

LAION-C: AN OUT-OF-DISTRIBUTION BENCHMARK FOR WEB-SCALE VISION MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) robustness is a desired property of computer vision models. Improving model robustness requires high-quality signals from robustness benchmarks to quantify progress. While various benchmark datasets such as ImageNet-C were proposed in the ImageNet era, most ImageNet-C corruption types are no longer OOD relative to today’s large datasets scraped from the web, which already contain common corruptions such as blur or JPEG compression artifacts. Consequently, these standard benchmarks are no longer well-suited for evaluating OOD robustness in the era of web-scale datasets. Indeed, recent models show saturating scores on ImageNet-era OOD benchmarks, indicating that it is unclear whether models trained on web-scale datasets truly become better at OOD generalization or whether they have simply been exposed to the test distortions during training. To address this, we here introduce LAION-C as a benchmark alternative for ImageNet-C. LAION-C consists of six novel distortion types specifically designed to be OOD, even for web-scale datasets such as LAION. In a comprehensive evaluation of state-of-the-art models, we find that the LAION-C dataset poses significant challenges to contemporary models, including MLLMs such as Gemini and GPT-4o. We additionally conducted a psychophysical experiment to evaluate the difficulty of our proposed corruptions for human observers, enabling a comparison of models to lab-quality human robustness data. We observe a paradigm shift in OOD generalization: from humans outperforming models, to the best models now matching or outperforming the best human observers.

1 INTRODUCTION

In recent years, large-scale vision models such as vision transformers (Dosovitskiy et al., 2021) and ConvNeXt (Liu et al., 2022), trained on large web-scale datasets like LAION (Schuhmann et al., 2022), have pushed the limits of performance on standard benchmarks. However, the continued advancement and reliable evaluation of these models depends on the availability of datasets that effectively challenge model robustness and generalization capabilities.

In the era of training models on curated datasets like ImageNet (Russakovsky et al., 2015), creating OOD-benchmarks was relatively straight-forward: By introducing visual corruptions that were absent from the training set, such as blur and noise, researchers could evaluate the robustness of their models in a controlled manner. If a model performs well on a corruption it has never seen, it can be said to be robust to this corruption. For example, ImageNet-C (Hendrycks & Dietterich, 2019), which introduces different parametric corruptions to the ImageNet validation set, has long stood as the de facto standard for OOD evaluation. Models that were trained on the (uncorrupted) ImageNet images must robustly generalize in order to perform well on ImageNet-C.

With the shift towards training models on vast, largely unfiltered image datasets, it is much less clear how to obtain test images that are truly OOD. Adding noise and blur to images can no longer be considered a distribution shift, because such images are already present in the training set, as demonstrated in Fig. 1. Models trained on LAION have seen the types of corruption in ImageNet-C, and are presumably exposed to all realistic corruptions. Therefore, the fact that models like CLIP (Radford et al., 2021) exhibit much better performance on classic OOD datasets than ImageNet-trained models might not be an indication of true OOD robustness, but rather a consequence of a smaller train-test gap. For distribution shifts defined by the style of an image, recent work empirically shows

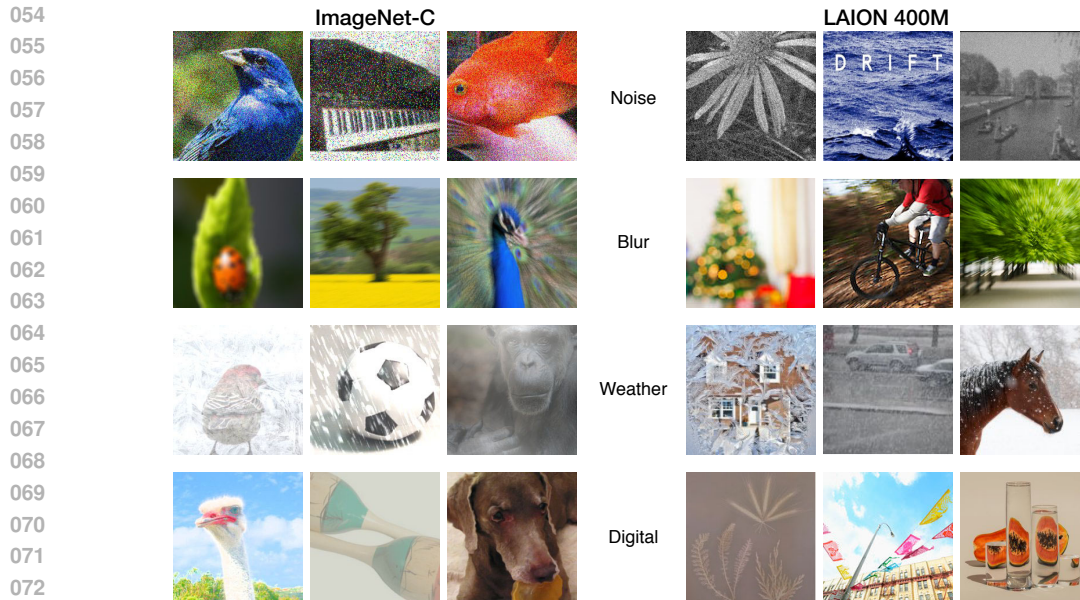


Figure 1: **ImageNet-C corruptions are not out-of-distribution (OOD) for web-scale datasets like LAION-400M.** Exemplary corrupted images from ImageNet-C (left) are similar to LAION-400M samples (right). Each row shows example corruptions and dataset images for one ImageNet-C corruption category (Noise, Blur, Weather, Digital). The presence of these distortions in web-scale datasets indicates the need for an OOD benchmark in the era of web-scale vision models.

that such datasets are indeed not OOD, but overlap with LAION-400M (Mayilvahanan et al., 2023; 2024). Hence, to measure the OOD robustness of modern models, a dataset containing truly new image corruptions is needed—even if this means that the corruptions must be highly artificial.

Our **contributions** are as follows:

1. Given that existing OOD datasets are often no longer OOD for models trained on web-scale datasets, we introduce LAION-C, a **novel benchmark** dataset with six manually designed corruption types and 16 superclasses to evaluate the robustness of web-scale vision models. This dataset serves as a proxy for unseen challenges, allowing us to probe the limits of current models’ robustness in a controlled but challenging environment.
2. We conduct a comprehensive performance analysis of various models on LAION-C and report a robust human OOD generalization baseline obtained through **psychophysical experiments** with 19 participants, collecting 11,400 trials in a highly controlled laboratory environment.
3. The resulting data serves as an OOD benchmark for current and future models, enabling not only an assessment of their generalization ability on truly OOD data but also providing insights into the **discrepancies between human and machine perception**, observing a paradigm shift in OOD generalization: from humans outperforming models to the best models now matching or outperforming the best human observers.

A detailed related work section can be found in [Appx. A.1](#).

2 METHODS

2.1 CONSTRUCTING NEW OOD DISTORTIONS

We introduce LAION-C, a novel dataset with six synthetic distortions designed to be OOD even for web-scale models, requiring genuine generalization beyond training distributions.

