Appendix Overview

A Discussion **12**
- A.1. Limitations . . . 12
- A.2. Broader Impact . . . 12
- A.3. Crowdsourced Study Design . . . 12

B Ablation Studies **13**
- B.1. Choice of γ . . . 13
- B.2. Sensitivity to β in BRAgg (Uniform Prior) . . . 13
- B.3. Activation-Guided Importance Sampling . . . 14
- B.4. Error Model for Rating Aggregation . . . 15
- B.5. SigLIP Only Evaluation Accuracy . . . 16
- B.6. Compositional/Complex Explanations . . . 16
- B.7. Model Ablation . . . 17
- B.8. Effect of Sampling Model Accuracy on MG-IS Performance . . . 18
- B.9. Comparison with Downstream Task Evaluation . . . 18

C Additional Details **20**
- C.1. MTurk Experimental Details: . . . 20
- C.2. Detailed MTurk results . . . 20
- C.3. Experimental Details . . . 25
- C.4. Automated Evaluation Details . . . 25
- C.5. Theorem 1 . . . 26
- C.6. Compute details . . . 27
- C.7. Inter-Annotator Consistency . . . 27

A. Discussion

A.1. Limitations

Due to time and budget constraints, our crowdsourced evaluation focuses on comparing descriptions on a few models/layers e.g. later layer neurons of ImageNet trained models. While we mostly find similar trends between the ResNet and ViT models, different description methods may have advantage on different types of neurons. For example, it is possible that Network Dissection based methods would perform better than they did on our current evaluation if we focus on lower layer neurons or on models trained on Places365 [34], as the labels in the Broden dataset are more suitable for these tasks. Similarly, we think LLM based methods [1, 29] might perform better when describing neurons of a sparse autoencoder [6] as these are more monosemantic and can be better described by highly activating inputs only.

Second, our crowdsourced evaluation relies on Amazon Mechanical Turk workers, who are not experts and often make errors in labeling. While we introduced principled measures to estimate the errors as well as mitigations for error, we cannot improve their domain knowledge, which means a crowdsourced evaluation might favor simpler descriptions over more complex concepts requiring domain knowledge. To compare more complex descriptions or neuron descriptions in a specific domain it may be necessary to recruit domain expert raters.

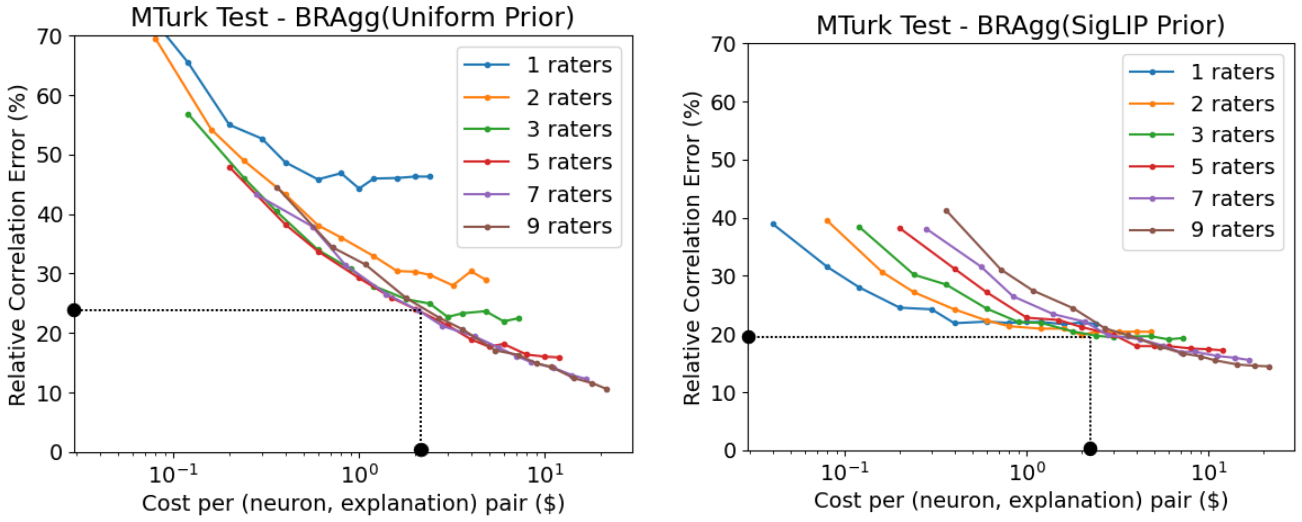
Similarly, our Mechanical Turk validation in Section 4 assumes that the ImageNet labels for these images are correct and unambiguous, but in reality the labels contain some errors and ambiguity. To mitigate this, we focused on ImageNet classes we deemed to have less ambiguity in our experiment, but some ambiguity still remains as discussed in Appendix C.7. Using an alternate rating setup that allows for rating indeterminancy such as [12] could potentially improve this validation.

Finally, our evaluation is focused on evaluating *input-based* neuron explanations that aim to explain the "input $\rightarrow$ neuron activation" function. Some recent work such as [10, 13] instead focus on *output-based* neuron explanations that aim to explain the "neuron activation $\rightarrow$ model output" relationship. Rigorously evaluating these *output-based* explanations will require different methodology and is an interesting problem for future work.

A.2. Broader Impact

This paper is focused on better understanding neural networks via interpretability, and as such we expect its impact to be largely positive as better understanding of neural networks can help us for example identify failure modes before deploying and enable better control of models. As our focus is in particular on rigorous evaluation of neuron explanations, this can help avoid interpretability illusions or users over-relying on unreliable explanations.

A.3. Crowdsourced Study Design



(a) Aggregation with Method 3a: BRAgg - Uniform Prior.

(b) Aggregation with Method 3b: BRAgg - SigLIP prior.

Figure 6. Comparing rating aggregation strategies in Setting 2 (Testing on MTurk) described in Sec 4. The x-axis represents the cost of evaluations as number of inputs per neuron multiplied by number of raters per input, times cost per input. The lowest curve represents the optimal number of raters for a particular budget.

In this section we discuss how our methodology testing/validation settings described in Section 4 can be used to help design the finer parameters of our study such as how many raters to use per input. In our current MTurk setup, the cost of obtaining a rating for a single image by a single rater is $\frac{\$0.06}{15} = \$0.004$ as we show 15 inputs per task. We can then plot the expected error as a function of evaluation cost by plotting $n_{inputs} \cdot n_{raters} \cdot \$0.004$ on the x-axis.

In Figure 6, we plot the cost vs expected error rate for different numbers of raters we have per input. As we can see, the optimal number of raters per input depends on your budget, with smaller number of raters working better with smaller budgets, while for a larger budget it is optimal to use more raters per input.

Given our budget, we are aiming for a cost of around \$2.2 per (neuron, explanation pair), and we can see the optimal number of raters and expected error rate for this budget following the black lines on our plot. Since we wish to get good results using both aggregation methods, we can see 3 raters (green line) is among the best for both methods, we choose to use 3 raters and 180 inputs per neuron for the crowd-sourced study. This gives us an expected correlation error of around 25% with Bayes - Uniform Prior and around 20% with Bayes - SigLIP prior based on the MTurk test.

