

Supplementary Material for *Do computer vision foundation models learn the low-level characteristics of the human visual system?*

Yancheng Cai, Fei Yin, Dounia Hammou, Rafal Mantiuk; University of Cambridge, UK

yc613@cam.ac.uk, fy277@cam.ac.uk, dh706@cam.ac.uk, mantiuk@gmail.com

This supplementary material provides detailed information on the following: (1) the formulas and examples of the experimental stimuli tested; (2) some further practical implications of our work; and (3) the detailed formula for model alignment scores, along with the model alignment scores for all models across all tests.

Please open webpage/index.html for the complete set of the results.

1. Test images

The achromatic Gabor patches used for tests are defined as:

$$G(x, y) = L_b \left(1 + c \sin \left(2\pi \frac{\rho x}{\text{ppd}} \right) e^{\left(-\frac{x^2 + y^2}{2 \text{ppd}^2 R^2} \right)} \right), \quad (1)$$

where L_b denotes the background/mean luminance in cd/m^2 , c represents the contrast, R is the Gabor radius in visual degree, and ρ is the spatial frequency in cycles-per-degree (cpd). x and y represent the image coordinates, where $x \in [-\frac{W}{2}, \frac{W}{2}]$ and $y \in [-\frac{H}{2}, \frac{H}{2}]$; $W = 224$ and $H = 224$ denote the image width and height, respectively.

To generate chromatic (RG and YV) Gabor stimuli, a single-channel Gabor patch is first created, the color direction is set, then converted to DKL color space, transformed to LMS color space, and finally converted back to RGB to check for gamut constraints. Note that the RGB channel values here are still represented in cd/m^2 units.

The luminance values, $G(x, y)$, were converted to RGB values using the sRGB display model, assuming the peak luminance of 400 cd/m^2 . The pixels-per-degree (ppd) was set to 60, which approximates the ppd value for a typical human observer viewing an Ultra HD display (3840×2160). The resolution of all test and reference images was set to 224×224 . Except for Supra-threshold Contrast Matching (Section 1.3), the references for all other eight experiments are uniform achromatic images with a luminance of 100 cd/m^2 .

1.1. Contrast detection

Spatial Frequency - Achromatic - Gabor The radius was set to 1° , and the background luminance was 100 cd/m^2 . Test examples are shown in Figure 1.

Spatial Frequency - Achromatic - Band-limited Noise

The background luminance was 100 cd/m^2 . Test examples are shown in Figure 2.

Spatial Frequency - Chromatic (RG) - Gabor

The radius was set to 1° , and the background luminance was 100 cd/m^2 . Test examples are shown in Figure 3.

Spatial Frequency - Chromatic (YV) - Gabor

The radius was set to 1° , and the background luminance was 100 cd/m^2 . Test examples are shown in Figure 4.

Luminance The radius was set to 1° , and the spatial frequency was 2 cpd. Test examples are shown in Figure 5.

Area The background luminance was 100 cd/m^2 , and the spatial frequency was 8 cycles per degree (cpd). Test examples are shown in Figure 6.

1.2. Contrast masking

Phase-Coherent Masking The test images contained Gabor patches with a spatial frequency of 2 cpd and a radius of 0.5° , while the masks were sinusoidal gratings at the same spatial frequency of 2 cpd. The background luminance was 32 cd/m^2 , following the parameters established in [1]. Test examples are shown in Figure 7.

Phase-Incoherent Masking The test images contained Gabor patches with a spatial frequency of 1.2 cpd and a radius of 0.8° , and the masks contained random noise with a frequency spectrum extending up to 12 cpd. The following equations outline the process of generating the noise mask $I_{\text{mask}} \in \mathbb{R}^{W \times H}$:

First, Gaussian noise $N(x, y)$ is generated:

$$N(x, y) \sim \mathcal{N}(0, 1), \quad x \in [0, W], y \in [0, H]. \quad (2)$$

Next, a two-dimensional Fast Fourier Transform (FFT) is applied to obtain the frequency domain representation $N_f(u, v)$:

$$N_f(u, v) = \mathcal{F}\{N(x, y)\}, \quad u \in [0, W], v \in [0, H]. \quad (3)$$

Subsequently, frequency filtering is performed:

$$N_f^{\text{filtered}}(u, v) = \begin{cases} N_f(u, v), & \rho(u, v) \leq 12 \text{ cpd} \\ 0, & \rho(u, v) > 12 \text{ cpd} \end{cases}, \quad (4)$$

$$\rho(u, v) = \sqrt{(K_u)^2 + (K_v)^2}, \quad (5)$$

$$K_u = 2 \rho_{\text{nyquist}} \left(\text{mod} \left(\frac{1}{2} + \frac{u}{W}, 1 \right) - \frac{1}{2} \right), \quad (6)$$

$$K_v = 2 \rho_{\text{nyquist}} \left(\text{mod} \left(\frac{1}{2} + \frac{v}{H}, 1 \right) - \frac{1}{2} \right), \quad (7)$$

where $\rho_{\text{nyquist}} = \frac{\text{ppd}}{2}$. The noise in the spatial domain is then obtained using the inverse Fourier transform:

$$N_{\text{bp}}(x, y) = \mathcal{F}^{-1} \{ N_f^{\text{filtered}}(u, v) \}. \quad (8)$$

Finally, the noise mask I_{mask} is generated:

$$I_{\text{mask}}(x, y) = L_b \left(1 + c_{\text{mask}} \frac{N_{\text{bp}}(x, y)}{\sigma_{N_{\text{bp}}}} \right), \quad (9)$$

where c_{mask} is the mask contrast, $\sigma_{N_{\text{bp}}}$ represents the standard deviation of N_{bp} . The background luminance L_b was 37 cd/m², consistent with the conditions in [2]. Test examples are shown in Figure 8.

1.3. Supra-threshold contrast matching

We followed the experimental setup from [3], where the reference was a sinusoidal grating with a spatial frequency of 5 cpd and a luminance of 10 cd/m², presented at eight distinct contrast levels c_r . The test stimulus was also a sinusoidal grating with a luminance of 10 cd/m², but presented at various spatial frequencies ρ_t . Examples are shown in Figure 9.

