Image-computable Bayesian model for 3D motion estimation with natural stimuli explains human biases

Anonymous

Abstract

Humans estimate the 3D motion of the self and of objects in the natural environment from the 2D images formed in the two eyes. To understand how this problem should be solved, we trained Bayesian observers on naturalistic binocular videos to solve two tasks: 3D speed estimation and 3D direction estimation. The resulting normative models leverage two cues used by humans–the changing disparity and the interocular velocity difference cues–and show that quadratic combination of linear filter responses is an optimal computation for speed estimation but not for direction estimation. The models reproduce the psychophysical response patterns that characterize human performance in 3D motion estimation tasks, including biases, discrimination thresholds, and counter-intuitive towards-away confusions. These results suggest that, rather than resulting from a haphazardly constructed system, the sometimes surprising performance patterns in human 3D motion perception result from optimal visual information processing.

1 Introduction

Interaction with the world entails the motion in depth of objects and surfaces relative to our eyes. Thus, accurate estimation of motion through depth is essential for successful interaction with the environment. Animals with binocular vision combine the 2D images formed by each eye to estimate 3D motion. However, even though estimating motion through depth is critical, it is not well understood how the brain does, and should, estimate 3D motion from binocular images (Cormack et al.).

To better understand the computations that optimize local 3D motion estimation, we developed an image-computable Bayesian ideal observer, constrained by the front-end properties of the human visual system (see Burge & Geisler; Jaini & Burge). We trained this model on groundtruth-labeled naturalistic binocular movies in two specific tasks: 3D speed estimation and 3D direction estimation. For each task, the model learns a small set of spatio-temporal linear filters that extract the most useful task-specific features from the binocular movies. Non-linear decoding of the noisy filter responses using probabilistic inference yields optimal estimates of the value of the latent variable (i.e. 3D speed or direction of motion).

We find that for the task of estimating 3D speed through depth, the optimal filter responses are approximately Gaussian-distributed when conditioned on the latent variable. This finding means that, given the natural statistics of filter responses, the optimal computations for 3D speed estimation can be described as an extended version of the energy model (i.e. linear filtering followed by quadratic combination). The energy model has been widely used to describe neural activity associated with many visual tasks, and has been described as a canonical neural computation (Burge; Jaini & Burge; Adelson & Bergen; Kouh & Poggio; Ohzawa; Bergen & Adelson). The current result suggests that energy-model-like computations are normative, in that they are required by natural scene statistics for optimal inference. Interestingly, for the task of estimating 3D direction, the conditional filter responses are not well-approximated as Gaussian, indicating that energy-model-like computations are non-optimal for this task. Additionally, for both 3D speed and direction estimation, we compare

4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at the Neural Information Processing Systems (NeurIPS) conference 2022. New Orleans.

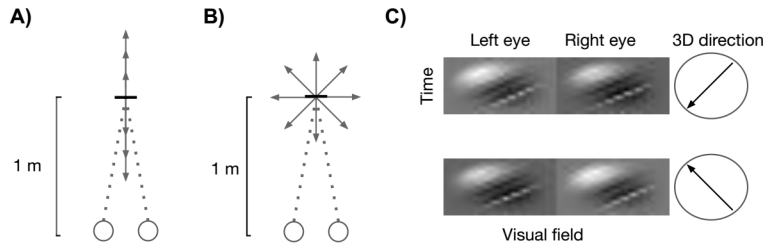

Figure 1: **Task design and stimuli. A)** Top-down view of the observer's eyes (the two bottom circles) and the target surface (the horizontal black bar). Grey arrows show the types of 3D motions used at this task. **B)** Same as A), but for the direction estimation task. **C)** Example of two binocular videos in the motion estimation dataset. The vertical dimension shows different frames of the 1D movie, and the horizontal dimension shows x-position. The 3D direction of the target relative to the observer (with the same layout as B) is also shown on the right.

the two Bayesian observer models against previously reported human psychophysical data (Rokers et al.; Bonnen et al.; Czuba et al.), and show that the models may provide a normative explanation of various aspects of human psychophysical performance.