B. Ablation Studies

B.1. Choice of γ

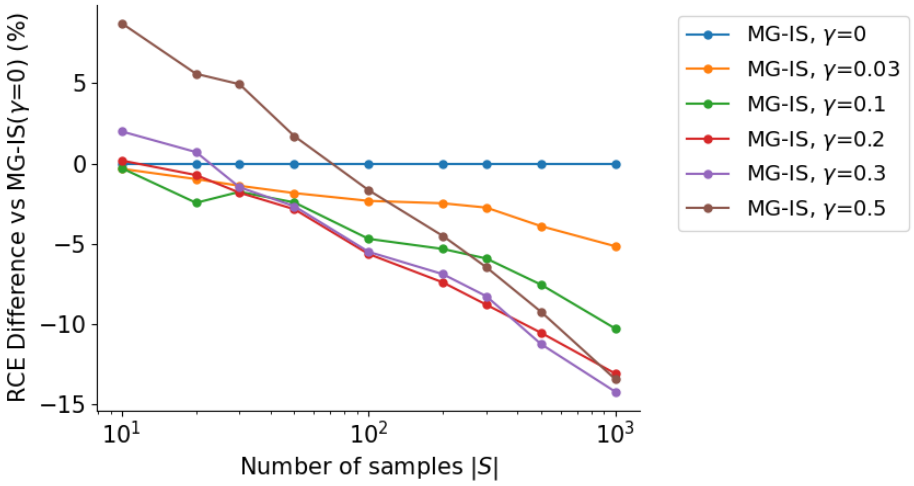


Figure 7. The Effect of the γ hyperparameter on Importance sampling performance.

Figure 7 shows the effects of the hyperparameter γ (Eq. 6) on importance sampling performance on our simulated setup with no label noise. For visual clarity, we report the difference compared to baseline with $\gamma = 0$ instead of absolute numbers, so for example reducing RCE from 25% to 22.5% would show as -10% on this plot. Overall we can see our method is not very sensitive to this choice, but using a reasonable γ does reduce error by around 10%, with more relative difference on larger sample sizes. We use $\gamma = 0.2$ as it performs the best on sample sizes around 200 that we use in our large study.

B.2. Sensitivity to β in BRAgg (Uniform Prior)

Our Rating Aggregation Method 3a Bayes - Uniform Sampling described in Section 3.2 depends on the β hyperparameter to select the uniform prior. In this section, we conducted a test on Setting 2 (MTurk test, using MG-IS sampling) comparing different values for β. As shown in Fig 8, the exact choice of β has little effect on the correlation error, with values between 0.02 and 0.2 performing well. In our experiments we used $\beta = 0.05$ which performs well overall. We think the small sensitivity is mostly because changing prior has a relatively linear effect on predicted c_t, and since correlation coefficient normalizes the c_t the scale of c_t does not change the correlation. Each datapoint in Figure 8 is the average over 30 samples with different random seeds.

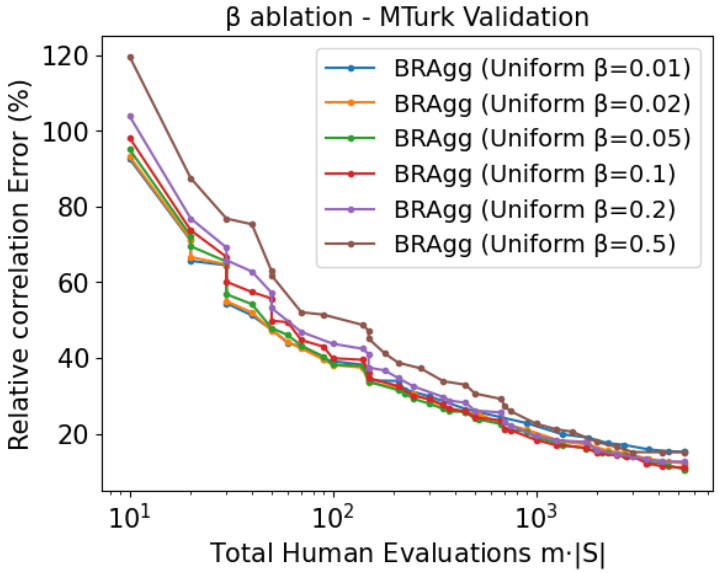


Figure 8. Comparing the error rate of using different β values for BRAgg(Uniform prior).

B.3. Activation-Guided Importance Sampling

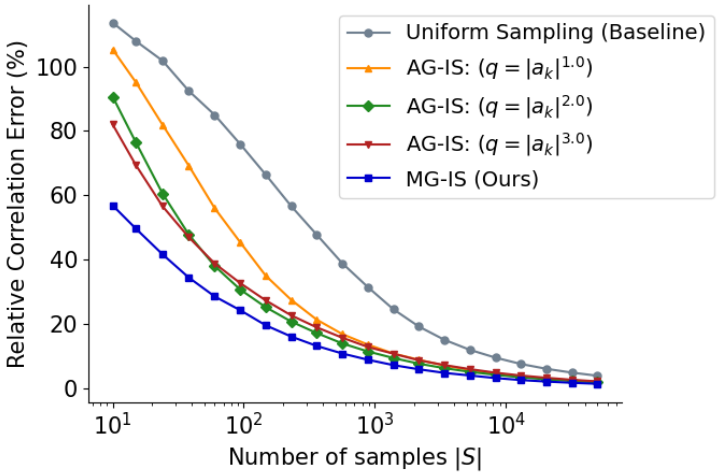


Figure 9. Comparing different sampling strategies on simulated data with no label noise. We can see that importance sampling without model guidance can still perform quite well and deliver significant gains over uniform sampling.

In addition to Model-Guided Importance Sampling (MG-IS), we also tested importance sampling guided purely by neuron activations, which we call activation guided importance sampling (AG-IS). This could be useful for cases where no cheap model is available to estimate concept presence. In particular we tested with importance sampling functions of the following form:

$$q^{act}(x_i) = \frac{|[\bar{a}_k]_i|^\alpha}{\sum_{x_i \in \mathcal{D}} |[\bar{a}_k]_i|^\alpha} \tag{16}$$

$$q^{\text{AG-IS}}(x) = (1-\gamma)q^{act}(x) + \gamma p(x) \tag{17}$$

As shown in Figure 9, we can see very significant gains over uniform sampling, with the best performing variant being

$\alpha = 2$, but overall performance is not as good as our model-guided importance sampling. We used $\gamma = 0.2$ for these experiments.

B.4. Error Model for Rating Aggregation

For our main results, our Bayes modeling uses a uniform error model, i.e. we assume all have concepts and inputs share the same error rate η as described in Section 3.2. However, this is not always a realistic assumption, as some inputs and concept are harder/more ambiguous, leading to higher error rates. To account for this, we also experimented using an alternative error model where the error rate η_{ti} is different for each pair of input i and concept t.

In this model $\mathbb{P}(r^j_{ti} = c^*_{ti}) = 1 - \eta_{ti}$, where $\eta_{ti} \sim \text{Beta}(0.75, 2.5)$. The hyperparameters $\alpha = 0.75$ and $\beta = 2.5$ were selected to maximize the probability of the MTurk user ratings we observed in Setting 2.