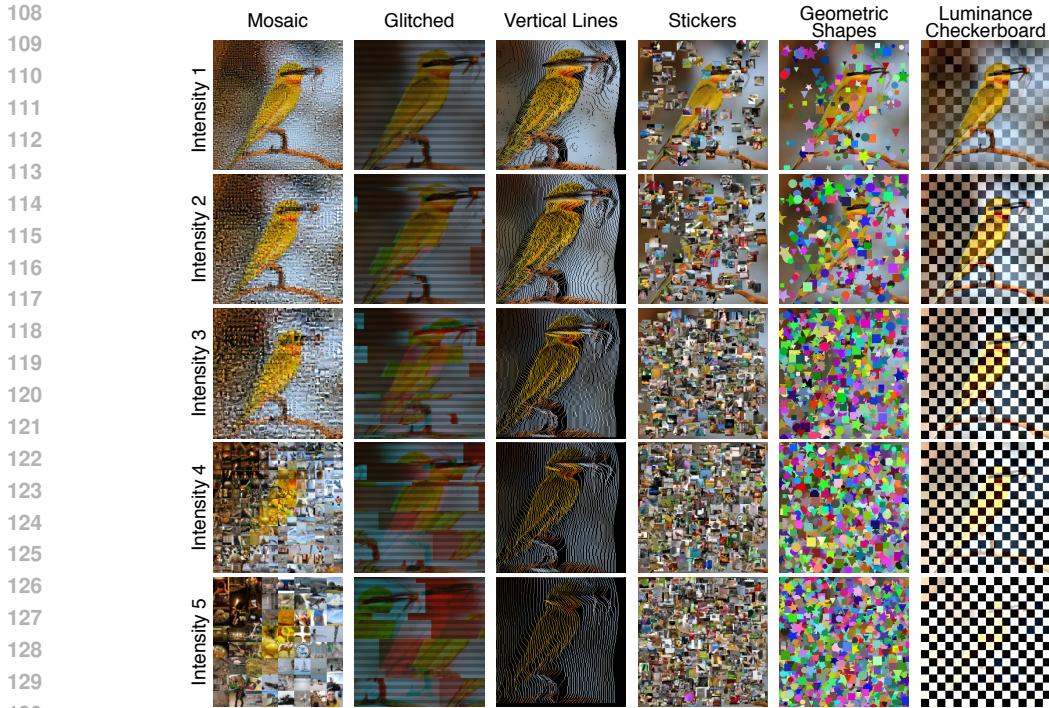


Figure 2: **LAION-C distortions, intended to be OOD even for web-scale datasets.** This figure illustrates the six LAION-C distortions at five intensity levels. Following the standard experimental paradigm from psychophysics, our dataset spans from near-perfect to chance-level difficulties, thoroughly testing models and leaving room for future model improvements. Best viewed on screen.

Distortions The core idea behind our distortions is to intentionally disrupt visual consistency and perceptual cues that models rely on for image classification, such as texture (Geirhos et al., 2019). LAION-C also complements existing benchmarks by including images that challenge human perception, instead of limiting ourselves to visual domains in which humans excel. Following ImageNet-C, each distortion consists of five different *intensity levels* of visual challenges, as illustrated in Fig. 2. In line with established psychophysical experimental paradigms, we calibrate the intensity levels of our distortions to range from near-perfect to chance-level difficulties, ensuring a progressive decline in model accuracy. We propose distortions that disrupt local image structure like texture and edges (Mosaic, Geometric Shapes, Stickers) and global image structure like color and visual consistency (Glitched, Vertical Lines, Luminance Checkerboard). See Appx. A.4 for more detailed descriptions of the distortions.

These distortions are applied to a carefully curated subset of images from the ImageNet validation dataset. To contextualize model performance, we compare it to human performance. Following prior work (Geirhos et al., 2018), we simplify the classification task to 16 superclasses—ball, bird, boat, bottle, butterfly, car & truck, cat, chair, dog, fish, fruit, instrument, primate, snake, timekeeping, and tool—derived from 285 ImageNet classes to ensure human evaluations remain feasible. Each superclass contains 273 images, enabling statistically significant detection of a 3% performance difference. Additionally, we manually filter the dataset to remove ambiguous labels and images requiring specific cultural knowledge for classification.

2.2 MEASURING MODEL PERFORMANCE

We use the generated datasets to evaluate the performance of a suite of 58 vision models. Our selection includes models trained on large-scale web datasets and fine-tuned on ImageNet-1k, such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022), and EVA (Fang et al., 2023; 2024). For comparison, we also evaluate the performance of smaller-scale model families such as ResNet (He et al., 2016) and MobileNet (Howard, 2017) and large-scale models trained

only on ImageNet-1k. Additionally, we also evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on a representative subset of LAION images. See Tab. 6 for a complete list of all models we evaluate.

3 RESULTS

3.1 How OOD is LAION-C?

Next, we empirically evaluate whether our LAION-C dataset is indeed OOD relative to the large-scale image datasets used to train modern vision models. Rigorously quantifying how OOD a test dataset is with respect to some training dataset requires a precise definition of the test and training domain (Mayilvahanan et al., 2024). As the distribution shifts introduced by the distortions of LAION-C and ImageNet-C are fuzzy in nature, we use three tools to compare the OOD-ness of our proposed dataset to the OOD-ness of ImageNet-C. First, we use a qualitative assessment. By searching for the name and related concepts of ImageNet-C corruptions, we easily find LAION samples with visual distortions akin to those of ImageNet-C samples (see Fig. 1).

Second, we use the difficulty of a test dataset (measured by the performance that models trained on a reference dataset yield on the test dataset) as a proxy for how much the test dataset differs from the reference dataset. Here, the reasoning is that if a test dataset can be solved almost perfectly by a model, it means that either the model has great generalization skills or the test dataset is not strictly OOD. If, at the same time, another dataset has much greater difficulty according to the same models, the second dataset is likely more OOD than the first. For the sake of comparability, we here use a version of ImageNet-C restricted to the same 16 superclasses that were used for LAION-C, where we implemented the ImageNet-C augmentations through the code by Michaelis et al. (2019). Indeed, a comparison of the performance achieved by our suite of models (see Fig. 3) suggests that LAION-C is more OOD to LAION than ImageNet-C is.

Third, we use the FID (Heusel et al., 2017; Kynkäänniemi et al., 2022) to quantify the difference between LAION and ImageNet-C and LAION-C, respectively. Specifically, we employ a CLIP-trained ViT-B as feature encoder and use the implementation by Parmar et al. (2022) to calculate FID-scores. In line with previous evidence, we find a FID of ≈ 70 between LAION and LAION-C, which is substantially higher than that between LAION and ImageNet-C (≈ 40). This means that features of LAION are closer to those of ImageNet-C than those of LAION-C, again highlighting the greater OOD-ness of LAION-C. In summary, we have presented three different kinds of evidence suggesting that LAION-C is more OOD than ImageNet-C to LAION.

3.2 MACHINE PERFORMANCE

In Fig. 4, we compare model performance on ImageNet-C against performance on LAION-C. Evidently, the average model performance on ImageNet-C stays above or close to 60%, indicating that current models are increasingly adept at handling the distortions in ImageNet-C. This observation reinforces our hypothesis that the challenge presented by ImageNet-C may no longer be sufficiently difficult to rigorously test the robustness of modern models.