2 Methods

2.1 Task design and dataset generation

We generated two binocular movie datasets, one for each of the different 3D-motion based tasks. In the 3D speed estimation task, planar target surfaces moved in depth along a straight line directly towards or away from the observer, with 3D speeds ranging from -2.5 m/s to 2.5 m/s in 0.125 m/s increments (Figure 1A). In the 3D direction task, the targets had fixed 3D speeds of 0.125 m/s, and moved along all directions in the XZ plane in 15°increments (Figure 1B). The target surfaces were textured with natural image patches sampled from a previously published natural scene dataset Burge et al.. Left- and right-eye movies were generated using the 3D motion geometry and perspective projection. The model also incorporated the optics of the human eye and basic retinal physiology (e.g. wavelength sensitivity of the human cones) (Burge & Geisler). Binocular movies subtended 1°of visual angle in each eye (60 pixels/°), and had a duration of 250 ms (60 Hz). Each movie was then contrast-normalized and vertically averaged (see Burge & Geisler). 500 naturalistic image movies were generated for each 3D speed or direction (Figure 1C).

2.2 Model structure and training

Accuracy Maximization Analysis (AMA), a Bayesian ideal observer method for task-specific dimensionality reduction, was used to learn the filters (receptive fields) that select the most useful stimulus features for each 3D motion task. Then, we determined the non-linear computations that optimize decoding performance for each task (Geisler et al.; Burge & Jaini; Jaini & Burge). AMA requires a filter response model. We chose a response model tightly linked to the properties of the primate visual system. Noisy filter responses were generated as follows:

$$R_i = \boldsymbol{f_i}^T \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|} + \eta \tag{1}$$

where R_i is the noisy response of filter f_i to natural stimulus $\boldsymbol{s}$, generated by filtering the stimulus and adding a sample of noise η. We model the noisy filter responses as Poisson-like, such that the noise variance is proportional to $\boldsymbol{f_i}^T \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|}$. Denoting the filter set $\boldsymbol{f} = [\boldsymbol{f}_1, ..., \boldsymbol{f}_n]$ and population filter response $\boldsymbol{R} = [R_1, ..., R_n]$, the noisy filter responses to a particular stimulus $P(R_i|\boldsymbol{s}) \sim \mathcal{N}(\boldsymbol{f_i}^T \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|}, \alpha \boldsymbol{f_i}^T \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|} + \sigma_0^2)$ are normally distributed where α is the constant of proportionality and σ_0^2 a baseline variance.

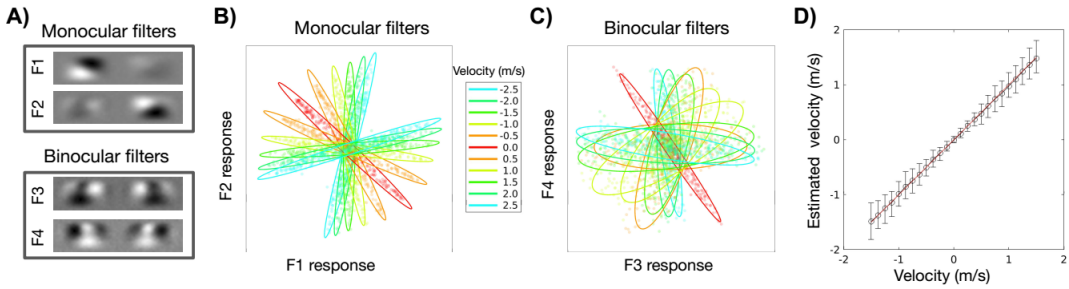

Figure 2: **3D speed estimation results. A)** Four (of eight) learned binocular filters for 3D speed estimation. Top: a pair of monocular filters, which extract information connected to the interocular velocity differences. Bottom: a pair of binocular filters, which relate to changing disparity. **B)** Conditional filter responses for two monocular filters, color coded by 3D speed. Ellipses show iso-probability contours corresponding to the best-fitting Gaussian. Each dot is the response to an individual stimulus. **C)** Conditional filter responses of two binocular filters. **D)** Distribution of estimated 3D speeds for each true 3D speed. Data points indicate the median estimate. Vertical bars show the 95% confidence intervals.