To do Bayesian inference with this model, we do inference on both c^*_{ti} and η_{ti} and report $[c_t]_i = \int_{\eta_{ti}} \mathbb{P}(C_{ti}, \eta_{ti}|R_{ti})d\eta_{ti}$. Applying the Bayes rule $\mathbb{P}(C_{ti}, \eta_{ti}|R_{ti}) = \mathbb{P}(R_{ti}|C_{ti}, \eta_{ti})\mathbb{P}(C_{ti}, \eta_{ti})/\mathbb{P}(R_{ti})$ we get:

$$[c_t]_i = \frac{\int_{\eta_{ti}} \mathbb{P}(R_{ti}|C_{ti}, \eta_{ti}) \cdot \mathbb{P}(C_{ti})\mathbb{P}(\eta_{ti})d\eta_{ti}}{\int_{\eta_{ti}} \mathbb{P}(R_{ti}|C_{ti}, \eta_{ti}) \cdot \mathbb{P}(C_{ti})\mathbb{P}(\eta_{ti}) + \mathbb{P}(R_{ti}|\neg C_{ti}, \eta_{ti})\mathbb{P}(\neg C_{ti})\mathbb{P}(\eta_{ti})d\eta_{ti}} \tag{18}$$

Since $\mathbb{P}(\eta_{ti})$ and $\mathbb{P}(C_{ti})$ are independent. In practice calculating exact solution for the integral is difficult, so we instead approximate it with 100 discrete bins.

Likelihood: $\mathbb{P}(R_{ti}|C_{ti}, \eta_{ti})$.

With $\alpha_{ti} = \sum_{j=1}^{m} r^j_{ti}$, we obtain the likelihood in below equations:

$$\mathbb{P}(R_{ti}|C_{ti}, \eta_{ti}) = (1 - \eta_{ti})^{\alpha_{ti}}(\eta_{ti})^{(m-\alpha_{ti})} \tag{19}$$

$$\mathbb{P}(R_{ti}|\neg C_{ti}, \eta_{ti}) = (\eta_{ti})^{\alpha_{ti}}(1 - \eta_{ti})^{(m-\alpha_{ti})} \tag{20}$$

Priors: For $\mathbb{P}(C_{ti})$ we use the same priors as in Section 3.2 (uniform or SigLIP prior). For $\mathbb{P}(\eta_{ti})$, we use the Beta distribution:

$$\mathbb{P}(\eta_{ti}) \propto \eta^{\alpha}_{ti} \cdot (1 - \eta_{ti})^{\beta} \tag{21}$$

The hyperparameters $\alpha = 0.75$ and $\beta = 2.5$ were selected to maximize the probability of the MTurk user ratings we observed.

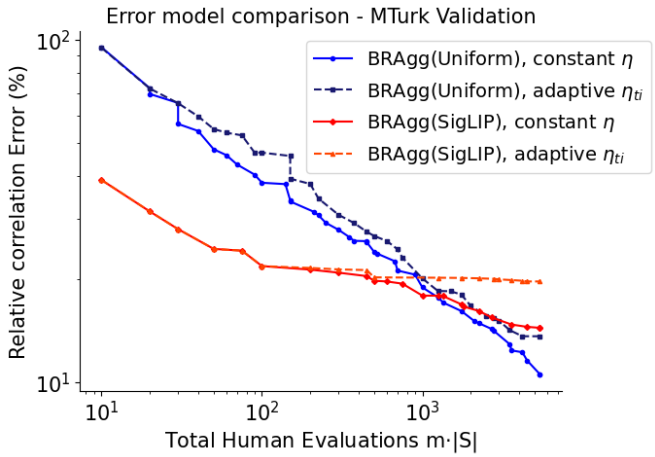


Figure 10. Comparing different error models for rating aggregation. Note log-scale on both axes.

Results As shown in Figure 10, the performance between different error models is similar, but overall the simplified uniform error model (solid lines) leads to slightly better results, so we use the simple error model for our main results. We are not fully certain why the simple model performs better, but one possibility is that even 9 ratings per user is not enough for the Bayes model to accurately infer per-input error rates.

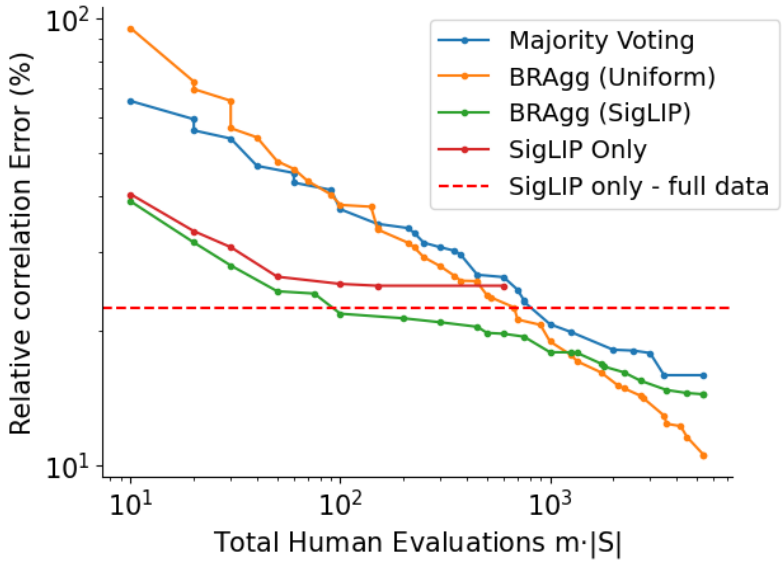


Figure 11. Comparison the error of different rating aggregation strategies and only relying on SigLIP instead of conducting a human study. Using our Setting 2 results from MTurk with MG-IS sampling.

B.5. SigLIP Only Evaluation Accuracy

Figure 11 shows comparison of different rating aggregation strategies vs a pure automated evaluation only relying on SigLIP scores. We can see that the optimal rating strategy depends on your budget. A fully automated evaluation provides similar accuracy as a hybrid evaluation (BRAgg(SigLIP)) with 100 samples per (neuron, explanation) pair, or a pure human evaluation with 800 samples per neuron. This implies that the optimal study strategy depends on your budget: If you cannot afford more than 100 human evaluations per neuron, it is best to rely on automated evaluation. For budgets between 100-1000 samples per neuron, a hybrid evaluation (BRAgg(SigLIP)) performs the best. For sufficiently large budget, i.e. >1000 evaluations per neuron, pure human evaluation (BRAgg(Uniform)) performs the best.

B.6. Compositional/Complex Explanations

A more complex explanation is typically more accurate, but it is harder to understand, and more expensive to evaluate as it requires labels for each of the concepts involved in the explanation. We use the length l to indicate explanation complexity, where l is the number of unique concepts in the explanation, e.g. "cat" OR "dog" would have length 2.

Complex Exp. (l>1)	Comp Exp [21] l=3	INVERT [7] l=3	CCE [19], l= 5	CCE [19], l=15	LE(label) [23] l=4.37/1.97	LE(SigLIP) [23] l=4.66/1.82
RN-50 (Layer4)	0.1399 ± 0.002	0.2341 ± 0.002	0.0993 ± 0.002	0.1510 ± 0.003	0.2924 ± 0.002	**0.3772 ± 0.002**
ViT-B-16 (Layer11 MLP)	0.0468 ± 0.002	0.1101 ± 0.003	0.0534 ± 0.003	0.0570 ± 0.006	0.3243 ± 0.005	**0.3489 ± 0.005**

Table 2. SigLIP based simulation with correlation scoring comparing complex explanations ($l > 1$). We can see Linear Explanation (LE) overall performs the best.

For fairness we split the explanations into two groups, simple explanations where $l = 1$ and complex explanations with $l > 1$. Table 1 in main text shows comparison of different simple explanations, while Table 2 compares complex explanation methods. For complex explanations, we can see Linear Explanations reaches the highest correlations with most methods having higher scores than their simple counterpart, highlighting the value of added complexity.