In contrast, models achieve much lower accuracy on LAION-C on average and exhibit more inter-model variability. This showcases our dataset’s ability to uncover nuances that remain hidden on more saturated benchmarks. These performance differences are particularly obvious at higher intensity levels, illustrating LAION-C’s potential to serve as a more challenging and insightful benchmark for evaluating robustness.

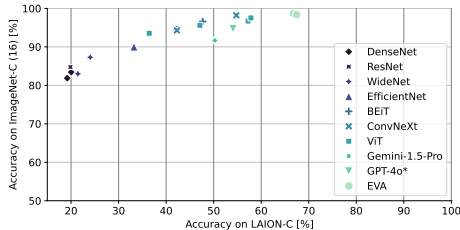
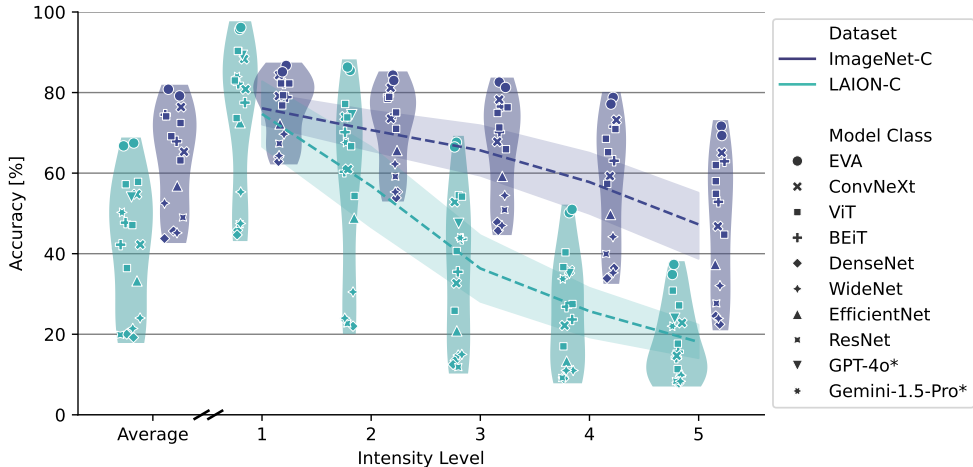


Figure 3: **Performance Divergence of Models on LAION-C and ImageNet-C.** Evaluating models on the 16-class versions of ImageNet-C and LAION-C produces a plateaued performance on ImageNet-C, while LAION-C still yields high variance across models.

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232



233
234
235
236
237

Figure 4: **LAION-C poses a greater challenge to model robustness than ImageNet-C.** We plot distortion intensity against each model’s average accuracy. Visual foundation models evaluated on ImageNet-C maintain high accuracy, with minimal drop across increasing intensity levels. On our LAION-C dataset, the models exhibit a sharper decline in accuracy, highlighting the benchmark’s effectiveness in measuring model robustness.

239
240
241
242

Table 1: **LAION-C is challenging but can be solved by fine-tuning on the exact distortions.** We compare the performance of ViT-H-P14-336-CLIP-LAION-IN12K before and after fine-tuning it on ImageNet-1k training images with LAION-C corruptions. As the performance after fine-tuning is high, this means that LAION-C retains enough signal to remain solvable.

Accuracy [%]	Mosaic	Vertical Lines	Glitched	Luminance	Geometric	Stickers
Before	45.2	51.2	69.8	88.2	64.4	24.6
After	79.0	93.5	95.8	97.7	90.2	61.0

244
245
246
247
248
249
250
251
252
253
254

We also provide a detailed breakdown of results in Fig. 10. We observe significant variability in the performance of different vision models across various datasets and distortion levels. These results further highlight the effectiveness of our datasets in eliciting different responses from models of different architectures, parameter scales, and training data sizes. This diversity is particularly valuable for understanding which model designs are more robust to specific types of distortions, offering insights that are beneficial for advancing the state-of-the-art model robustness.

3.3 IS LAION-C A PROXY FOR MORE REALISTIC DISTRIBUTION SHIFTS?

256
257
258
259
260
261
262
263
264

To demonstrate that model performance on LAION-C is indicative of real-world performance despite the highly synthetic nature of our corruptions, we analyze the correlations between model accuracy on LAION-C and on several well-established OOD benchmark datasets such as ImageNet-R, ImageNet-A and ImageNet-Sketch in Tab. 5. Clearly, models that achieve high accuracy on LAION-C are also robust to other distribution shifts. However, our main goal is not to measure real-world performance, but to measure a model’s ability to generalize beyond its training data, which requires a truly OOD test set - a requirement that might even be incompatible with the requirements of a real-world distribution shift.

3.4 CAN LAION-C BE SOLVED?

265
266
267
268
269

Given the low performance of current state-of-the-art models on LAION-C, one might wonder whether the heavy distortions make classification impossible. To disprove this hypothesis and highlight the validity of LAION-C as a benchmark for evaluating model robustness, we fine-tune a ViT-Huge model (originally pretrained on LAION-2B and later fine-tuned on ImageNet-22k and

270 ImageNet-1k) using 336,000 ImageNet-1K images uniformly augmented with LAION-C distor-
 271 tions.

272 As shown in Tab. 1, fine-tuning the model results in substantial accuracy gains, which define an up-
 273 per bound on LAION-C accuracy that no normal model can be expected to achieve. Notably, these
 274 accuracy gains are particularly pronounced on higher-intensity distortions, as detailed in Tab. 4,
 275 where accuracy is broken down by distortion intensity. The fine-tuned model likely achieves such
 276 good performance by employing un-human-like (or “spurious”) features, but the purpose of this ex-
 277 periment is *not* to suggest that fine-tuning on LAION is a sensible approach to improve OOD robust-
 278 ness, but to quantify how much learnable signal is left. LAION-C provides meaningful robustness
 279 tests without being intractable, making it a valuable tool for a more comprehensive evaluation of
 280 model performance under difficult conditions.

281 282 283 3.5 HUMAN AND MACHINE VISION DISCREPANCY

284
285 In addition to its utility as an OOD-benchmark, LAION-C can be used as a tool to investigate
 286 differences between vision models and human observers. Since LAION-C consists of 16 high-level
 287 classes, we can collect human reference data following the paradigm of Geirhos et al. (2018). For
 288 this, 19 human subjects are briefly presented with a distorted image and are asked to classify it into
 289 one of 16 classes, reminiscent of how a DNN might be evaluated on a classification task. Further
 290 methodological details are provided in Appx. A.2. To quantify the alignment between human and
 291 machine visual perception, we adopt the error consistency metric proposed in Geirhos et al. (2020b),
 292 see Appx. A.3 for details.

293
294 For a detailed analysis of human-machine differences, we refer to Appx. A.6, but to summarize
 295 results, we find that current state-of-the-art vision models now match or even outperform human
 296 observers in OOD scenarios, including on our synthetic distortions, which they likely have never
 297 encountered during training—a stark contrast to just a few years ago, when humans were still vastly
 298 outperforming models (Geirhos et al., 2018; Dodge & Karam, 2019; Taori et al., 2020; Jang &
 299 Tong, 2024). There is a high degree of variability in error consistency between human observers and
 300 different vision models ranging from 0 to 0.4. This indicates that while model families such as ViT
 301 and EVA rival or surpass human performance, they are approaching the task with different strategies
 302 than humans, demonstrating less human-like behaviors. The observed value range matches the one
 303 found in previous work for older models and different image data (Geirhos et al., 2021). This again
 304 suggests that while recent developments have boosted model performance, these models have not
 305 become more human-like, as they follow alternative strategies.

