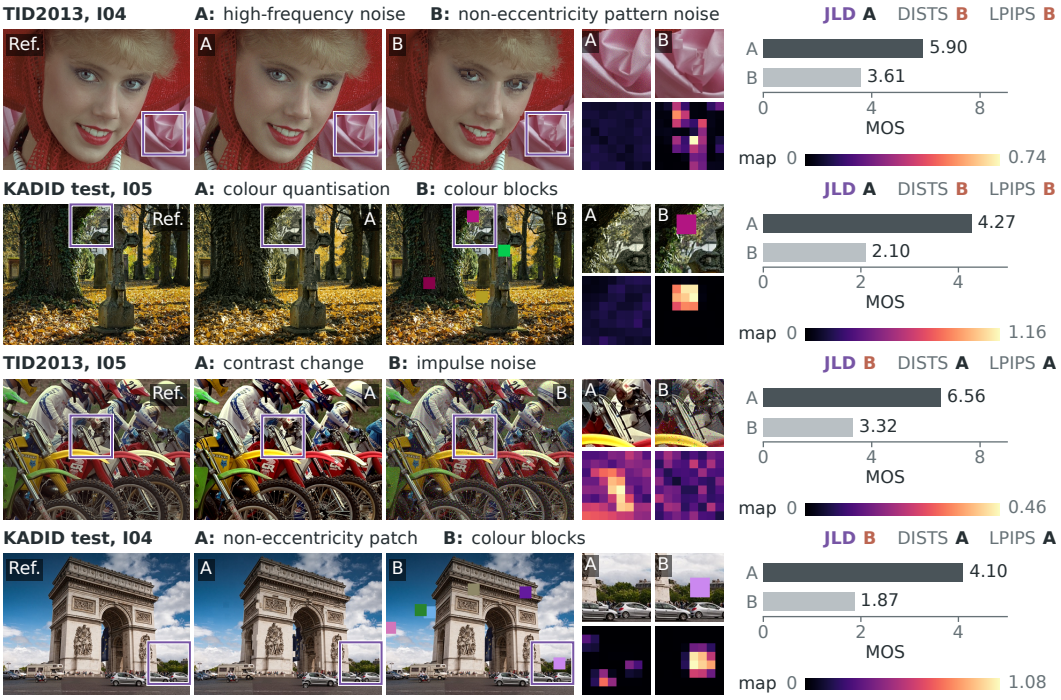

Figure A17: Four same-reference pairs in which humans prefer A. JLD agrees in the top two rows and reverses the preference in the bottom two. Maps share one scale within each row and show the lens term only.

despite using no subjective video scores during fitting. Figure A18 shows representative cases where JLD-video and VMAF agree or disagree with human scores.

M Matched pixel error

Figure A17 shows representative same-reference pairs with similar pixel error where human judgements distinguish the distortions. JLD agrees with the human preference in the top two examples and fails in the bottom two, providing both representative successes and failure cases.

N Limitations

JLD inherits limitations of its encoder. It is weaker on restoration outputs (0.624 SRCC on PIPAL) and on global contrast changes (0.378 on TID2013; Appendix G). The lens term is a pseudometric with a non-trivial null space: displacements in the 320 discarded feature directions are invisible to it, so direct optimisation against the lens can exploit these directions (Ding et al., 2021). Although JLD is differentiable, using it as a training loss therefore requires care. The output term removes some of this ambiguity but sacrifices the triangle inequality, with violations on at most 107 per million audited triplets; applications requiring a strict pseudometric can use the lens term alone. The lens itself uses no human labels, but CSIQ and KADID-10k development data select the block, rank, and output weight; LIVE, KADID-10k test, and PIPAL are held out. On held-out video, JLD-video reaches 0.912 versus 0.932 for VMAF (Appendix L). Full JLD is also slower than DISTS;



Figure A18: Four Waterloo examples. Each row shows the reference and two distorted videos at synchronized frames; row titles report MOS and the JLD-video and VMAF choices. Coral marks disagreement with MOS and visible artifacts.

the approximately four-fold speed advantage applies to JLD-fast. Finally, the human study contains only ten participants, and JLD assumes spatially aligned inputs.