B.7. Model Ablation

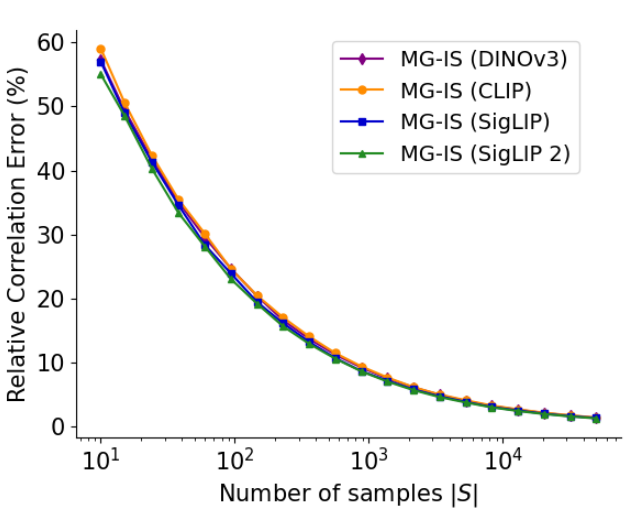
In our main results we used the SigLIP [33] ViT-L-16-384 model both as the guide for importance sampling (MG-IS), and to provide the prior for our Bayes rating aggregation. For the Automated Evaluation described in Section 5.1 we used the SigLIP ViT-SO400M-14-384 model.

The purpose of this section is to analyze the sensitivity of our results to the model choice. In particular, we test replacing the SigLIP ViT-L-16-384 model for sampling and rating aggregation with the following models:

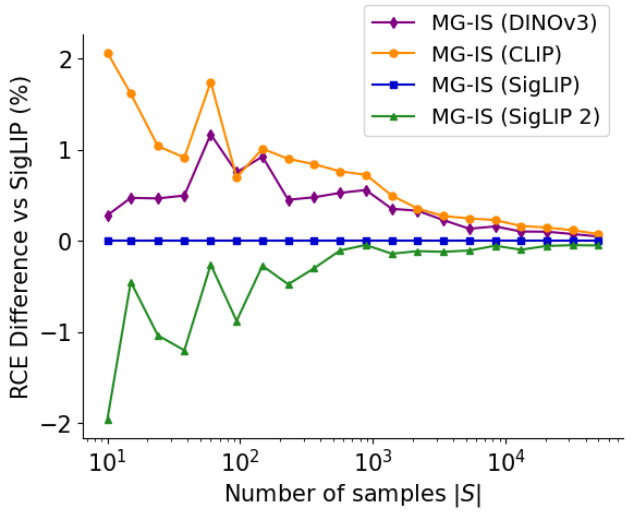
1. CLIP [26] ViT-L14-336, which is an older model with slightly weaker benchmark performance.
2. SigLIP2 [31] ViT-gopt-16-384, which is a bigger and more recent model with slightly better benchmark scores.
3. DINOv3 [30] - ViT-L-16 with pretrained head for zero-shot classification (*dinov3_vitl16_dinotxt_tet1280d20h24l*).

For all models we used the weights and implementation from the open_clip package [17].

Results - Sampling Model: Figure 12 shows a comparison of different models for guiding sampling. We can see better models improve results but the differences are quite small. On average using CLIP instead of SigLIP leads to around 5 % (relative) increase in error with the same number of samples, while using SigLIP 2 leads to around 2 % (relative) reduction in error. DINOv3 performs worse than SigLIP but better than original CLIP on this task.



(a) Comparing different models for guiding sampling in our simulated test (Setting 1). Stronger models improve sampling but the differences are small.



(b) Same plot, but zooming in on the difference in RCE using SigLIP vs. other models reveals the relative ranking.

Figure 12. Model ablation for guided sampling. (a) All models converge similarly; (b) the per-model difference from SigLIP shows the relative ranking. Note much smaller y-axis range in (b).

Results - Model for Bayes Prior: When changing the model used for creating the prior in BRAgg, we observe similar trends, with larger models generally performing better than smaller ones, but the results are a bit less clear and depend on the number of evaluations. Figure 13 shows the results of using different models as the Bayes prior We notice a somewhat bigger performance difference in this setting, with for SigLIP 2 reducing RCE at (3 raters per input and 200 inputs) from 19.70%(of SigLIP) to 17.95% for around 10% relative reduction in error. Interestingly, using CLIP as the prior performs worse with small sample sizes but beats SigLIP with large evaluation budget. Inspired by these results, we also analyze our main experiment results using SigLIP 2 instead of SigLIP as the Bayes prior and report the results in Table 3. Overall we can see the results are very similar to our results using SigLIP as the prior reported in Table 6. The only difference we notice is the correlation scores are slightly higher for all methods on ViT-B-16 neurons. This highlights our conclusions regarding the relative performance of different explanation methods are not sensitive to model choice.

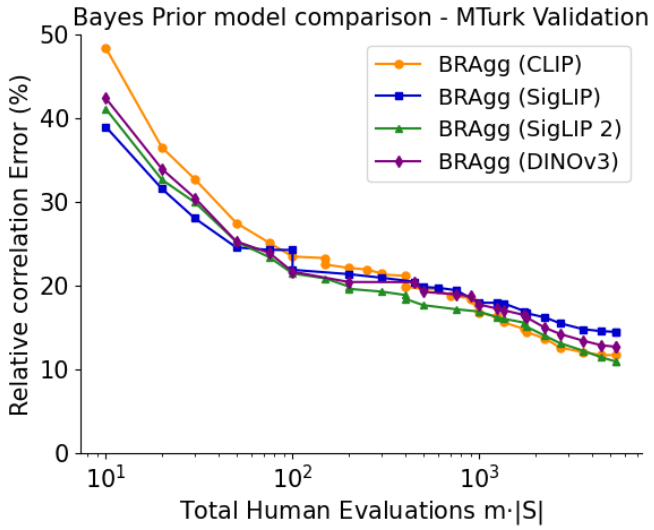


Figure 13. Comparing different choices for the Bayes prior. We can see SigLIP2 performs the best overall.

Model	Aggregation	CLIP-Dissect	INVERT l=1	DnD	MAIA	LE(SigLIP) l=1
ResNet-50 (Layer4)	BRAgg (SigLIP 2 Prior)	0.1392 ± 0.008	0.1490 ± 0.005	0.1242 ± 0.008	0.1072 ± 0.008	**0.1786 ± 0.009**
ViT-B-16 (Layer11 mlp)	BRAgg (SigLIP 2 Prior)	0.1714 ± 0.019	0.1250 ± 0.021	0.1039 ± 0.016	0.1355 ± 0.019	**0.2591 ± 0.023**
Average	BRAgg (SigLIP 2 Prior)	0.1553	0.1370	0.1141	0.1201	**0.2189**

Table 3. Large scale crowdsourced study results using SigLIP 2 as the prior for BRAgg.

AUC Range: (#neurons)	0-0.6 (217)	0.6-0.8 (209)	0.8-0.95 (204)	0.95-0.99 (359)	0.99-1 (1059)
Uniform	0.271	0.408	0.447	0.513	0.677
MG-IS	**0.268**	**0.351**	**0.299**	**0.210**	**0.089**

Table 4. Comparing Sampling performance based on SigLIP AUC in detecting the concept. We report average RCE.