2. Practical implications

We checked whether our alignment scores (Fig. 6 in the paper) can indicate how well a model can perform on computer vision tasks. In Figure 10, we show scatter plots of the alignment scores and different performance indicators for DINO, DINOv2, and OpenCLIP (data were not available for other models). The correlations (absolute value 0.55–0.8) suggest that good alignment with the contrast masking/matching characteristic can improve model’s performance. Such results were consistent for the alignment of contrast masking and contrast matching, less so for detection (as expected). We did not find a strong correlation between alignment scores and the parameters of the model architecture (model size, number of parameters) or computational GFlops. We hope that future work can provide stronger evidence for the benefits of model-HVS alignment and spark interest in using low-level human vision models to introduce invariances or constraints into the training of the foundation models (via architectural changes, loss functions, or data augmentation).

3. Model alignment scores

Section 3.2 in the main text briefly describes the computation of Spearman rank-order correlation coefficients for model alignment scores. This section provides further details and formulas.

Specifically, for each contour plot, N points were selected along the x-axis $X_{1 \dots N}$, where X represents dimensions such as area, luminance, or mask contrast. Based on the predictions of castleCSF, we obtain the ground truth $Y_{1 \dots N}$, where Y represents sensitivity in contrast detection and test contrast in contrast masking.

We then scaled each $Y_j (j = 1 \dots N)$ by multipliers $m_i (i = 1 \dots M)$:

$$m_i = 10^{\log_{10}(0.5) + \frac{i-1}{M-1} \cdot \log_{10}(\frac{2}{0.5})}, \quad (10)$$

$$Y'_{ij} = m_i Y_j, \quad (11)$$

producing $Y'_{1 \dots NM}$ and their respective $S_{1 \dots NM}$ (S_{ac})¹. Given that psychometric functions near the threshold typically exhibit uniform shapes across all conditions in psychophysical experiments, we hypothesized that the trend of scaled scores would remain consistent across all $Y_{1 \dots N}$. The Spearman’s rank correlation coefficient r_s was calculated as the similarity metric:

$$r_s = \frac{\text{cov}(R(m_{1 \dots NM}), R(S_{1 \dots NM}))}{\sigma_{R(m_{1 \dots NM})} \sigma_{R(S_{1 \dots NM})}}, \quad (12)$$

where $m_{1 \dots NM} = \bigcup_{k=1}^N m_{1 \dots M}$, $R(*)$ denotes ranked data, $\text{cov}(*)$ represents covariance, and $\sigma(*)$ stands for standard deviation. For all models and tests, higher r_s (closer to 1) reflects a greater model alignment. In the contrast detection experiment, $N = 20$, $M = 10$. In the contrast masking experiment, $M = 10$ and N is equal to the number of human data points.

In the main text, we presented the experimental results for all models in the form of bar charts. To provide higher decimal precision, the results are presented in Table 1.

References

- [1] John M Foley. Human luminance pattern-vision mechanisms: masking experiments require a new model. *JOSA A*, 11(6): 1710–1719, 1994. 1
- [2] Karl R Gegenfurtner and Daniel C Kiper. Contrast detection in luminance and chromatic noise. *JOSA A*, 9(11):1880–1888, 1992. 2
- [3] MA Georgeson and GD Sullivan. Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of physiology*, 252(3):627–656, 1975. 2, 7

¹Specifically, for these NM conditions, a test and reference signal pair is generated for each condition (X_j, Y'_{ij}) , and S_{ac} is computed using the method described in the main text.

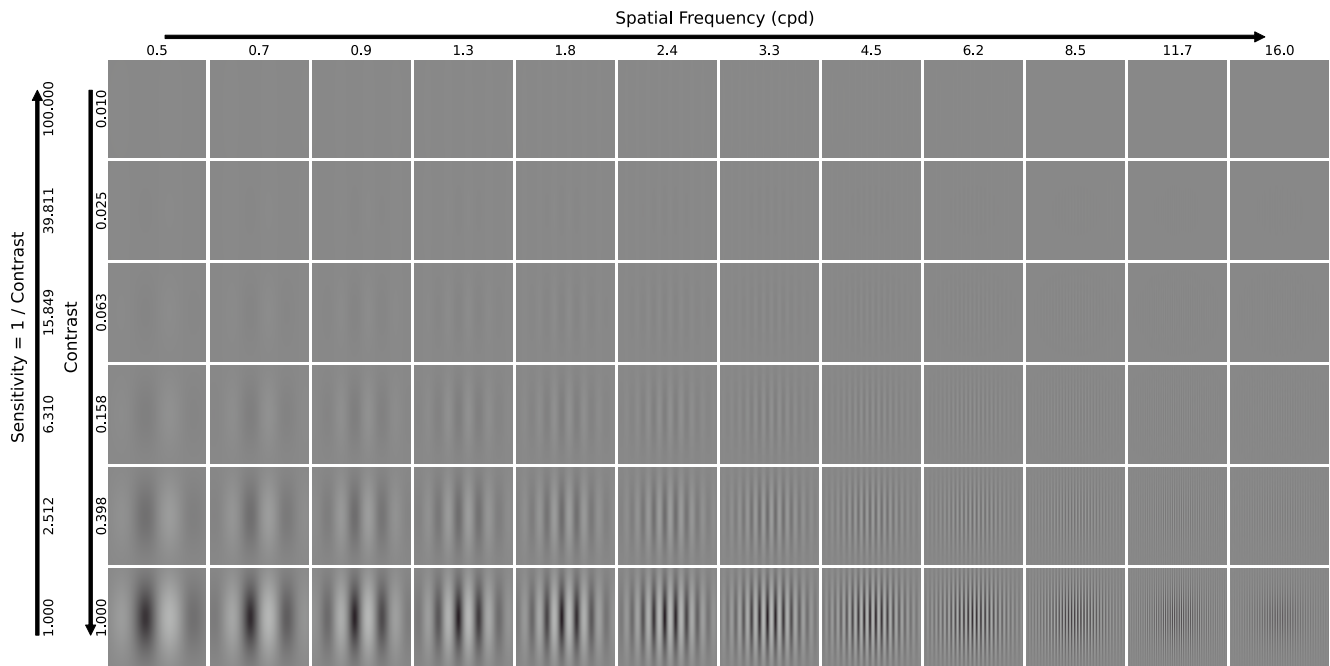


Figure 1. Achromatic Gabors with different spatial frequencies (x-axis) and contrast (y-axis) used as the test images in the contrast detection tests. “cpd” denotes cycles per degree. High-frequency patterns may introduce aliasing artifacts on screens or prints, so we display up to 16 cpd here (no such artifacts were present in our tests). Observations indicate that the human eye is indeed most sensitive to achromatic Gabor patterns with spatial frequencies around 2–4 cpd.