306 307 4 DISCUSSION

308
309
310 Given that existing OOD benchmarks are often no longer OOD for models trained on web-scale
 311 datasets like LAION, we here introduce LAION-C: A novel benchmark designed to evaluate the
 312 robustness and generalization capabilities of modern vision models. We demonstrate that LAION-
 313 C is more challenging for a representative suite of vision models than ImageNet-C, particularly
 314 at higher distortion intensity levels. Given that LAION-C, by virtue of its construction, serves
 315 as a better proxy for a model’s ability to recognize objects despite an unexpected distortion, we
 316 recommend it as an OOD benchmark for current and future web-scale vision models. Furthermore,
 317 our thorough human evaluation shows that *the best models often outperform even the best human*
 318 *observers*. While they do not always follow similar strategies, this reassuring finding indicates that
 319 the best models have indeed substantially progressed in their ability to handle unexpected input
 320 and are not just getting better on in-distribution distortions. This marks a paradigm shift in OOD
 321 generalization: From humans outperforming models to models outperforming humans, from relying
 322 on a single strategy to a diverse set of multiple robust strategies, and from OOD benchmarking
 323 measuring progress towards human-like object recognition to better performance now indicating
 super-human (in other words, *less human-like*) vision models.

324 CODE AND DATASET AVAILABILITY
325

326 We will publicly release the dataset and the code to generate distortions and evaluate models af-
327 ter acceptance of the paper. During the anonymous review period, the code is available from the
328 supplementary material.

329
330 ETHICS STATEMENT

331 We confirm that all experimental procedures involving human subjects in our study had IRB ap-
332 proval. In addition, we ensured that all participants gave informed consent prior to their inclusion in
333 the study. Detailed information was provided to each participant beforehand, outlining the study’s
334 purpose, procedures and benefits, ensuring they were fully informed before agreeing to participate.
335 Participants were also informed that they could abort the study at any time, without incurring any
336 negative consequences. Experimental data and contact information for the participants was stored
337 in accordance with GDPR. Participants were compensated with an hourly base rate of 12 EUR and
338 received bonus payments based on classification performance, as is customary in psychophysical
339 experiments, so that the final reimbursements exceeded the local minimum wage.

340
341 REFERENCES
342

343 Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers.
344 In *ICLR*, 2022.

345 Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are
346 we done with imagenet? *arXiv preprint*, 2020. URL <https://arxiv.org/abs/2006.07159>.

347
348 Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological mea-*
349 *surement*, 1960.

350
351 Samuel Dodge and Lina Karam. Human and dnn classification performance on images with quality
352 distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)*, 2019.

353
354 Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lind-
355 say, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The
356 neuroconnectionist research programme. *Nature Reviews Neuroscience*, 2023.

357 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
358 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszko-
359 reit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at
360 scale. In *ICLR*, 2021.

361 Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong
362 Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale.
363 In *CVPR*, 2023.

364
365 Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A
366 visual representation for neon genesis. *Image and Vision Computing*, 2024.

367
368 Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cyber-*
369 *netics*, 1975.

370 Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach,
371 Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64
372 (12):86–92, 2021.

373 Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and
374 Felix A. Wichmann. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018.

375
376 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and
377 Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias im-
proves accuracy and robustness. In *ICLR*, 2019.

- 378 Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel,
379 Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature*
380 *Machine Intelligence*, 2020a.
- 381 Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-
382 trial behaviour of cnns and humans by measuring error consistency. In *NeurIPS*, 2020b.
- 383 Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge,
384 Felix A. Wichmann, and Wieland Brendel. Partial success in closing the gap between human and
385 machine vision. In *NeurIPS*, 2021.
- 386 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
387 nition. In *CVPR*, 2016.
- 388 Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common
389 corruptions and perturbations. In *ICLR (Poster)*, 2019.
- 390 Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul
391 Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical
392 analysis of out-of-distribution generalization. In *ICCV*, 2021a.
- 393 Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial
394 examples. In *CVPR*, 2021b.
- 395 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
396 Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*,
397 2017.
- 398 Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applica-
399 tions. *arXiv preprint*, 2017. URL <https://arxiv.org/abs/1704.04861>.
- 400 Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected
401 convolutional networks. In *CVPR*, 2017.
- 402 Hojin Jang and Frank Tong. Improved modeling of human vision by incorporating robustness to
403 blur in convolutional neural networks. *Nature Communications*, 15(1):1989, 2024.
- 404 Max Kaufmann, Daniel Kang, Yi Sun, Steven Basart, Xuwang Yin, Mantas Mazeika, Akul Arora,
405 Adam Dziedzic, Franziska Boenisch, Tom Brown, Jacob Steinhardt, and Dan Hendrycks. Testing
406 robustness against unforeseen adversaries. *arXiv preprint*, 2019. URL <https://arxiv.org/abs/1908.08016>.
- 407 Mario Kleiner, David Brainard, Denis Pelli, Allen Ingling, Richard Murray, and Christopher Brou-
408 sard. What’s new in psychtoolbox-3. *Perception*, 2007.
- 409 Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of
410 imagenet classes in fréchet inception distance. *arXiv preprint*, 2022. URL <https://arxiv.org/abs/2203.06026>.
- 411 Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie.
412 A convnet for the 2020s. In *CVPR*, 2022.
- 413 Prasanna Mayilvahanan, Thaddäus Wiedemer, Evgenia Rusak, Matthias Bethge, and Wieland Brendel.
414 Does clip’s generalization performance mainly stem from high train-test similarity? *arXiv*
415 *preprint*, 2023. URL <https://arxiv.org/abs/2310.09562>.
- 416 Prasanna Mayilvahanan, Roland S. Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila
417 Juhos, Matthias Bethge, and Wieland Brendel. In search of forgotten domain generalization. In
418 *ICML 2024 Workshop on Foundation Models in the Wild*, 2024. URL <https://openreview.net/forum?id=Bc2p8T4V32>.
- 419 Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander
420 S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detec-
421 tion: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

- 432 Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and cor-
433 rruptions in natural corruption robustness. *Advances in Neural Information Processing Systems*,
434 34:3571–3583, 2021.
- 435 Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A. Vandermeulen, and Simon Kornblith.
436 Human alignment of neural network representations. In *ICLR*, 2023.
- 437
- 438 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024. Accessed:
439 2024-11-19.
- 440
- 441 Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in
442 GAN evaluation. In *CVPR*, 2022.
- 443 Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling
444 with vector-quantized visual tokenizers. *arXiv preprint*, 2022. URL [https://arxiv.org/
445 abs/2208.06366](https://arxiv.org/abs/2208.06366).
- 446 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
447 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
448 models from natural language supervision. In *ICML*, 2021.
- 449
- 450 Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers
451 generalize to imagenet? In *ICML*, 2019.
- 452 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng
453 Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual
454 recognition challenge. *International journal of computer vision*, 2015.
- 455
- 456 Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Ko-
457 hitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which
458 artificial neural network for object recognition is most brain-like? *BioRxiv*, 2018.
- 459 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
460 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An
461 open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.
- 462 Chase T. Glitch this, 2020. URL [https://github.com/TotallyNotChase/
463 glitch-this](https://github.com/TotallyNotChase/glitch-this). MIT License.
- 464
- 465 Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural
466 networks. In *ICML*, 2019.
- 467
- 468 Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig
469 Schmidt. Measuring robustness to natural distribution shifts in image classification. *NeurIPS*,
2020.
- 470
- 471 Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer,
472 Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal under-
473 standing across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- 474 Felix A Wichmann and Robert Geirhos. Are deep neural networks adequate behavioral models of
475 human visual perception? *Annual Review of Vision Science*, 2023.
- 476
- 477 Ross Wightman. The timm (pytorch image models) leaderboard. [https://huggingface.co/
478 spaces/timm/leaderboard](https://huggingface.co/spaces/timm/leaderboard), 2024.
- 479
- 480 Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training
481 procedure in timm. *arXiv preprint*, 2021. URL <https://arxiv.org/abs/2110.00476>.
- 482
- 483 Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J
484 DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual
485 cortex. *Proceedings of the national academy of sciences*, 2014.
- 486
- 487 Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint*, 2016. URL
<https://arxiv.org/abs/1605.07146>.