B.8. Effect of Sampling Model Accuracy on MG-IS Performance

To study how the performance of our sampling model (SigLIP) affects the efficiency of MG-IS, we repeated the experiment in Sec 3.1.2 and split neurons into groups based on how accurately SigLIP can detect their concepts. Table 4 shows the RCE of different neuron groups when sampling 180 inputs using different methods. As we can see, even on neurons where SigLIP accuracy is close to random chance (AUC<0.6), our MG-IS still produces slightly lower error than pure uniform sampling baseline. This is likely because MG-IS samples from neurons with a high product $\hat{a}_k \cdot \hat{c}_t$, so even when $\hat{c}_t$ is inaccurate, oversampling high activations can still reduce variance. For concepts that are both rare and low AUC, a larger sample size is needed to estimate correlation with low variance. However as seen in Table 4, MG-IS still outperforms the uniform sampling baseline in those cases. Conversely, on neurons where SigLIP accuracy is high (most neurons) MG-IS reduces error $\sim 7\times$.

B.9. Comparison with Downstream Task Evaluation

To test whether our correlation evaluations give signal on downstream task performance of neuron explanation methods, we ran the evaluation introduced by [1](Appx. A.6), where they measure whether neuron explanation methods can be used to find neurons that make good classifiers for unseen classes. In particular, we tested whether we can find neurons in ResNet-50 layer4 (trained on ImageNet) that make good classifier for CIFAR or PLaces365 classes. For each class, we find the neuron whose

Target Classes	CD	DnD	INVERT	LE
CIFAR10	0.763	0.697	0.751	**0.787**
CIFAR100	0.727	0.723	0.733	**0.746**
Places365	0.726	0.692	0.679	**0.727**

Table 5. Using generated explanations to find neurons in RN-50 that work as classifiers for unseen classes. We report Average AUC.

explanation has the highest text embedding similarity to the class name, and then measure it's performance as a classifier on the new dataset using AUC. We report the average AUC across classes in Table 5. The intuition is that a good explanation method should be able to identify better neurons for each task. Overall we see that the results of this experiment align with our results and the methods that scored highest in our correlation evaluation (LE > CD, INVERT > DnD) also find the best classifier neurons.

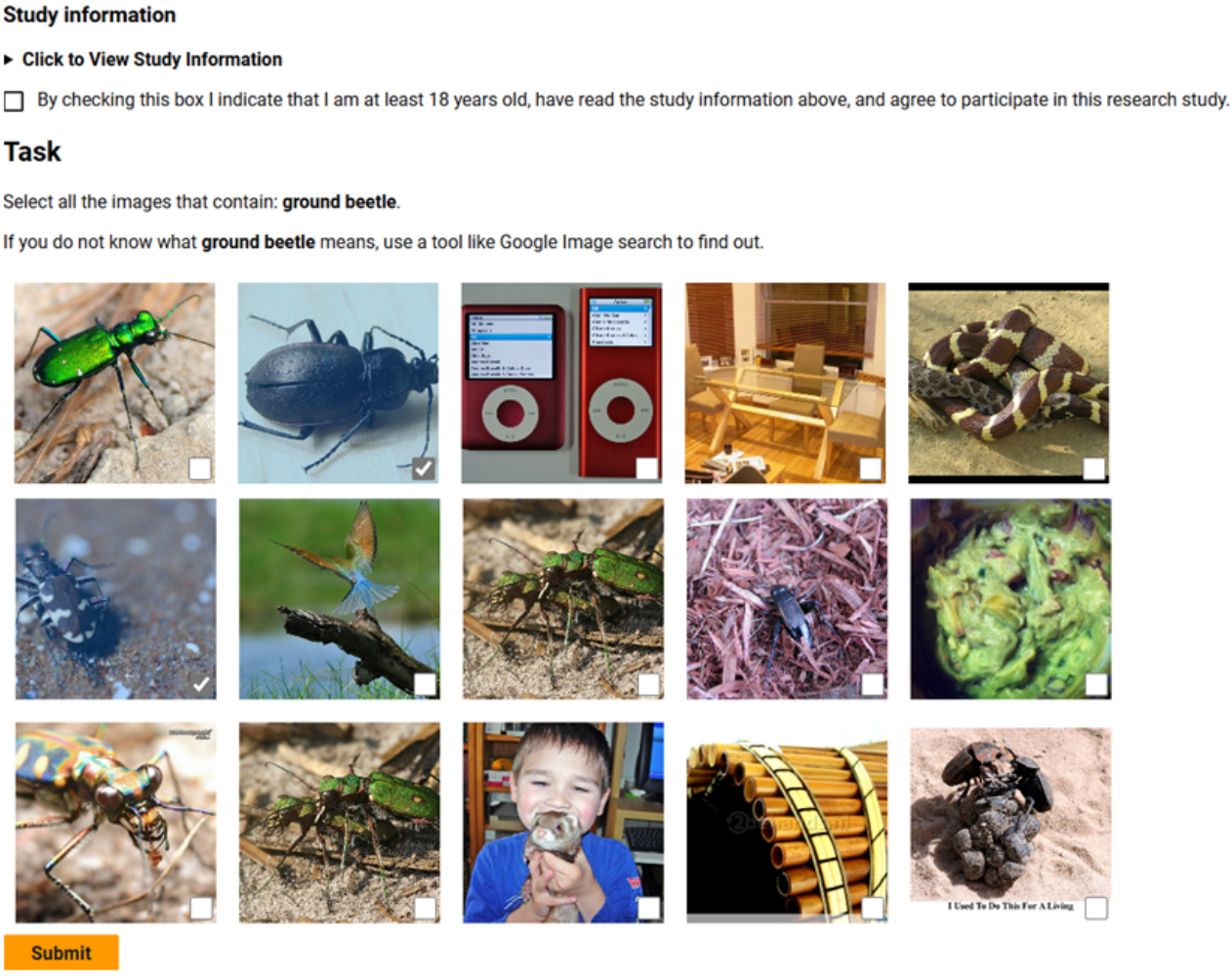


Figure 14. Our user study interface.

C. Additional Details

C.1. MTurk Experimental Details:

Figure 14 showcases the full user interface displayed to the raters. We selected raters based in the US with over 10,000 tasks approved and $> 98\%$ task approval rate. Each rater was paid \$0.05 per task (and we paid another \$0.01 per task in fees to MTurk), which takes around 15 seconds to complete based on our testing, for an estimated \$12/hr earnings.

C.2. Detailed MTurk results

Table 6 shows more detailed results of our crowdsourced evaluation, including multiple rating aggregation methods. Comparing different rating aggregation methods, we can see that the correlation scores themselves vary a lot depending on the aggregation method used, but the order between explanations methods is quite consistent, for example Linear Explanations [23] score the highest regardless of aggregation method used.

Statistical Significance Table 7 reports the statistical significance of our results using the two-sample Student's t-test. We can see we have statistically significant evidence that Linear Explanations performs better than all the other methods, with the exception of CLIP-Dissect when using BRAgg with uniform prior, and all methods with the SigLIP prior.