Given a particular response $\boldsymbol{R}$ of the filter population, the task is to estimate the value of the latent variable X (speed or direction) associated with the input stimulus. To achieve this, the posterior probability $P(X = X_j|\boldsymbol{R})$ must be obtained for each value X_j, which can be computed as the product of the likelihood $L(X_j; \boldsymbol{R}) = P(\boldsymbol{R}|X_j)$ and the prior $P(X_j)$. However, the conditional probability distributions of filter responses $P(\boldsymbol{R}|X_j)$–and thus the likelihoods–are *a priori* unknown. Re-expressing the conditional probability of filter response as $P(\boldsymbol{R}|X_j) = \sum_l P(\boldsymbol{R}|\boldsymbol{s}_l)P(\boldsymbol{s}_l|X_j)$ where l indexes each stimulus $\boldsymbol{s}$ in the training set at latent variable value X_j presents a way forward. The first term in the summation is specified by the filter response model above (Equation 1). The second term, $P(\boldsymbol{s}_l|X_j) = 1/N_j$ is specified by the training set, where N_j is the number of training stimuli for each value of the latent variable. Thus, given the response model, a training set, and an arbitrary set of filters, the posterior probabilities can be computed and read out into an estimate, from which a cost can be obtained. Here, we use (in a slight abuse of terminology) an 'information-theoretic cost function': the KL-divergence between the actual posterior probability distribution and an idealized posterior probability distribution with all its probability mass at the correct value of the latent variable. The objective of the training procedure is to find the filters $\boldsymbol{f}$ that minimize the average cost $\arg\min_{\boldsymbol{f}} \overline{C} = \frac{1}{N}\sum_{kl} C_{kl}$. The cost associated with an individual stimulus is $C_{kl} = E_{\boldsymbol{R}_{k,l}}[-lnP(X_k|\boldsymbol{R}_{k,l})]$ where l indexes a particular stimulus belonging to latent variable level k. With this procedure, we trained 8 filters for each model. Training with AMA (Geisler et al.; Burge & Jaini), which assumes no *a priori* structure over the conditional filter response distributions, and training with AMA-Gauss (Jaini & Burge), which assumes Gaussian-distributed conditional filter response distributions $P(\boldsymbol{R}|X_j)$, yields near-identical results in other similar low-level vision tasks (Jaini & Burge). Here, for computational convenience, we use AMA-Gauss to learn the optimal filters and non-linear decoding rules for estimating 3D speed and direction.

3 Results

3.1 3D motion speed estimation

Figure 2A shows four of eight binocular filters learned for 3D speed estimation. Figures 2B and 2C show the corresponding response distributions. The filter responses to naturalistic binocular movies, conditioned on the latent variable (speed of the stimulus), are well-approximated by Gaussian distributions. Hence, the log-likelihood of each value X_j of the latent variable can be estimated by a quadratic combination of the filter responses, with weights given by the mean μ_j and covariance Σ_j of the posterior distribution $P(\boldsymbol{R}|_j) \sim \mathcal{N}(\boldsymbol{R}; \mu_j, \Sigma_j)$. Thus, we find these computations of linear filtering followed by the quadratic combination of filter responses, to be optimal for decoding speed in depth from natural scenes. Importantly, these computations are closely related to the quadratic computations posited by the energy model, a popular descriptive model of complex cells. The current