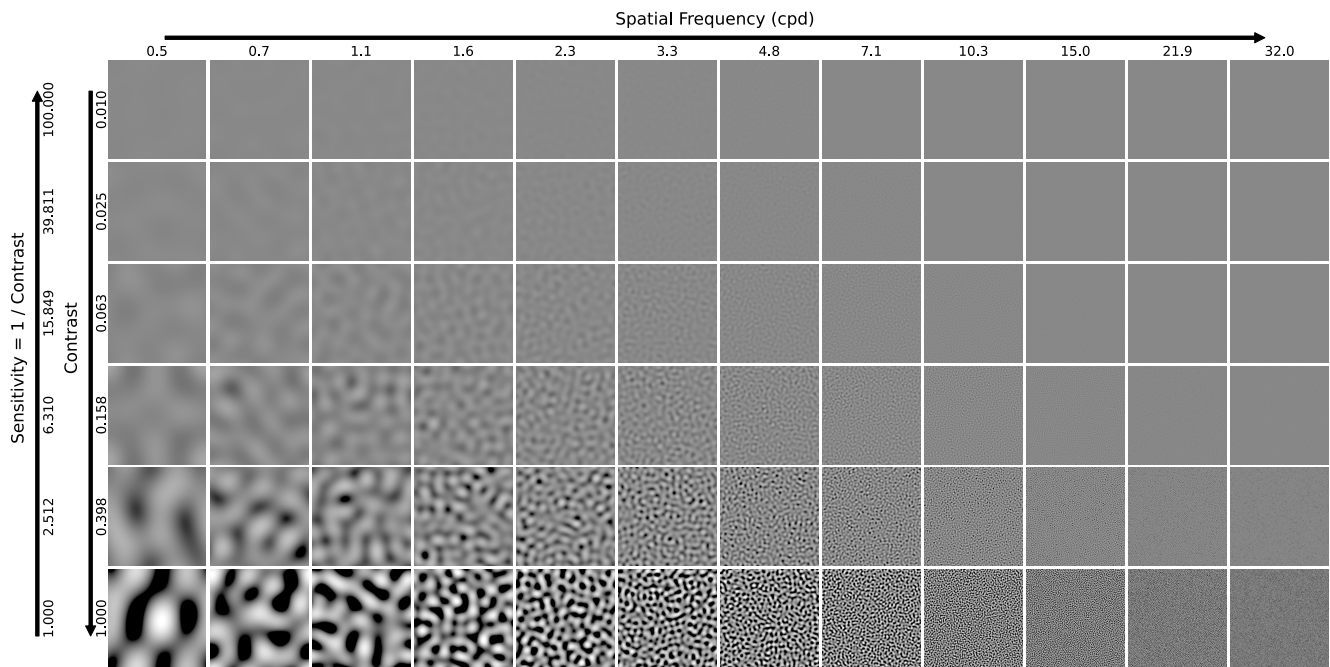


Figure 2. Achromatic band-limited noise signals with different spatial frequencies (x-axis) and contrast (y-axis) used as the test images in the contrast detection tests. Human observers were most sensitive to frequencies in the 2–4 cpd range.

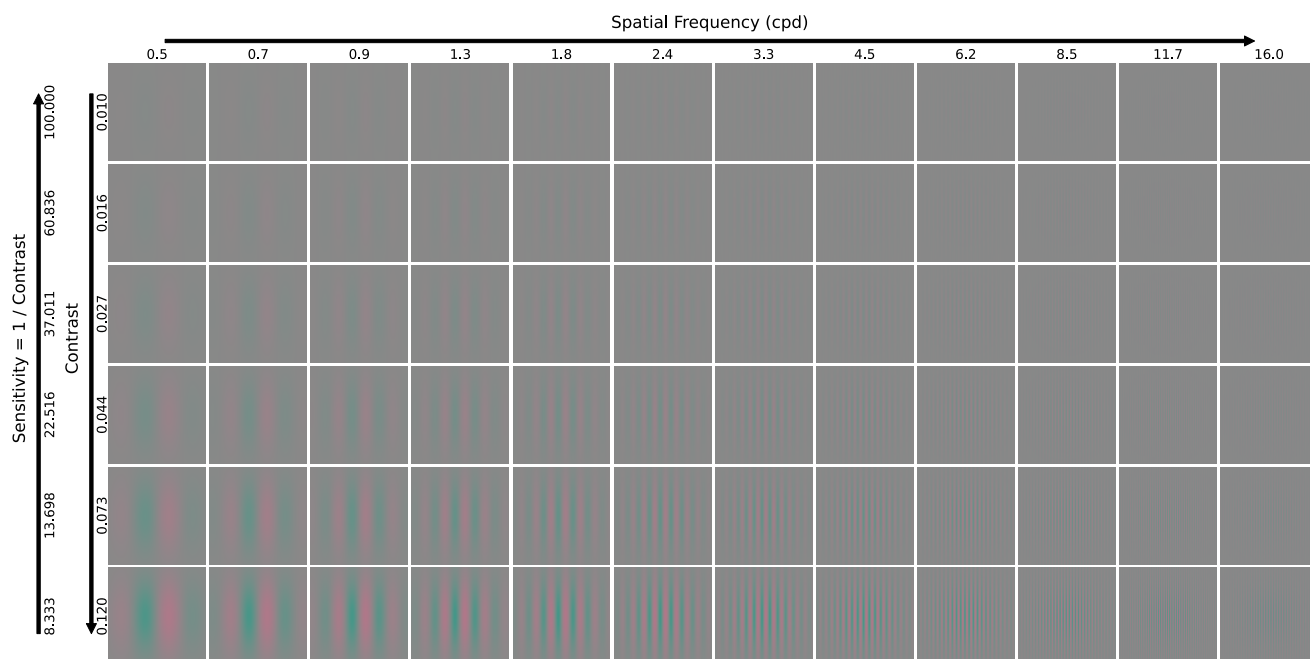


Figure 3. Red-Green (RG) Gabors with different spatial frequencies (x-axis) and contrast (y-axis) used as the test images in the contrast detection tests. Due to gamut limitations, the maximum achievable contrast is capped at 0.2. It was observed that, compared to achromatic Gabor patterns, humans are more sensitive to low frequencies when viewing red-green Gabor patterns.