486 Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers.
487 In *CVPR*, 2022.
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

A APPENDIX

A.1 RELATED WORK

OOD generalization ability of vision models. As deep learning has advanced to the point where models can reliably generalize to data that matches their training distribution or even exceed the quality of the original labels (Beyer et al., 2020), OOD-robustness, as measured by specifically designed OOD test sets, has moved to the center stage of computer vision research. In particular, ImageNet-C (Hendrycks & Dietterich, 2019), a dataset containing images from the test-set of ImageNet to which various fairly natural corruptions such as blurring and pixelation were applied, has long been the gold standard for assessing OOD-performance, to the point where data augmentations proposed to increase OOD robustness were found to only work well because they are similar to the ImageNet-C corruptions (Mintun et al., 2021). In contrast, ImageNet-R (Hendrycks et al., 2021a) instead provides a more complex distribution shift by collecting different renditions of the target classes such as sculptures and paintings, instead of photos. A more subtle distribution shift which still caused considerable drops in model performance for ImageNet-trained models, was proposed by Recht et al. (2019). They collected ImageNetV2, a new test set for ImageNet that should theoretically not differ from the ImageNet test set at all, because it was collected with a very similar methodology, but revealed that models do perform slightly worse on ImageNetV2 than on the original test set. Hendrycks et al. (2021b) proposed two other OOD-test sets which do not rely on synthetic image manipulations but instead consist of natural images that are in some sense OOD relative to ImageNet, either by virtue of displaying object classes not present in ImageNet (ImageNet-O) or by showing an object of an ImageNet-class in a scene that is weird enough to fool most models (ImageNet-A). What all of these datasets have in common is that, by design, they provide distribution shifts *relative to ImageNet*. But with the rapid evolution of deep learning, these traditional benchmarks have become less challenging for state-of-the-art vision models trained on expansive web-scale datasets (Radford et al., 2021). While it is to some degree possible to predict a model’s OOD generalization from its training set performance (Taori et al., 2020), the only reliable measurements of this capability stem from empirical evaluations of models on OOD datasets. Our work addresses this need by introducing LAION-C, a dataset that incorporates novel and complex synthetic distortions tailored to challenge even advanced vision systems.

Advancement in visual foundation models The rise of visual foundation models such as Vision Transformers (ViT) (Dosovitskiy et al., 2021), ConvNeXt (Liu et al., 2022) and BeiT (Bao et al., 2022) has redefined what constitutes standard performance across many visual tasks. These improvements in performance partially stem from architectural innovations and parameter optimization, but were mostly powered by the effective leveraging of unprecedented dataset sizes (Zhai et al., 2022). However, because visual foundation models were trained on web-scale datasets, the extent of their generalization capability remains underexplored.

Comparing human vs. machine perception. Deep Neural Networks were originally conceived as models of human vision (Fukushima, 1975) and were found to be the best available models for neuronal activity in the primate visual cortex (Yamins et al., 2014), even if not trained for this task. Today, there is a growing body of research dedicated to evaluating the adequacy of neural networks as behavioral models of human core object recognition (Doerig et al., 2023; Schrimpf et al., 2018; Wichmann & Geirhos, 2023; Muttenthaler et al., 2023). Building upon the findings of Geirhos et al. (2021), who illustrate the narrowing of the behavioral gap between humans and machines in terms of their error consistency, our study further explores this dynamic utilizing LAION-C. We conducted a systematic analysis of differences in perception between human and machine observers, and assessed if the behavioral gap is closing further, as well as highlighting the persistent cognitive differences between humans and machines.

A.2 EXPERIMENT PROCEDURE AND PARTICIPANT INCENTIVES

Participant recruitment and setup. We recruited 20 participants (10 female) from the university student body via mailing lists. All participants were screened to ensure normal or corrected vision and no color blindness, and gave informed consent to participate. One participant was excluded post-hoc due to reporting extreme tiredness. Our experiments were conducted in a darkened cabin,

594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647



Figure 5: **Interface presented to participants.** This figure illustrates the icon layout as displayed to participants during the study. The grid is adapted from (Geirhos et al., 2018), while most of the categories and therefore symbols are different.

using a 22” VIEWPixx 3D light LCD monitor (VPixx Technologies, Saint-Bruno, Canada) at a refresh rate of 120 Hz (scanning backlight mode on). The screen measures 484×302 mm, at a resolution of 1920×1200 pixels. Stimuli were presented foveally in the center of the screen, with a viewing distance of 65 cm, resulting in 5° of visual angle. In line with earlier experiments, the background was set to a grey value of 0.454 in the $[0, 1]$ range. A chin rest was used to maintain a fixed viewing distance and angle. The experiment was implemented using the Psychophysics Toolbox (Kleiner et al., 2007, version 3.0.12) in MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States) using a 12-core desktop computer (AMD HD7970 graphics card “Tahiti” by AMD, Sunnyvale, California, United States) running Kubuntu 14.04 LTS.

The entire classification task, including both the warm-up and main experiment phases, was organized into 12 blocks. In each block, participants were shown images from the 16 superclasses for 2.5 seconds—a duration empirically determined to balance efficient overall experiment length with sufficient exposure time allowing for viewing and consideration time. After each image, the 16 corresponding class icons were displayed on screen, allowing participants 2 seconds to classify each image into one of these categories. The icons were organized in a layout that roughly grouped them by size and general category (e.g., animals and objects), as illustrated in image Fig. 5. To encourage responses rather than leaving selections blank, a message was displayed at the top of the screen 0.75 second before icon display time ended, prompting participants to make a choice. At the end of each block, if a participant surpassed the 90% accuracy threshold calibrated using internal baseline performance data, they received an encouraging on-screen message (“Congratulations! You just earned some extra money!”) along with a \$0.50 bonus per block to incentivize higher performance.

Warm-up session and main experiment. The experiment began with a 10-minute warm-up session, allowing participants to familiarize themselves with the icon layouts and the classification task procedure through two practice blocks. Each practice block contained 45 images, with one block consisting of clean images and the other of augmented images. To avoid test-time adaptation, the augmentations used during the warm-up phase differed from those in the actual trials. The images used for the practice trials were also a subsample of the ImageNet validation dataset, but did not overlap with those used in the main experiment.