Model	Aggregation	CLIP-Dissect[22]	INVERT l=1[7]	DnD[1]	MAIA[29]	LE(SigLIP) l=1[23]
ResNet-50 (Layer4)	Majority	0.0978 ± 0.010	0.0981 ± 0.010	0.0978 ± 0.010	0.0834 ± 0.010	**0.1216 ± 0.011**
	BRAgg (Uniform Prior)	0.1017 ± 0.009	0.0977 ± 0.010	0.1011 ± 0.008	0.0876 ± 0.010	**0.1261 ± 0.011**
	BRAgg (SigLIP Prior)	0.1356 ± 0.008	0.1472 ± 0.005	0.1179 ± 0.007	0.0992 ± 0.008	**0.1803 ± 0.008**
ViT-B-16 (Layer11 mlp)	Majority vote	0.1035 ± 0.016	0.0754 ± 0.016	0.0825 ± 0.017	0.0723 ± 0.013	**0.1208 ± 0.015**
	BRAgg (Uniform Prior)	0.0934 ± 0.014	0.0637 ± 0.014	0.0689 ± 0.013	0.0555 ± 0.012	**0.1018 ± 0.012**
	BRAgg (SigLIP Prior)	0.1568 ± 0.017	0.1167 ± 0.020	0.0904 ± 0.015	0.1257 ± 0.017	**0.2450 ± 0.021**
Average	BRAgg (SigLIP Prior)	0.1462	0.132	0.1042	0.1125	**0.2127**
	All	0.1148	0.0998	0.0931	0.0873	**0.1493**

Table 6. Detailed results of our large scale crowdsourced study. For each model we tested 100 random neurons, and we report the average correlation score of the explanations produced by each method, as well as the standard error of the mean. Comparing different rating aggregation methods, we can see that the correlation scores themselves vary a lot depending on the aggregation method used, but the order between explanations methods is quite consistent.

		Hypothesis			
Model	Aggregation	LE >CLIP-Dissect	LE >INVERT	LE >DnD	LE >MAIA
RN-50	BRAgg(Uniform)	0.0601	**0.0079**	**0.0198**	**0.0010**
	BRAgg(SigLIP)	**0.0020**	**0.0098**	**0.0000**	**0.0000**
ViT-B-16	BRAgg(Uniform)	0.4295	**0.0037**	**0.0431**	**0.0062**
	BRAgg(SigLIP)	**0.0039**	**0.0001**	**0.0000**	**0.0001**

Table 7. Statistical significance of our results according to two-sample students t-test. We report the p-values of different hypothesis, with statistically significant p-values in bold.

Example Ratings Figures 15, 16 and 17 showcase example neurons and the descriptions assigned by different explanation methods, as well as the correlation scores estimated by our crowd-sourced study for those explanations. We colored explanations based on the estimated correlation coefficient, with Green: $\rho > 0.25$, Yellow: $0.25 \geq \rho > 0.10$ and Red: $0.10 \geq \rho$.

We can see sometimes generative methods like MAIA [29] produce the best explanations, for example "Vibrant Green Elements in Nature" in Figure 15(Neuron 1126), but other times are too specific or fail to find the right concept 16.

Overall we did not observe very high correlation scores for any neurons in ResNet50, likely due to them being more densely activated and polysemantic, see for example neuron 108 (Fig. 15 that activates for both trains and invertebrates. On the other hand, for ViT-B-16 we observed several neurons that were extremely interpretable (correlation scores >0.5), such as Neuron 2187 (Fig. 17) that seems to only activate on Alpine ibex/Mountain goats.

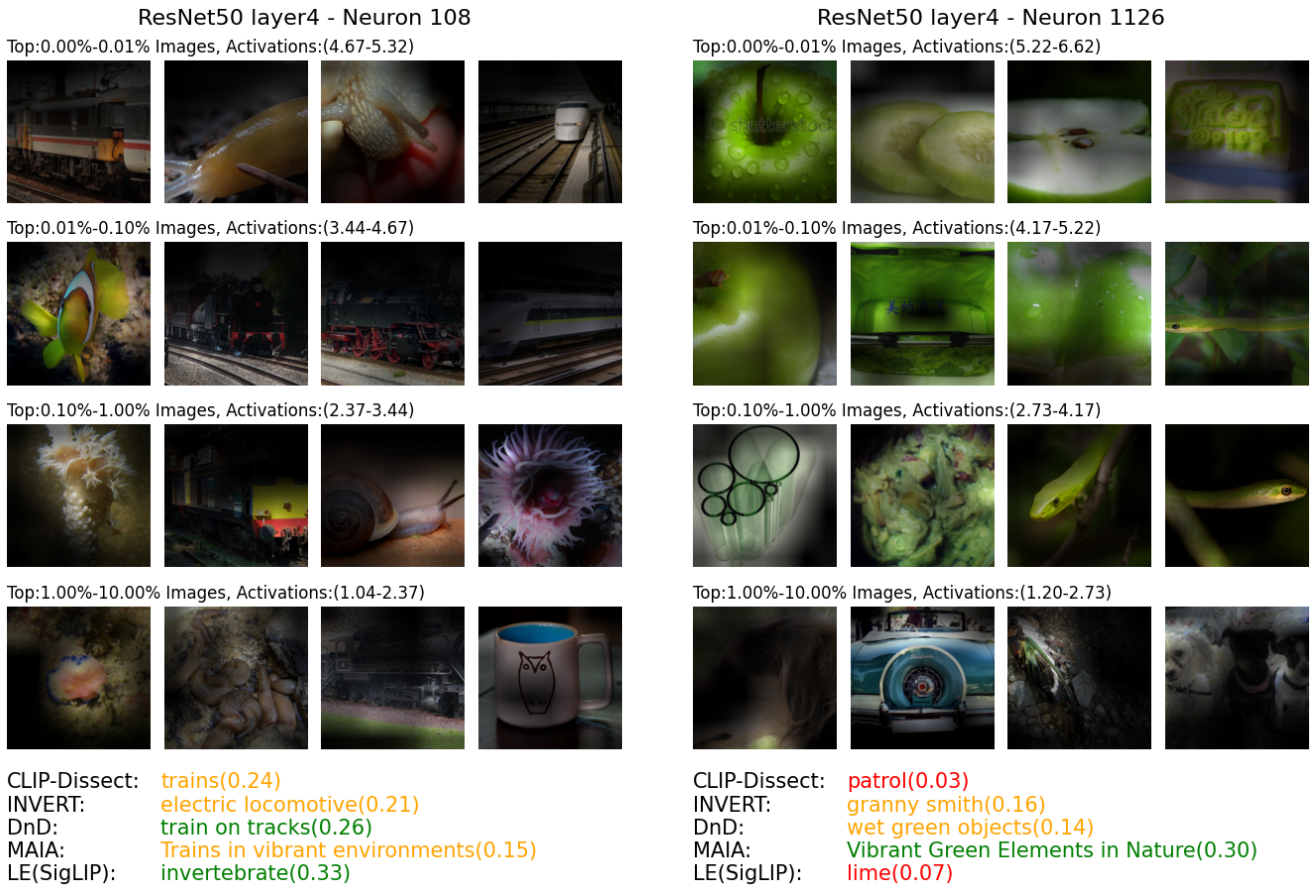


Figure 15. Visualization of example neurons, their descriptions and correlations scores from our crowdsourced evaluation (BRAgg with SigLIP prior). We have colored the descriptions based on the correlation score.