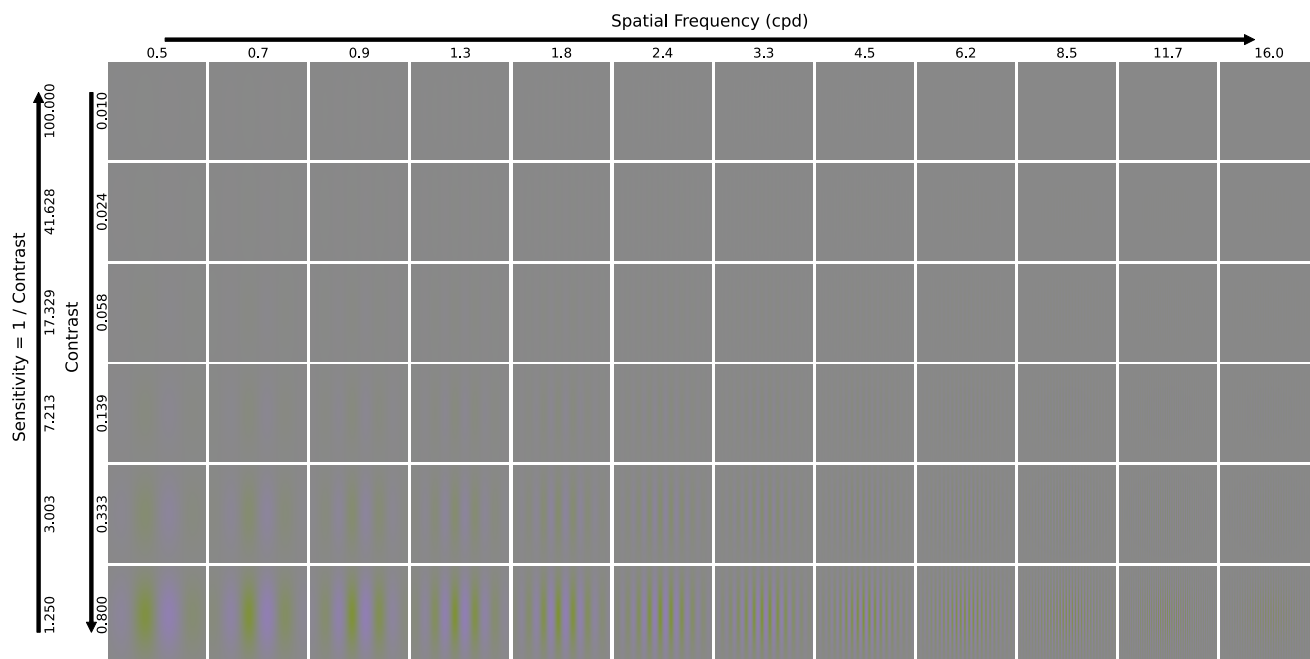


Figure 4. Yellow-Violet (YV) Gabors with different spatial frequencies (x-axis) and contrast (y-axis) used as the test images in the contrast detection tests. Due to gamut limitations, the maximum achievable contrast is capped at 0.2. Similar to RG Gabors, humans are also more sensitive to low-frequency YV Gabors.

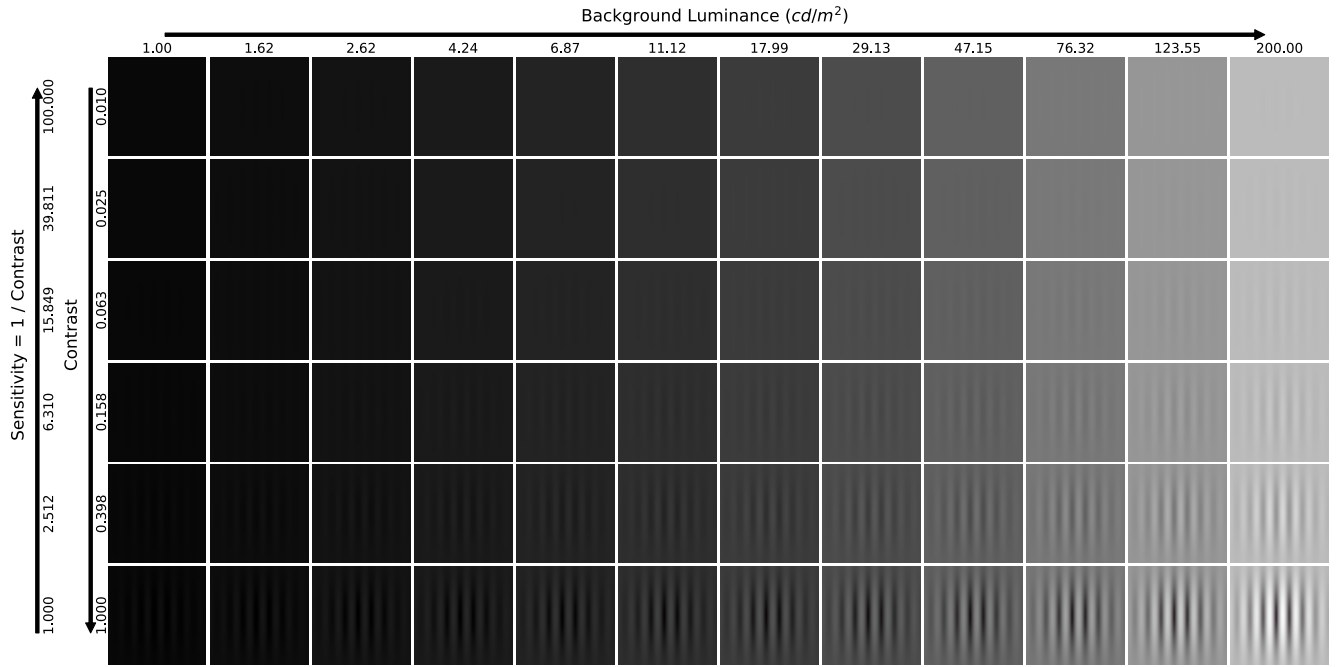


Figure 5. Achromatic Gabor patches with different background luminance (x-axis) and contrasts (y-axis) used as the test images in the contrast detection tests. Note that very low luminance levels cannot be displayed; therefore, a minimum of 1 cd/m^2 is used here. In the experiment, this limitation is not present as we use floating-point inputs.