Following the warm-up, the main experiment proceeded consisting of 10 blocks, each block comprising 60 images. Each set of 5 blocks was augmented using a consistent method, with a balanced distribution across different intensity levels and superclasses. To avoid fatigue, participants were allowed an unlimited amount of time to rest between blocks and encouraged to rest their eyes or accommodate elsewhere.

A.3 ERROR CONSISTENCY

Here, we provide a more detailed explanation of error consistency (EC), summarizing Geirhos et al. (2020b). The EC between two observers which both classified a sequence of samples is obtained by first using the necessary ground-truth labels to assess which images each observer classified correctly. A trial increases the agreement if both observers solved it correctly, or if they both failed (and decreases it if only one of them got the trial correct while the other one failed). One then calculates how much more agreement was observed between the two observers relative to the agreement expected by chance. This is done by calculating Cohen’s Kappa (Cohen, 1960), which is defined as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the (empirically measured) proportion of agreement-trials and p_e is the (theoretical) expected agreement when modeling both observers as independent binomial observers. κ takes on values between -1 and 1 , with a higher κ signifying higher levels of agreement, and a κ of 0 indicating that a pair of observers does not agree more frequently than one would expect by chance, given their marginal correctness probabilities.

In this work, we calculate the error consistency between model responses and human classification data. To do this, we first collect all human responses. Since each human participant saw a fresh set of stimuli, we thus obtain exactly one human response per image. We then calculate each model’s EC to this list of human responses.

A.4 AUGMENTATION DESIGNS

- **Mosaic:** The original image is broken down into smaller tiles, each replaced by a chromatically similar picture. This patchwork creates a mosaic effect that disrupts edges and textures while introducing contextually irrelevant information.
- **Glitched:** The original image undergoes an artistic digital corruption with horizontal lines overlaying shifted image segments and color channel shifts. This dislocates the global contextual structure of the image. While the concept of such glitchy images has been explored in earlier work (Kaufmann et al., 2019), our transformation introduces a more intense corruption.
- **Vertical Lines:** The original image is deconstructed into bent vertical line segments. This distortion retains the original colors but strips away local information, disrupting the contours and edges of the image and introducing visual discontinuity.
- **Geometric Shapes:** The original image is overlaid with overlapping geometric figures such as squares, circles, and stars. This visual clutter introduces local noise that obscures the main object, like the Kaleidoscope corruption from Kaufmann et al. (2019).
- **Stickers:** The original image is augmented with assorted image patches. This addition of visual elements masks features of the primary object.
- **Luminance Checkerboard:** The original image is divided into a grid, with the luminance of each cell altered in a checkerboard pattern. The stark luminance contrast between adjacent tiles and artificial grid boundaries makes this distortion challenging.
- **Mosaic:** The image is divided into an $n \times n$ grid, where each tile is replaced by a patch from the ImageNet validation set whose average color best matches that of the tile. The values of n per intensity level are:
 - Level 1: $n = 4$
 - Level 2: $n = 6$
 - Level 3: $n = 8$
 - Level 4: $n = 16$
 - Level 5: $n = 28$
- **Glitched:** Alternating rows are replaced with black pixels to create a scan line effect. Pixel shifts and color channel offsets are applied to random regions as follows:
 - Level 1: Shift up to 8% of image width in 4 regions, ± 4 pixel channel offset.
 - Level 2: Shift up to 32% of image width in 8 regions, ± 8 pixel channel offset.
 - Level 3: Shift up to 50% of image width in 10 regions, ± 10 pixel channel offset.
 - Level 4: Shift up to 128% of image width in 16 regions, ± 16 pixel channel offset.

Table 2: **Occlusion ratio of objects in Stickers and Geometric Shapes distortions.** We calculated the object occlusion ratio for the Stickers and Geometric Shapes corruptions as an additional quantitative measurement of the distortion strength.

Intensity Level	Geometric Shapes (%)	Stickers (%)
1	61.88	65.83
2	72.51	76.52
3	85.35	86.19
4	90.16	89.54
5	93.21	91.63

- Level 5: Shift up to 200% of image width in 20 regions, ± 20 pixel channel offset.

The implementation is inspired by T (2020)

- **Vertical Lines:** The image is divided into vertical sections, each of which is further subdivided into smaller sections along the y-axis (called y-steps). A vertical line is drawn within each y-step with a slight x-offset based on the intensity level. The line color is determined by the average color of that section. The parameters for each intensity level are:
 - Level 1: 224 vertical sections, with 1-pixel steps along the y-axis.
 - Level 2: 178 vertical sections, with 2-pixel steps along the y-axis.
 - Level 3: 112 vertical sections, with 4-pixel steps along the y-axis.
 - Level 4: 84 vertical sections, with 6-pixel steps along the y-axis.
 - Level 5: 60 vertical sections, with 8-pixel steps along the y-axis.
- **Luminance Checkerboard:** The image is divided into a 14×14 grid, and the luminance of each tile is altered in a checkerboard pattern. The luminance variation per intensity level is:
 - Level 1: ± 50 .
 - Level 2: ± 50 –100.
 - Level 3: ± 100 –125.
 - Level 4: ± 125 –150.
 - Level 5: ± 150 –255.
- **Stickers:** 16×16 pixel image patches from the ImageNet validation set are randomly placed with uniform probability across the image. The number of patches per intensity level is:
 - Level 1: 100 patches.
 - Level 2: 200 patches.
 - Level 3: 400 patches.
 - Level 4: 600 patches.
 - Level 5: 1200 patches.

For an estimate of the occlusion ration of the objects per intensity level, see Tab. 2.

- **Geometric Shapes:** Random geometric shapes (triangle, square, star, circle) of varied colors and sizes are introduced. The number of shapes per intensity level is:
 - Level 1: 150 shapes.
 - Level 2: 300 shapes.
 - Level 3: 600 shapes.
 - Level 4: 800 shapes.
 - Level 5: 1000 shapes.

For an estimate of the occlusion ration of the objects per intensity level, see Tab. 2.

Table 3: **LAION-C benchmark results.** Numbers show the top-1 accuracy in percent. *ImageNet* refers to model accuracy on the (uncorrupted) ImageNet validation set, with values sourced from the timm leaderboard (Wightman, 2024). For each corruption, we report the mean top-1 accuracy across all intensity levels, with *LAION-C* as the overall benchmark metric (averaged across corruption types). GPT-4o and Gemini 1.5 Pro were evaluated on 48,000 images, 100 for each class. For full model names and descriptions, see Tab. 6 in the Appendix.