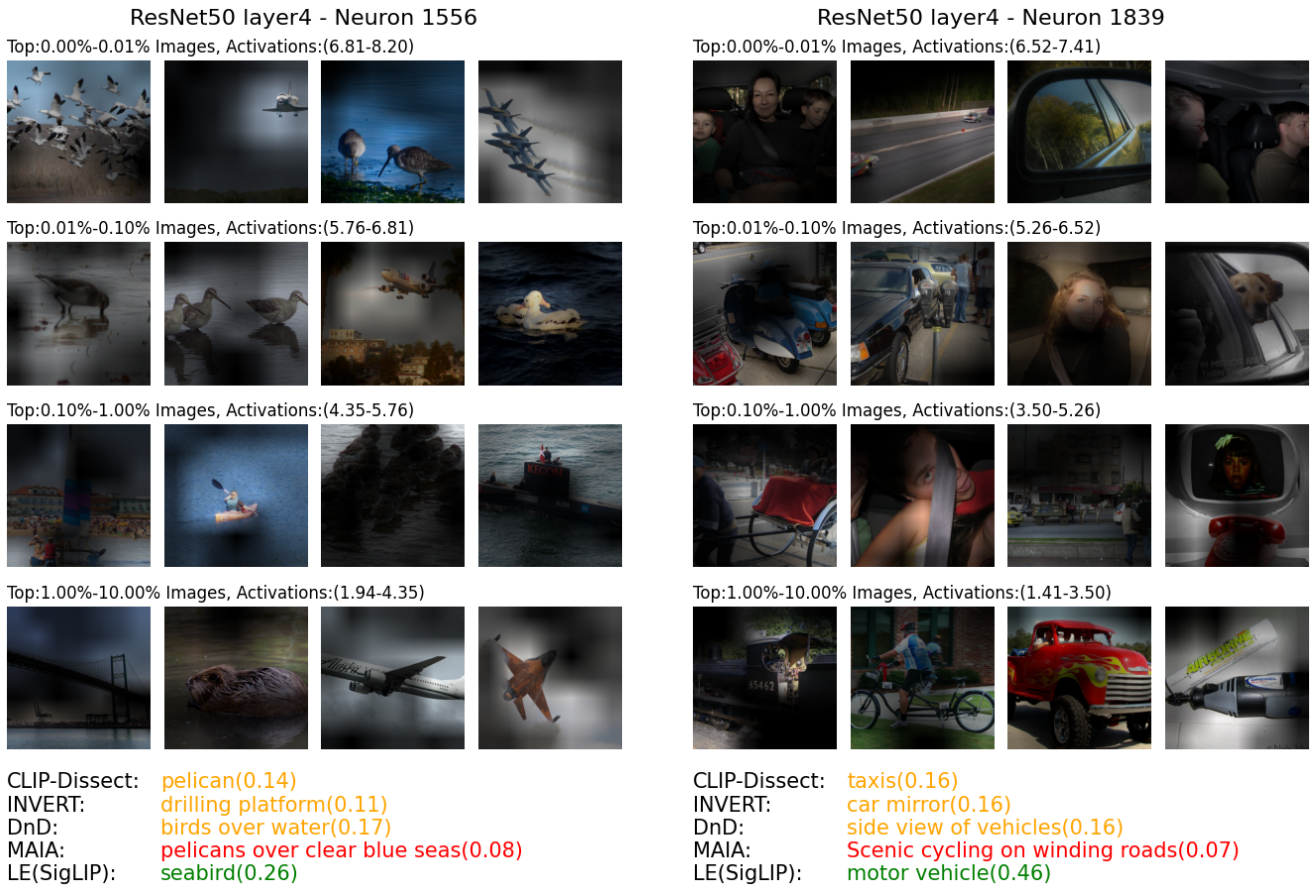


Figure 16. Visualization of example neurons, their descriptions and correlations scores from our crowdsourced evaluation (BRAgg with SigLIP prior). We have colored the descriptions based on the correlation score.

Figure 17. Visualization of example neurons, their descriptions and correlations scores from our crowdsourced evaluation (BRAgg with SigLIP prior). We have colored the descriptions based on the correlation score.

C.3. Experimental Details

Our evaluation focused on two models/layers:

1. Layer4 (end of residual block 4) of ResNet-50 trained on ImageNet. For simulation we use the neuron activation after average pooling, giving scalar activations, but for methods designed for 2d-activations, activations before average pooling are given as input.
2. MLP-activations in Layer 11 of ViT-B-16 [9] Encoder trained on ImageNet. We focus only on the activations of the CLS-token as this is the last layer and other tokens do not affect the prediction.

We use the full ImageNet validation dataset as $\mathcal{D}$ for the human study for all methods, and for generating explanations unless the method requires a specific dataset for explanation generation such as Broden [2].

For all methods using SigLIP guidance we use the SigLIP-SO400M-14-386 model.

C.3.1. Baseline Implementation Details

For practical purposes we made slight modifications to some baseline methods. The changes are discussed in detail below:

DnD [1]: The original implementation uses GPT-3.5 Turbo through the OpenAI API. Given the high cost of using the API and that recent open-source LLMs have strong performance compared to even closed-source LLMs like GPT-3.5, we replace GPT-3.5 Turbo with Llama 3.1-8B-Instruct [11]. DnD [1] showed that the older Llama 2 model was already better than GPT-3.5 Turbo and on-par with GPT-4 Turbo for neuron description, so our choice of Llama 3.1 does not degrade DnD's performance w.r.t. GPT-3.5 Turbo.

MAIA [29]: Compared to the original implementation, we replace GPT4-vision-preview with the newer GPT4o-2024-08-06 [25] since it has lower API costs and better performance. The method is still quite expensive, costing us approximately $65 and $116 to generate descriptions for 100 randomly selected neurons of ResNet-50 Layer4 and ViT-B-16 Layer11 MLP respectively. Note that this cost also includes repeating the experiment for $\sim$10 and $\sim$20 neurons from ResNet and ViT respectively which did not yield any neuron description in the first run. Initially, we also tried open-source LLMs with support for vision inputs (*i.e.* VLMs) like Llama-3.2-11B-Vision-Instruct [11], Llava-OneVision-Qwen2-7B-ov-hf [20], and CogVLM2-Llama3-Chat-19B [16]. However, these VLMs do not work well with MAIA since they fail to generate executable code, forget the image tokens and focus only on the last few text tokens if given a long prompt, and allow only one image input at a time. This is likely because these VLMs are geared towards visual question answering and do not possess the more generalized capabilities of GPT4/4o.

Methods relying on 2d activations: Many methods are designed to explain entire channels of CNNs with 2d activations [2, 15, 19, 21]. For ResNet-50 layer4 we fed the pre-avg pool activations to these methods for proper 2d input. However, for ViT-B-16 last layer only the CLS-token activations affect the output, and as such we are explaining neurons with scalar activations. In this case, we broadcasted the scalar activations into a 2d-tensor with the same value in all spatial locations. However this is not the intended way to use these methods and may partially explain poorer performance of some methods on ViT-neurons as observed in Tables 1 and 2.

CCE [19]: For Clustered Compositional Explanations, we tested two different versions: the $l = 15$ version corresponds to the default version with length 3 explanations for each of the 5 activation clusters. For the $l = 5$ version we used explanation length=1 with 5 clusters of activations. We also used the implementation of [19] to reproduce the results of Compositional Explanations [21] by setting explanation length=3 and number of clusters=1.

C.3.2. Subset of Neurons

For most methods in Table 1 and 2, we report the average correlation scores across all 2048 neurons for ResNet-50 layer4 and 3072 neurons for ViT-B-16 layer11 mlp. However for certain methods due to high computational and/or API cost we were only able to explain a subset of these neurons and report the average score of these subsets in Tables 1 and 2. We report the results for a subset of neurons for the following methods:

- MAIA [29]: Randomly selected subset of 100 neurons each for both RN50 and ViT-B-16.
- CCE [19] $l = 5$: For ViT-B-16 we used a subset of 1420 neurons. RN50 evaluated on all neurons.
- CCE [19] $l = 15$: RN50: subset of 984 neurons. ViT-B-16: subset of 422 neurons.