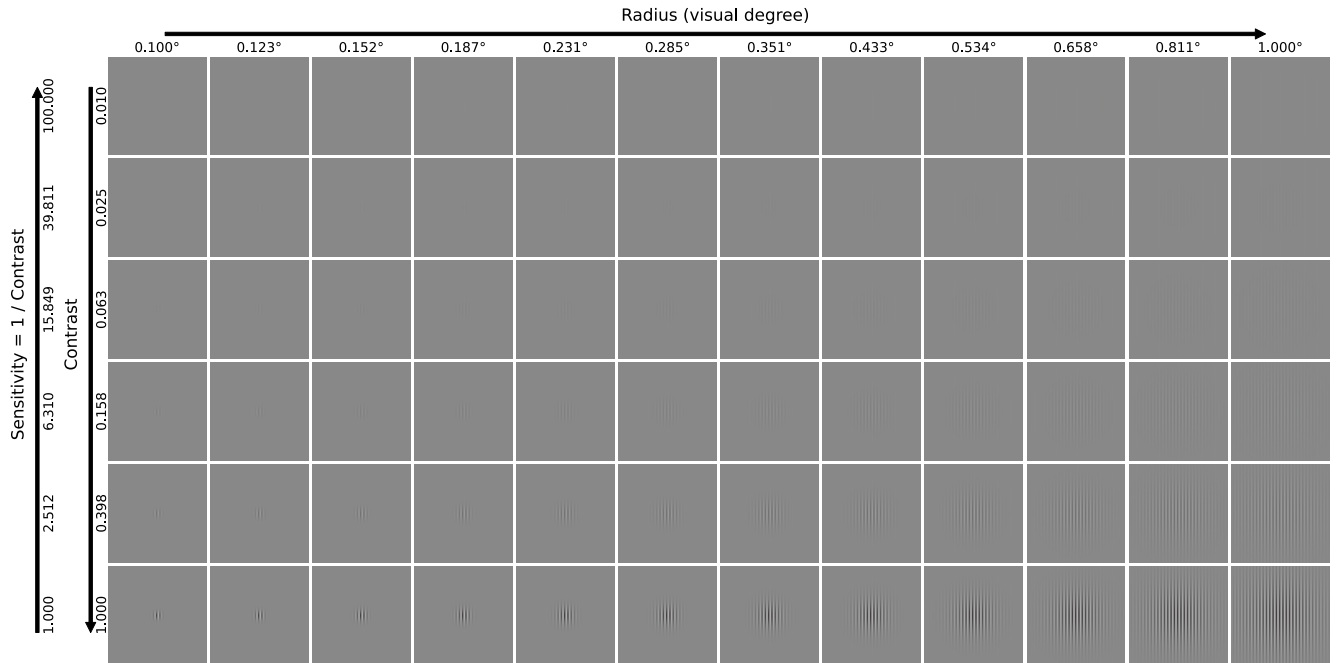


Figure 6. Achromatic Gabor patches with different area (radius) (x-axis) and contrasts (y-axis) used as the test images in the contrast detection tests. In this experiment, we selected a higher spatial frequency (8 cpd); otherwise, it would be impossible to observe a complete Gabor signal within small areas.

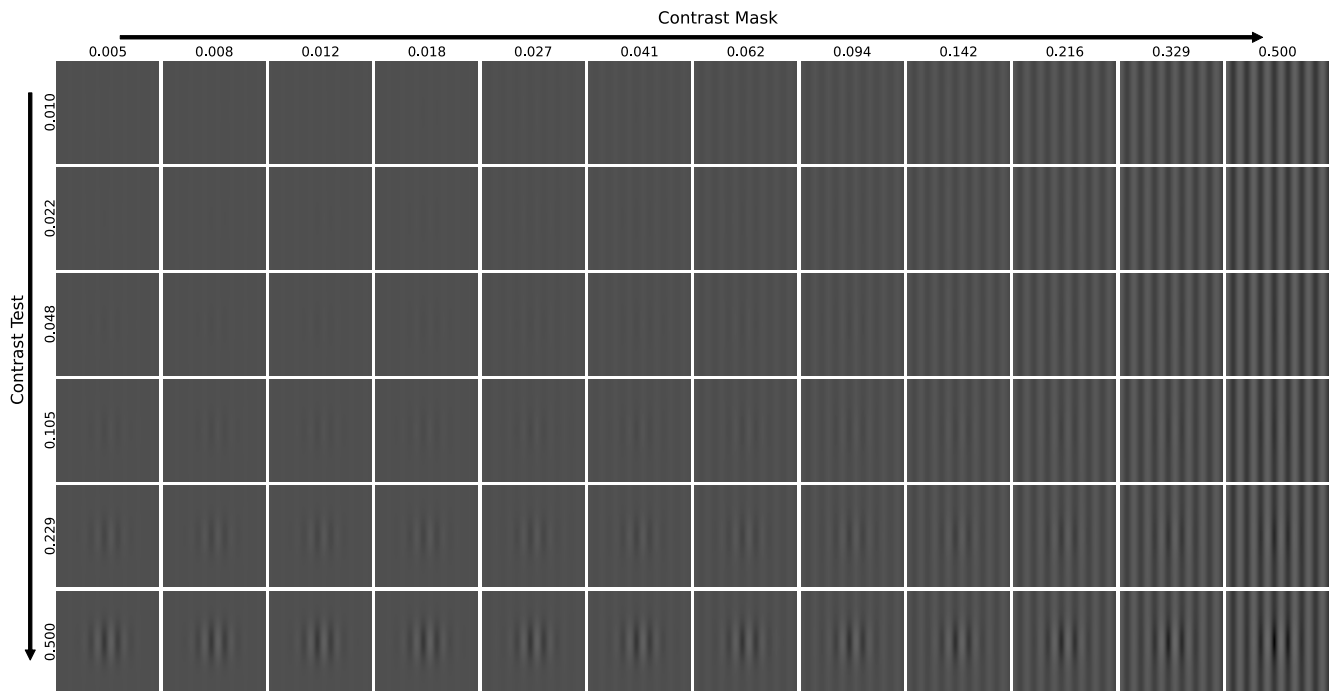


Figure 7. Images from the Phase-Coherent Masking experiment with varying Contrast Mask (x-axis) and Contrast Test (y-axis). The masks are sinusoidal gratings, while the test stimuli are Gabor patterns, set against a background luminance of 32 cd/m^2 .