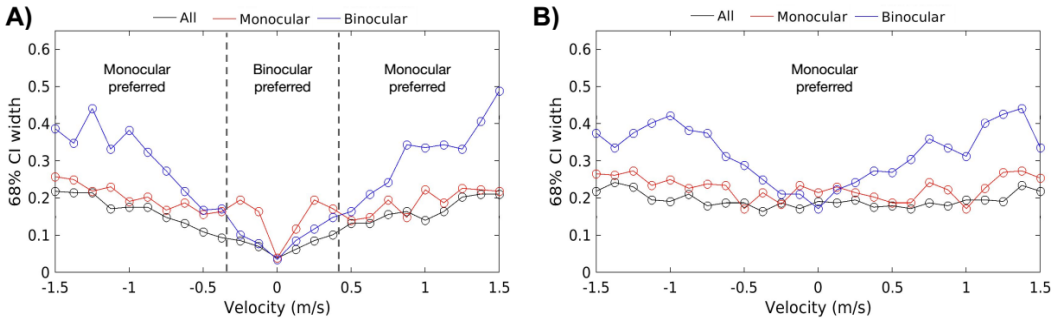

Figure 3: **3D speed estimation results. A)** Confidence intervals (68%) of the speed estimates when using all the filters (black), monocular filters only (red), or binocular filters only (blue). Binocular filters are more reliable for slow speeds. Monocular filters are more reliable for fast speeds **B)** Same as A, but with variability in the binocular disparity of the initial frame of the test movie. Monocular filters are now more reliable for all speeds. The initial disparity was selected from a Gaussian distribution with mean 0 and standard deviation of 5 arcmin.

results suggest that quadratic computations–which have previously been described as canonical neural computations (Kouh & Poggio)–are optimal for the task of estimating 3D speed with natural movies. We note that when training the model on unnormalized stimuli, the conditional response distributions are not Gaussian (data not shown), indicating that these resulting Gaussian distributions are not a trivial result of the model architecture.

Figure 2A shows that the model learns two types of filters: monocular filters and binocular filters. Monocular filters respond near-exclusively to the input in one eye. Binocular filters require matching input in both eyes for robust response. These two filter types appear to be related to two different 3D motion cues that are computed in human and non-human primate visual systems: the *inter-ocular velocity difference* (IOVD) cue and the *changing disparity over time* (CDOT) cue. The IOVD cue is thought to be computed by first performing a time derivative (on the monocular inputs) to obtain local velocities (v_L, v_R) , and then performing an binocular comparison $(v_R - v_L)$. The CDOT cue is thought to be computed by first performing a binocular comparison to obtain a binocular disparity δ, and then performing a time-derivative $\frac{\partial \delta}{\partial t}$. The CDOT cue can be estimated from the frame-by-frame comparison performed by the binocular filters over time.

To examine whether the normative model uses these two cues in a similar fashion to humans, we examined model performance with monocular filters alone, binocular filters alone, and both monocular and binocular filters together, as a function of 3D speed. Humans rely more on the IOVD cue (monocular filters) at high speeds and in the peripheral visual field, while they rely more on the CDOT cue (binocular filters) for estimating low speeds near the fovea (Czuba et al.). Figure 3A shows that, at slow speeds, the confidence intervals are smaller for the model estimates obtained from binocular filters than for estimates obtained from monocular filters. The opposite is true at larger speeds. The confidence intervals from the full model are more similar to those of the monocular filters at fast speeds, and more similar to the binocular filters at slow speeds, reproducing the patterns from human psychophysics.

Next, we examined whether the model accounts for differences between foveal and peripheral 3D motion processing. In viewing natural scenes, binocular disparity is usually close to zero in the central visual field (i.e. the fovea), because binocular eye movements tend to accurately fixate the surfaces of targets. In the visual periphery, the distribution of binocular disparities are much more variable (Iyer & Burge; Sprague et al.). To mimic the increased disparity variation in the peripheral visual field, we introduced variability in the disparity of the target surface on the initial frame of each movie. The expectation was that the binocular filters become less useful relative to the monocular filters under these conditions. Figure 3B shows that, with increased disparity variability, the usefulness of the binocular filters (i.e. the CDOT cue) is reduced relative to the monocular filters (i.e. the IOVD cue). This result dovetails with human psychophysics, and suggests that the adaptive use of motion cues across conditions reflects the optimal computations for speed estimation.