All other methods were evaluated on all neurons of each layer.

C.4. Automated Evaluation Details

For our automated evaluation(Sec. 5.1), we use Simulation with Correlation Scoring as described by [23]. This evaluation was originally proposed for language model neurons by [4].

The basic idea of simulation evaluation to use the *explanation* to predict neuron activations on unseen inputs. With correlation scoring we then evaluate the correlation coefficient ρ between the predicted activations s and actual neuron activations a_k on the entire test split of 10,000 inputs as done by [23].

For **simple explanations**, the predicted activation s is simply the presence of concept on that input.

$$s(x_i, t) = [c_t]_i \tag{22}$$

For a **linear explanation**, $E = \{(w_1, t_1), ..., (w_l, t_l)\}$ the predicted activation s is calculated following [23] as:

$$s(x_i, E) = \sum_{w_j, t_j \in E} w_j [c_{t_j}]_i \tag{23}$$

For **compositional explanations** [21], we calculate the predicted activation as follows using probabilistic logic. Different basic logical operators are calculated as:

$$s(x_i, t_1 \text{ AND } t_2) = [c_{t_1}]_i \cdot [c_{t_2}]_i \tag{24}$$

$$s(x_i, t_1 \text{ OR } t_2) = 1 - (1 - [c_{t_1}]_i) \cdot (1 - [c_{t_2}]_i) \tag{25}$$

$$s(x_i, \text{NOT } t) = 1 - [c_t]_i \tag{26}$$

Predictions for more complex compositions are then calculated by iteratively applying these rules.

Clustered Compositional Explanations: CCE [19] explanations are of the form $E = \{(l_1, u_1, F_1), ..., (l_r, u_r, F_r)\}$ where r is the number of activation clusters, and l_j, u_j are the lower and upper bound of activations for that cluster and F_j is a compositional explanation for activations of that cluster. To predict neuron activation based on this explanation, we use the following formula:

$$s(x_i, E) = \sum_{l_j, u_j, F_j \in E} \frac{l_j + u_j}{2} s(x_i, F_j) \tag{27}$$

This means if the concepts according to the formula of a cluster are present, we predict the neuron's activation will be in the middle of the clusters activation range.

For all automated evaluations we use SigLIP-SO400M-14-386 model to predict c_t following [23].

C.5. Theorem 1

Suppose we are estimating the expected value of function $h(x)$ when $x \sim \mathcal{P}$. Let $\mathcal{X}$ be the support of $\mathcal{P}$.

Theorem 1 ([28], Sec 3.3.2, Theorem 3.12)**.** *For importance sampling with sampling distribution q:*

$$\mathbb{E}_{x \sim \mathcal{P}}[h(x)] = \int_{\mathcal{X}} h(x) \frac{p(x)}{q(x)} q(x) dx \approx \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{h(x_i) p(x_i)}{q(x_i)}.$$

The choice of q that minimizes the variance satisfies $q(x) \propto |h(x)|p(x)$.

Proof: Reproduced from [28].

$$Var\left[\frac{h(x)p(x)}{q(x)}\right] = \mathbb{E}_q\left[\left(\frac{h(x)p(x)}{q(x)}\right)^2\right] - \left(\mathbb{E}_q\left[\frac{h(x)p(x)}{q(x)}\right]\right)^2 \tag{28}$$

Since the the second term $\left(\mathbb{E}_q\left[\frac{h(x)p(x)}{q(x)}\right]\right)^2 = \left(\int_{\mathcal{X}} h(x)p(x)dx\right)^2$ does not depend on q, in order to minimize variance we only need to minimize the first term.

From Jensen's inequality it follows that:

$$\mathbb{E}_q\left[\left(\frac{h(x)p(x)}{q(x)}\right)^2\right] \geq \left(\mathbb{E}_q\left[\frac{|h(x)|p(x)}{q(x)}\right]\right)^2 = \left(\int_{\mathcal{X}} |h(x)|p(x)dx\right)^2 \tag{29}$$

Giving us a lower bound for the first term. If we set

$$q(x) = \frac{|h(x)|p(x)}{\int_{\mathcal{X}} |h(z)|p(z)dz} \tag{30}$$

Which is a valid probability distribution, we get:

$$\mathbb{E}_q \left[\left(\frac{h(x)p(x)}{q(x)} \right)^2 \right] = \left(\int_{\mathcal{X}} |h(z)|p(z)dz \right)^2 \tag{31}$$

This exactly matches the lower bound, proving that minimum variance is attained by setting

$$q(x) = \frac{|h(x)|p(x)}{\int_{\mathcal{X}} |h(z)|p(z)dz} \propto |h(x)|p(x) \tag{32}$$

C.6. Compute details

Our main contribution is focused on efficient crowd-sourced evaluation and as such our method is not computationally costly. The main computational cost associated with our method is calculating the SigLIP image encoder outputs for the entire $\mathcal{D}$, as these are needed for both importance sampling and the Bayes - SigLIP prior. However this is a relatively cheap onetime cost of around 20 minutes on a single NVIDIA RTX 6000 Ada Generation GPU.

The main computational expense associated with this paper involved running the baseline methods. In Tables 8 and 9 below, we report the approximate runtime to explain all neurons of a layer using different description methods using a single NVIDIA RTX 6000 Ada Generation GPU.

Simple Exp. (l=1)	ND [3]	MILAN [15]	CD [22]	INVERT l=1 [7]	DnD [1]	MAIA [29]	LE(label) l=1 [23]	LE(SigLIP) l=1 [23]
RN-50 (Layer4)	$\sim$ 1 hr	$\sim$ 1 hr	$\sim$ 5 mins	$\sim$ 1 hr	$\sim$ 55 hrs	$\sim$ 255 hrs	$\sim$ 1 hr	$\sim$ 1 hr

Table 8. Approximate runtime of different $l = 1$ baseline methods to explain neurons.

Complex Exp. (l>1)	Comp Exp [21] l=3	INVERT [7] l=3	CCE [19], l= 5	CCE [19], l=15	LE(label) [23] l=4.37/1.97	LE(SigLIP) [23] l=4.66/1.82
RN-50 (Layer4)	$\sim$ 92 hrs	$\sim$ 24 hrs	$\sim$ 87 hrs	$\sim$ 275 hrs	$\sim$ 1 hr	$\sim$ 1 hr

Table 9. Approximate runtime of different complex explanation baseline methods.

C.7. Inter-Annotator Consistency

To measure the quality of our ratings, we also measure inter-annotator consistency through Fleiss's Kappa. This measures the amount of agreement above chance, with 1 meaning perfect agreement and 0 random chance.

For MTurk raters in our test Setting 2, we observe a Kappa of 0.395. On the real study we observe Fleiss's Kappa of 0.191 on ViT neuron explanations and 0.274 on RN50 neurons. Overall we can see these agreement rates are relatively low, highlighting the need for error correction trough rating aggregation. While we think some of the noise in responses is underlying poor quality responses, some disagreement is also caused by underlying ambiguity in the task or poor explanations like MAIA sometimes outputting "Undetermined Selectivity".

To compare against higher quality annotations, we had 3 authors rate 150 samples each for the 10 neurons in Setting 2. For authors we observed a Fleiss Kappa of 0.774. This shows that obtaining higher quality raters has potential to improve study signal, but as we show we can still efficiently utilize noisy ratings with our methods.