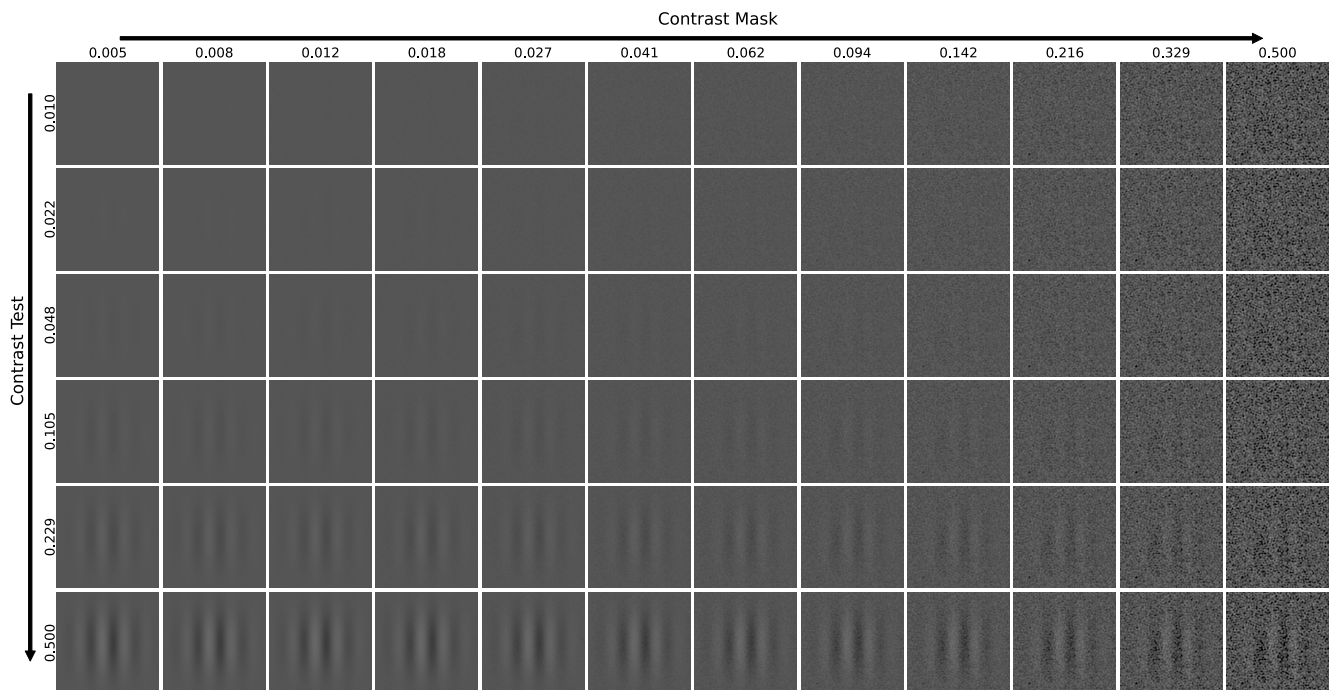


Figure 8. Images from the Phase-Incoherent Masking experiment with varying Contrast Mask (x-axis) and Contrast Test (y-axis). The masks consist of random noise with a frequency spectrum extending up to 12 cpd , while the test stimuli are Gabor patches, presented against a background luminance of 37 cd/m^2 .