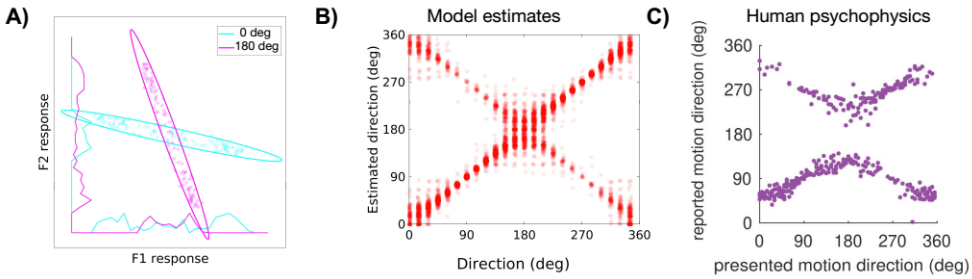

Figure 4: **3D direction estimation results A)** Conditional response distribution for two learned filters. Points show individual stimuli and ellipses the iso-probability contours from a fitted Gaussian, color coded like in Figures 2B and 2C. The lines on the left and on the bottom show the marginal densities of each filter activation, where bimodal conditional distributions can be appreciated. We only show two motion directions for clarity. B) Bayesian ideal observer estimates of 3D motion direction in the XZ plane vs. the true direction of motion. The 0motion direction corresponds to a target moving to the right of the observer and the 90motion direction indicates motion away from the observer. C) Human estimates of target motion. Figure modified from online open data provided in Bonnen et al..

3.2 3D motion direction estimation

Here, we use the Bayesian ideal observer to determine the optimal computations for 3D direction estimation in the XZ plane, and we examine whether the ideal observer accounts for neural computations and human psychophysical performance in the task. Unlike the previous task, the responses of the filters learned for direction estimation are not well-approximated by Gaussian distributions, as can be seen in the bimodal histograms of the marginal response distributions (Figure 4A). This result suggests that the energy model is not the optimal computation for this task.

Next, we analyzed how the performance of the model, using quadratic decoding, compares to human psychophysical performance. Note that for this task, quadratic decoding under a Gaussian approximation of the conditional filter responses is non-optimal. In some conditions–left and right 3D motion directions–model performance is very good. Interestingly, however, the model confuses towards and away motion directions a non-negligible proportion of the time. These towards-away confusions generate a characteristic X-shaped pattern of responses, a pattern of response that is characteristic of human performance in laboratory-based 3D motion tasks Bonnen et al.; Rokers et al., (Figures 4B,C). Thus, although energy model-like computations are not optimal for this task, the model reproduces the patterns of human psychophysical performance.

4 Conclusions

Optimal estimation of 3D motion speed from natural image movies is supported by the quadratic combination of task-optimized spatio-temporal filter responses. The ideal observer for 3D speed estimation makes adaptive use of two closely related 3D motion cues (IOVD and CDOT) similar how those same cues are used by humans (Czuba et al.), and produces a pattern of performance that matches human performance. The current results provide a normative account of the human psychophysical results.

Optimal estimation of 3D direction is supported by more sophisticated computations. But a model that uses (non-optimal) quadratic computations exhibits performance similar to humans in a direction estimation task Bonnen et al.; Rokers et al.. Specifically, like humans, the model makes estimates marred by towards-away direction confusions. Future work will analyze whether an optimal image-computable model still shows this pattern of towards-away confusion.

Here, we presented an analysis of 3D motion estimation from naturalistic binocular movies that shows that a canonical computation that is used in descriptive models of complex cells is the optimal computation for 3D speed estimation, but not 3D direction estimation. These findings highlight the potential utility of incorporating canonical quadratic computations into neural networks for some tasks, and the need to consider their potential in a task-specific manner.

References

References

Edward H. Adelson and James R. Bergen. Spatiotemporal energy models for the perception of motion. 2(2):284–299. ISSN 1520-8532. doi: 10.1364/JOSAA.2.000284. URL https://www.osapublishing.org/josaa/abstract.cfm?uri=josaa-2-2-284. Publisher: Optical Society of America.

James R. Bergen and Edward H. Adelson. Early vision and texture perception. 333(6171):363–364. ISSN 0028-0836, 1476-4687. doi: 10.1038/333363a0. URL http://www.nature.com/articles/333363a0.