Model	ImageNet	LAION-C	Mosaic	Vertical	Glitched	Luminance	Geometric	Stickers
EVA-G-P14-560-M30M-IN22K	89.8	67.5	48.8	53.6	70.8	97.2	81.0	53.4
EVA02-L-P14-448-MIM-M38M-IN22K	90.1	66.8	53.6	58.2	78.2	93.6	76.4	40.6
ViT-H-P14-336-CLIP-LAION-IN12K	88.6	57.3	45.2	51.2	69.8	88.2	64.4	24.6
ViT-L-P14-224-CLIP-OpenAI-IN12K	88.3	57.8	52.6	49.8	68.2	98.6	55.4	22.4
ViT-B-P32-384-CLIP-LAION-IN12K	85.4	36.4	36.8	35.2	35.8	54.0	37.6	19.2
ViT-B-P16-224-AugReg-IN21K	85.5	47.1	46.4	42.8	62.0	71.4	42.4	17.6
BEiT-v2-L-P16-224-IN1K	87.4	47.7	52.4	44.8	63.2	70.2	11.8	43.8
BEiT-v2-B-P16-224-IN1K	85.6	42.2	46.2	40.4	52.6	68.2	11.4	34.6
ConvNeXt-XXL-CLIP-LAION-IN1K	88.6	54.8	53.0	53.4	71.8	77.4	52.2	20.8
ConvNeXt-B-CLIP-LAION-AugReg-IN12K	87.6	42.3	37.6	43.8	44.4	54.2	50.4	23.2
WRN101-2-TV-IN1K	78.8	21.4	30.4	28.4	22.0	22.8	18.2	6.8
WRN50-2-RACM-IN1K	81.5	24.0	26.8	21.4	17.0	45.0	24.6	9.4
RN50-A1-IN1K	81.2	19.9	28.0	18.8	20.8	23.4	21.2	7.0
EFF-B3-RA2-IN1K	82.3	33.2	32.4	31.8	40.2	45.2	37.6	12.2
DN201-TV-IN1K	77.3	19.2	28.6	26.2	13.2	23.2	16.8	7.2
DN161-TV-IN1K	77.3	20.0	31.0	26.8	15.2	25.2	15.4	6.6
GPT-4o	-	54.1	42.8	45.4	65.1	80.1	54.2	36.5
Gemini 1.5 Pro	-	50.2	34.9	37.0	46.2	84.4	60.9	38.1
Best human observer	-	55.2	58.0	55.3	78.7	63.4	40.4	35.7
Average human observer	-	47.0	50.8	43.6	71.0	53.1	34.3	29.1

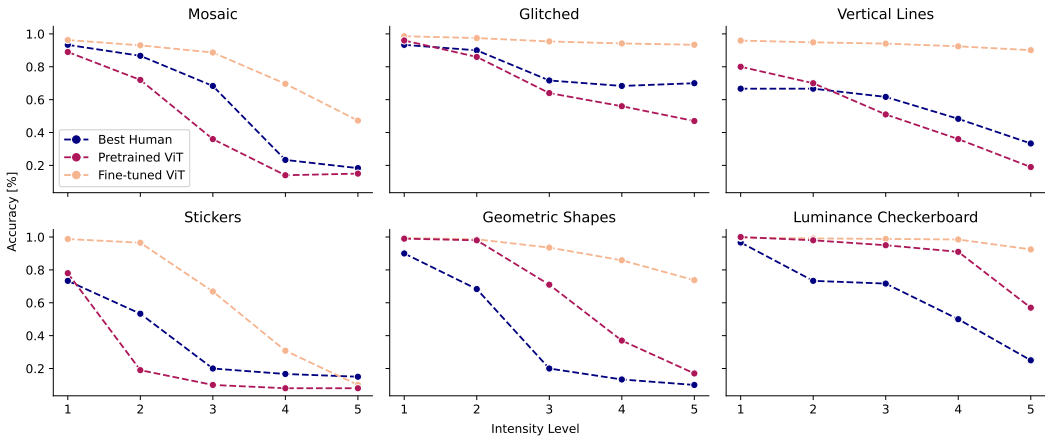


Figure 6: **LAION-C can be solved.** For every distortion, we plot the accuracy of our reference model (ViT-H-P14-336-CLIP-LAION-IN12K) before and after fine-tuning, in comparison to the best human participant for reference. Most distortions can be learned perfectly, only the Stickers and Mosaic distortions might have been too difficult at the highest intensity levels. Further performance gains might be possible with more careful fine-tuning.

A.5 ACCURACY

In Tab. 3, we present a comprehensive evaluation of our suite of models on LAION-C. We report each model’s top-1 accuracy on the (undistorted) ImageNet validation set as a baseline (*Clean Accuracy*) and the average top-1 accuracy on LAION-C averaged across distortion types and intensity levels (*LAION-C*). We then break the latter down into the six distortion types. This enables a thorough comparison of model performance, highlighting which architectures generalize best.

Table 4: **Model (ViT) Accuracy Before and After Fine-Tuning on LAION-C.** The high accuracies after fine-tuning indicate that even though the dataset is challenging, there is, in principle, enough signal left to perform well on LAION-C.

	Intensity Level	Accuracy Before (%)	Accuracy After (%)
Mosaic	1	89.0	96.3
	2	71.9	93.0
	3	35.8	88.7
	4	14.3	69.6
	5	14.7	47.7
Vertical Lines	1	79.9	95.9
	2	70.1	94.9
	3	50.8	94.1
	4	36.1	92.4
	5	19.4	90.0
Glitched	1	95.9	98.6
	2	86.2	97.5
	3	63.6	95.4
	4	55.6	94.2
	5	47.1	93.4
Luminance Checkerboard	1	99.7	99.6
	2	98.4	99.2
	3	95.1	98.8
	4	90.7	98.5
	5	56.6	92.5
Geometric Shapes	1	30.9	99.4
	2	11.2	98.6
	3	6.7	93.6
	4	6.6	85.9
	5	6.3	73.7
Sticker	1	97.3	98.8
	2	77.8	96.5
	3	28.7	63.7
	4	14.9	31.8
	5	8.1	14.3

To demonstrate the value of LAION-C as a benchmark for evaluating model robustness, we analyze how model performance on LAION-C correlates with that on ImageNet-C. Grounding our comparison in models that have demonstrated a baseline level of robustness on well-established benchmarks, we apply a threshold to include 40 models that achieved at least 60% accuracy on ImageNet-C.

As shown in Fig. 9, the majority of data points lie above the identity line representing performance alignment on LAION-C and ImageNet-C. The gradual slope of the data points, combined with their positioning, indicates that models generally perform better on ImageNet-C, while their performance on LAION-C is more dispersed and often substantially lower.

This broader distribution of performance highlights that LAION-C introduces more challenging distortions, prompting models to exhibit greater variability in robustness. The moderate Kendall’s tau coefficient ($\tau = 0.66$) between the models’ performances on LAION-C and ImageNet-C further underscores this, indicating notable pairwise differences in how models rank across these two benchmarks, unearthing vulnerabilities that are less pronounced on ImageNet-C. These results demonstrate the necessity of LAION-C as a complementary benchmark for a more comprehensive evaluation of model robustness.