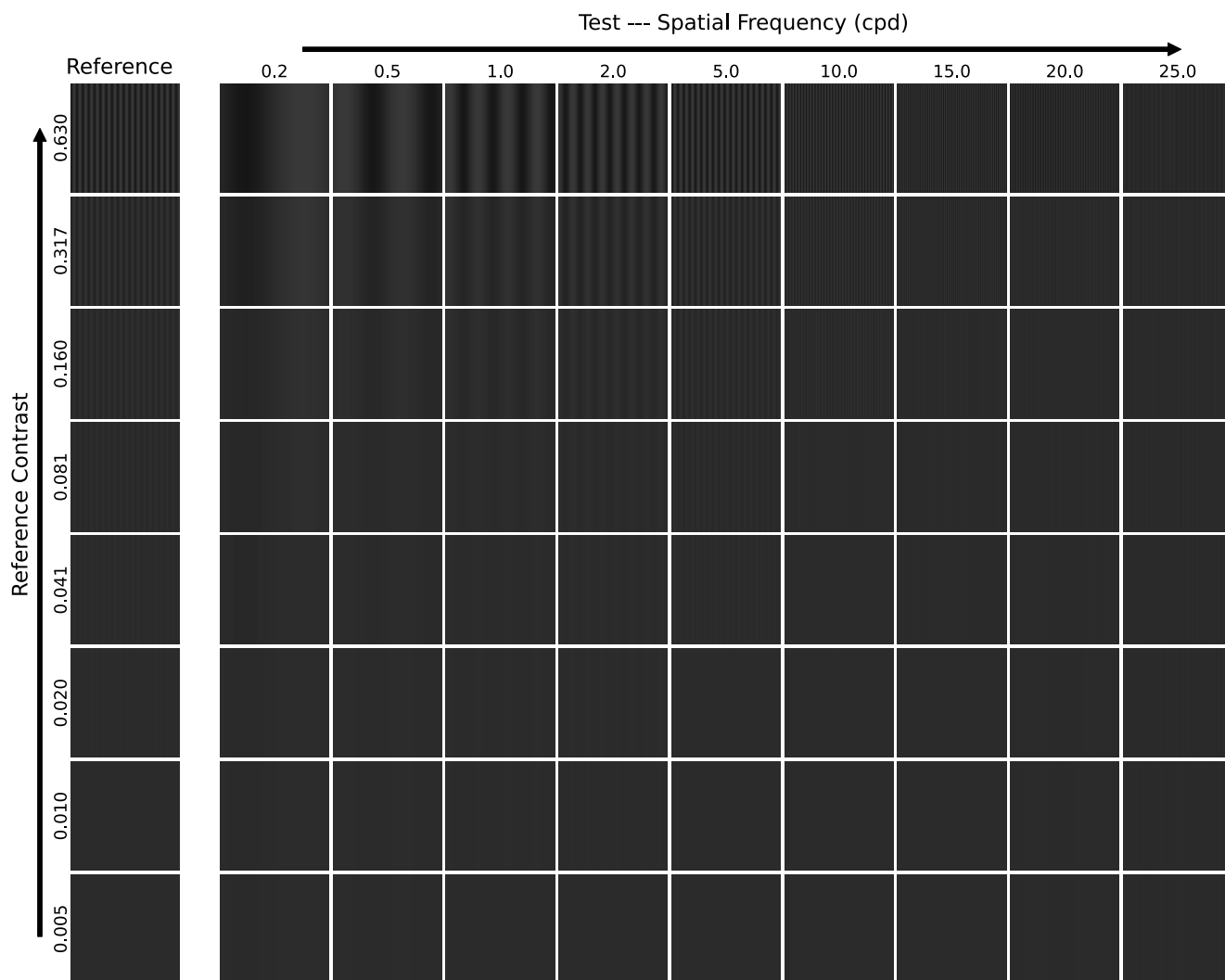


Figure 9. Example images from the Contrast Matching experiment. The first column on the left displays the references, which are sinusoidal gratings at 5 cycles per degree (cpd) with varying reference contrasts. The remaining images on the right are tests with different spatial frequencies matched to the references. Note that the contrast levels of the references, as well as the contrasts and spatial frequencies of the tests, are based on the experimental results in [3]. Following the experimental conditions outlined in [3], the background luminance is set to 10 cd/m^2 .

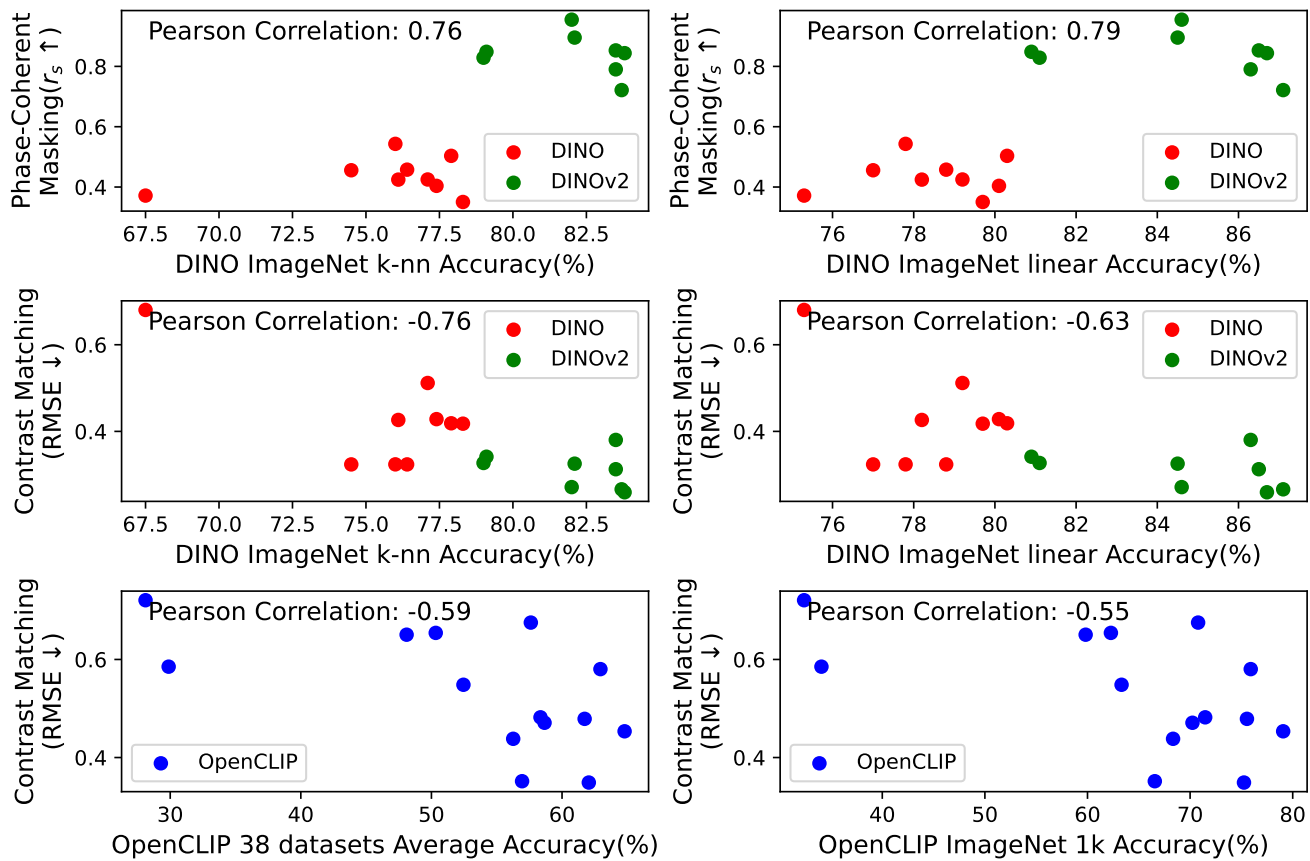


Figure 10. The performance of DINO/DINOv2/OpenCLIP on their classification tasks (from their GitHub repo) shows a potential correlation with the alignment scores in our masking/matching tests.

Table 1. The model alignment scores for all 45 models across nine test types. Spearman’s rank correlation coefficient r_s is used as the evaluation metric for the contrast detection and contrast masking experiments, with higher values (approaching 1) indicating greater similarity between the model and the human visual system. For the contrast matching experiment, Root Mean Square Error (RMSE) is employed as the metric, where lower values (approaching 0) signify a closer match to the human visual system. For each model series, its best score on each test has been highlighted in bold.