Kathryn Bonnen, Thaddeus B. Czuba, Jake A. Whritner, Adam Kohn, Alexander C. Huk, and Lawrence K. Cormack. Binocular viewing geometry shapes the neural representation of the dynamic three-dimensional environment. 23(1):113–121. ISSN 1546-1726. doi: 10.1038/s41593-019-0544-7. URL https://www.nature.com/articles/s41593-019-0544-7. Number: 1 Publisher: Nature Publishing Group.

Johannes Burge. Image-computable ideal observers for tasks with natural stimuli. 6:22.1–22.27. doi: 10.1146/annurev-vision-030320-041134. URL https://www.annualreviews.org/doi/abs/10.1146/annurev-vision-030320-041134. Publisher: Annual Reviews.

Johannes Burge and Wilson S. Geisler. Optimal speed estimation in natural image movies predicts human performance. 6(1):7900. ISSN 2041-1723. doi: 10.1038/ncomms8900. URL https://www.nature.com/articles/ncomms8900. Number: 1 Publisher: Nature Publishing Group.

Johannes Burge and Priyank Jaini. Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. 13(2):e1005281. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005281. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005281. Publisher: Public Library of Science.

Johannes Burge, Brian C. McCann, and Wilson S. Geisler. Estimating 3d tilt from local image cues in natural scenes. 16(13):2. ISSN 1534-7362. doi: 10.1167/16.13.2. URL https://doi.org/10.1167/16.13.2.

Lawrence K. Cormack, Thaddeus B. Czuba, Jonas Knöll, and Alexander C. Huk. Binocular mechanisms of 3d motion processing. 3:297–318. ISSN 2374-4642. doi: 10.1146/annurev-vision-102016-061259. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5956901/.

Thaddeus B. Czuba, Bas Rokers, Alexander C. Huk, and Lawrence K. Cormack. Speed and eccentricity tuning reveal a central role for the velocity-based cue to 3d visual motion. 104(5):2886–2899. ISSN 0022-3077. doi: 10.1152/jn.00585.2009. URL https://journals.physiology.org/doi/full/10.1152/jn.00585.2009. Publisher: American Physiological Society.

Wilson S. Geisler, Jiri Najemnik, and Almon D. Ing. Optimal stimulus encoders for natural tasks. 9(13):17. ISSN 1534-7362. doi: 10.1167/9.13.17. URL https://doi.org/10.1167/9.13.17.

Arvind V. Iyer and Johannes Burge. Depth variation and stereo processing tasks in natural scenes. 18(6):4. ISSN 1534-7362. doi: 10.1167/18.6.4. URL https://doi.org/10.1167/18.6.4.

Priyank Jaini and Johannes Burge. Linking normative models of natural tasks to descriptive models of neural response. 17(12):16. ISSN 1534-7362. doi: 10.1167/17.12.16. URL http://jov.arvojournals.org/article.aspx?doi=10.1167/17.12.16.

Minjoon Kouh and Tomaso Poggio. A canonical neural circuit for cortical nonlinear operations. 20(6):1427–1451. ISSN 0899-7667. doi: 10.1162/neco.2008.02-07-466. URL https://doi.org/10.1162/neco.2008.02-07-466.

Izumi Ohzawa. Mechanisms of stereoscopic vision: the disparity energy model. 8(4):509–515. ISSN 0959-4388. doi: 10.1016/S0959-4388(98)80039-1. URL https://www.sciencedirect.com/science/article/pii/S0959438898800391.

Bas Rokers, Jacqueline M. Fulvio, Jonathan W. Pillow, and Emily A. Cooper. Systematic misperceptions of 3-d motion explained by bayesian inference. 18(3):23. ISSN 1534-7362. doi: 10.1167/jov.18.3.23. URL https://doi.org/10.1167/jov.18.3.23.

William W. Sprague, Emily A. Cooper, Ivana Tošić, and Martin S. Banks. Stereopsis is adaptive for the natural environment. 1(4):e1400254. doi: 10.1126/sciadv.1400254. URL https://www.science.org/doi/full/10.1126/sciadv.1400254. Publisher: American Association for the Advancement of Science.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, see the end of paragraph 2 from the conclusions.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]