Models	Architecture	Training dataset	Spatial Freq. Gabor Ach. $r_s \uparrow$	Spatial Freq. Noise Ach. $r_s \uparrow$	Spatial Freq. Gabor RG $r_s \uparrow$	Spatial Freq. Gabor YV $r_s \uparrow$	Luminance Gabor Ach. $r_s \uparrow$	Area Gabor Ach. $r_s \uparrow$	Phase Coherent Masking $r_s \uparrow$	Phase Incoherent Masking $r_s \uparrow$	Contrast Matching RMSE \downarrow
No Encoder	-	-	0.4688	0.4594	0.5235	0.6582	0.4188	0.8981	0.5057	0.6746	0.2657
DINO	ResNet-50	ImageNet	0.3428	0.2506	0.2795	0.3359	0.5623	0.9610	0.3716	0.8913	0.6803
	ViT-S/16	ImageNet	0.4769	0.4951	0.5316	0.4517	0.4260	0.9686	0.4556	0.8844	0.3238
	ViT-S/8	ImageNet	0.4129	0.4248	0.4211	0.4114	0.4048	0.9358	0.3504	0.6262	0.4178
	ViT-B/16	ImageNet	0.5281	0.4883	0.6148	0.4699	0.4554	0.9584	0.4246	0.6351	0.4264
	ViT-B/8	ImageNet	0.4213	0.4446	0.4398	0.4184	0.3655	0.8929	0.4039	0.4761	0.4283
	Xcit-S-12/16	ImageNet	0.5721	0.4783	0.5312	0.4315	0.4436	0.9418	0.5431	0.7120	0.3239
	Xcit-S-12/8	ImageNet	0.4337	0.3961	0.3862	0.3064	0.4373	0.8409	0.4250	0.7618	0.5117
	Xcit-M-24/16	ImageNet	0.5424	0.4848	0.4651	0.4330	0.4916	0.9408	0.4576	0.6458	0.3238
	Xcit-M-24/8	ImageNet	0.4852	0.3848	0.3742	0.4030	0.5172	0.8149	0.5034	0.8028	0.4187
DINOv2	ViT-S/14	LVD-142M	0.3976	0.4348	0.4114	0.5207	0.4865	0.7071	0.8288	0.9593	0.3271
	ViT-B/14	LVD-142M	0.4286	0.5032	0.4898	0.7148	0.5580	0.9216	0.8955	0.8514	0.3255
	ViT-L/14	LVD-142M	0.5256	0.4843	0.5152	0.4934	0.5363	0.9493	0.7902	0.7968	0.3803
	ViT-g/14	LVD-142M	0.4304	0.5198	0.5277	0.6687	0.4935	0.7856	0.8530	0.8846	0.3127
	ViT-S/14 + reg	LVD-142M	0.4508	0.4484	0.4254	0.4252	0.5201	0.6942	0.8484	0.9254	0.3415
	ViT-B/14 + reg	LVD-142M	0.4408	0.5180	0.4837	0.6488	0.5705	0.4752	0.9549	0.9539	0.2714
	ViT-L/14 + reg	LVD-142M	0.4423	0.5645	0.4799	0.7996	0.5167	0.5289	0.8439	0.9814	0.2595
	ViT-g/14 + reg	LVD-142M	0.4351	0.5732	0.4518	0.6670	0.5168	0.5600	0.7214	0.9295	0.2663
OpenCLIP	ResNet-50	OpenAI	0.3499	0.2903	0.2972	0.3310	0.5703	0.6855	0.4981	0.7349	0.6505
	ResNet-50	YFCC-15M	0.3604	0.2562	0.3206	0.2729	0.5629	0.9671	0.4506	0.6802	0.7208
	ResNet-101	OpenAI	0.4130	0.3001	0.3403	0.3530	0.5070	0.9568	0.4031	0.5697	0.6542
	ResNet-101	YFCC-15M	0.3414	0.2275	0.3620	0.3479	0.4569	0.9046	0.5136	0.8588	0.5853
	ConvNext-B-w	LAION-2B	0.3957	0.2787	0.3772	0.4233	0.4802	0.9590	0.3461	0.6583	0.6752
	ConvNext-B-w	LAION-2B+	0.4649	0.3487	0.4458	0.5648	0.3854	0.7026	0.4887	0.5971	0.4820
	ConvNext-L-d	LAION-2B+	0.3419	0.2835	0.3911	0.6443	0.1530	0.7690	0.4919	0.4983	0.5803
	ConvNext-XXL	LAION-2B+	0.4136	0.3642	0.3890	0.6212	0.1112	0.8422	0.4699	0.4586	0.4535
	ViT-B/32	OpenAI	0.4132	0.5696	0.3748	0.7197	0.1772	0.9434	0.4047	0.7736	0.5484
	ViT-B/32	LAION-2B	0.5108	0.6556	0.3146	0.4108	0.3063	0.8673	0.8837	0.9620	0.3517
	ViT-B/16	OpenAI	0.4798	0.5429	0.4138	0.7141	0.2655	0.7887	0.4321	0.6451	0.4382
	ViT-B/16	LAION-2B	0.4654	0.5740	0.4144	0.6922	0.4242	0.7869	0.5235	0.7725	0.4709
	ViT-L/14	OpenAI	0.4625	0.5026	0.4357	0.6229	0.4585	0.8892	0.5496	0.7050	0.4789
	ViT-L/14	LAION-2B	0.5917	0.5240	0.4801	0.7338	0.3084	0.6430	0.7678	0.8799	0.3490
SAM	ViT-B-SAM	SA-1B	0.3545	0.3287	0.3577	0.3617	0.3074	0.9714	0.4144	0.4353	0.5877
	ViT-L-SAM	SA-1B	0.3061	0.2769	0.3160	0.3140	0.3090	0.9598	0.4077	0.3787	0.6354
	ViT-H-SAM	SA-1B	0.3651	0.3234	0.3316	0.3863	0.5243	0.9533	0.3983	0.5105	0.5489
SAM-2	SAM2.1-hiera-tiny	SA-V	0.4058	0.3483	0.4315	0.4693	0.4966	0.9660	0.4137	0.4646	0.4805
	SAM2.1-hiera-S	SA-V	0.4544	0.3705	0.3936	0.4608	0.5389	0.9533	0.398	0.4941	0.4472
	SAM2.1-hiera-B+	SA-V	0.3728	0.3195	0.4949	0.4993	0.5396	0.9296	0.4062	0.4882	0.4852
	SAM2.1-hiera-L	SA-V	0.3872	0.3259	0.3695	0.4935	0.5631	0.9431	0.4613	0.7361	0.4686
MAE	ViT-B-MAE	ImageNet	0.4471	0.4903	0.4410	0.5008	0.5812	0.9036	0.5446	0.7277	0.4223
	ViT-L-MAE	ImageNet	0.4284	0.4560	0.4126	0.4803	0.5647	0.8874	0.6849	0.7043	0.4344
	ViT-H-MAE	ImageNet	0.4250	0.3969	0.3995	0.4737	0.6335	0.6466	0.5003	0.6088	0.4964
SD-VAE	SD-v1-5	LAION-5B	0.3527	0.4226	0.8447	0.8051	0.4993	0.8402	0.5394	0.5579	0.4177
	SD-xl-base-1.0	LAION-5B	0.2465	0.1662	0.3811	0.3132	0.4273	0.3561	0.4996	0.4962	0.6727
ColorVideoVDP	HVS-based	XR-DAVID+	0.5545	0.7817	0.7455	0.9339	0.9020	0.8937	0.7418	0.7626	0.2604