A.6 DETAILS ON HUMAN AND MACHINE VISION DISCREPANCY

Occlusion and Luminance Manipulations. For distortions involving occlusions, such as Stickers and Geometric Shapes, models usually match or exceed human performance (see second row of Fig. 8). One possible hypothesis is that this can be attributed to the robustness that models develop after e.g., masked image modeling (MIM) (Fang et al., 2023; 2024). The fact that models perform so much better than humans on partially occluded images implies that models use different features

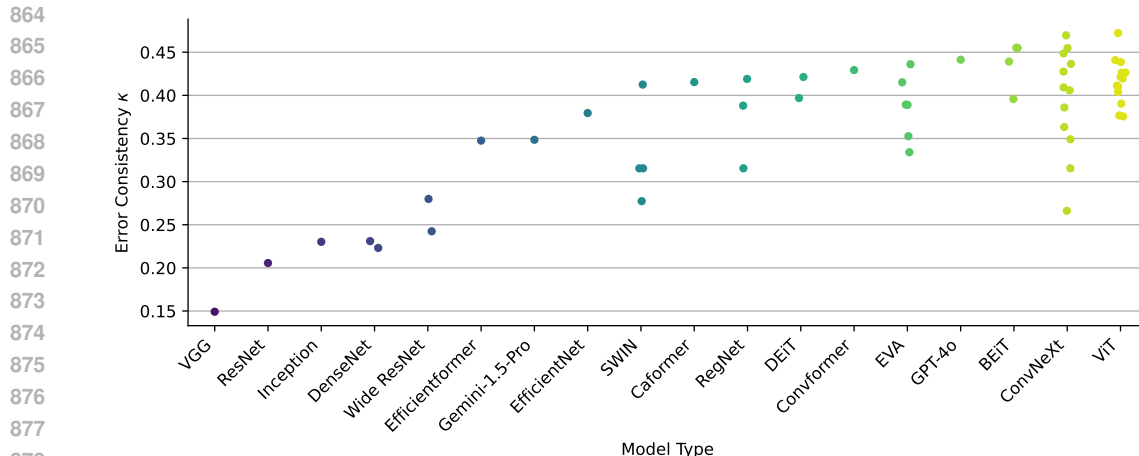


Figure 7: **Humans and models make different mistakes.** We analyze the agreement of error patterns between different families of vision models (see Tab. 6 for a complete list) and human observers. The error consistency (κ) could theoretically achieve a maximum value of 1, but in line with earlier work (Geirhos et al., 2021), the EC values range between 0 and 0.4, indicating that behavioral differences between humans and machines are still quite large. Marker colors encode model families.

than humans. For example, for the Stickers distortion, certain ViT models outperform humans, likely due to their ability to focus on those parts of the image background that remain visible despite the occlusions. As shown in Fig. 1, the stickers occlude nearly the entire image on higher intensity levels, and little to no meaningful object information is retained. Nevertheless, certain models are still able to correctly classify the image based on subtle background cues. This indicates that while models are performing well, they may be doing so by leveraging unintended shortcuts (Geirhos et al., 2020a), such as exploiting background information, when faced with severely occluded images. For the Luminance Checkerboard distortion, we observe that models from the ViT and EVA families outperform humans by a large margin. This advantage could potentially stem from their architectural features, such as self-attention mechanisms and patch-based processing (Fang et al., 2023; Dosovitskiy et al., 2021), which enable them to extract meaningful information from both light and dark regions independently, as well as handle subtle luminance variations. These capabilities give them a clear edge over humans and older models.

Performance on Complex Distortions. When analyzing more complex distortions such as Mosaic, Vertical Lines, and Glitched images (first row of Fig. 8), we observe that human performance is generally on par with the best-performing models. Especially at greater intensity levels, humans perform competitively, e.g., outperforming all models for the strongest Vertical Lines distortions. As we show in Fig. 10, the gap between humans and older models like the ResNet variants is particularly large on these complex distortions. However, modern model classes demonstrate substantial progress, approaching human-level performance even at higher intensity levels. While some margin for improvement remains, the narrowing gap suggests that achieving human-level robustness on classification tasks is no longer the primary challenge for state-of-the-art architectures.

A.7 BREAKDOWN OF MODEL PERFORMANCE

Evaluating VLMs To evaluate GPT-4o (OpenAI, 2024) and Gemini 1.5 Pro (Team et al., 2024) on LAION-C, we decided to test a random subsample of the full dataset, consisting of 100 images per category, which were then tested on all corruptions and intensity levels, resulting in a total of 48,000 images. For ImageNet-C, we limited ourselves to only 10 images per class, to get an initial ballpark estimate of performance.

We employed the following system prompt, in line with our human experiments, during which participants were also shown examples:

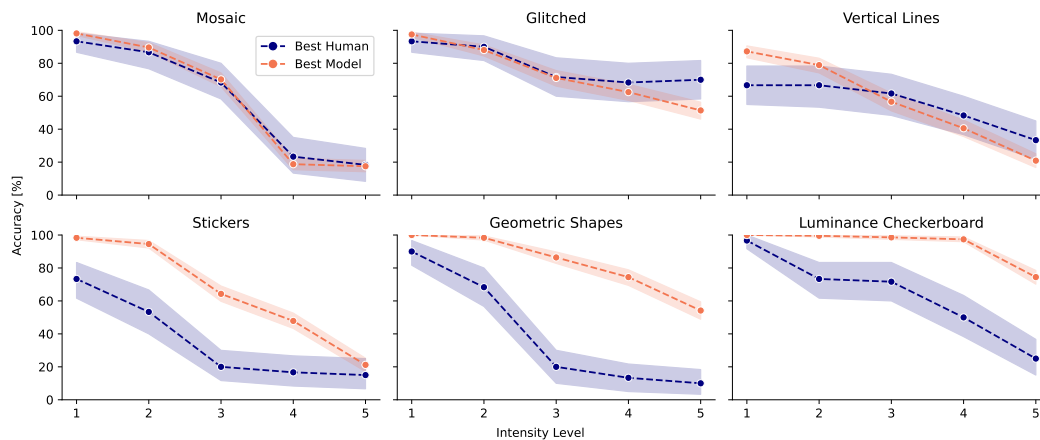


Figure 8: **Human vs. machine accuracy on all distortions.** For each LAION-C distortion, we plot the distortion intensity against the accuracy of the best human and the best model in this condition. The shaded regions indicate the 95% confidence intervals around the means. On the Mosaic, Glitched and Vertical Lines distortions, humans and machines perform similarly, whereas the best model vastly outperforms the best human observer on the Stickers, Geometric Shapes, and Luminance Checkerboard distortions.

System Prompt:

You are an image-recognition API.

You are always asked to classify the main object of images into one of 16 mutually exclusive categories.

In some images, the distortion may be so strong that you might not recognize anything. If you're unsure, provide your best guess - you always have to pick exactly one of the 16 categories.

The 16 categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.

Here is a list of characterizations of every such category:

primate: a primate, like e.g. monkeys, chimpanzees, Orang-Utans etc.

dog: a dog, like e.g. german shepherd, labrador, golden retriever etc.

cat: a cat, like e.g. domestic cat, lion, cheetah etc.

bird: a bird, like e.g. songbird, eagle, chicken etc.

fish: a fish, like e.g. trout, shark, whale etc.

snake: a snake, like e.g. viper, cobra, seasnake etc.

butterfly: a butterfly, like e.g. monarch, cabbage butterfly, ringlet etc.

fruit: a fruit, like e.g. apple, orange, pineapple etc.

boat: a boat, like e.g. ship, gondola, fireboat etc.

vehicle: a vehicle, like e.g. truck, van, sports car etc.

chair: a chair, like e.g. bench, throne, couch etc.

ball: a ball (or a person playing with a ball), like e.g. soccer ball, football, tennis ball etc.

bottle: a bottle, like e.g. water bottle, jug, pill bottle etc.

instrument: a musical instrument (or a person playing an instrument), like e.g. sax, flute, harp etc.

timekeeper: a timekeeper, like e.g. clock, watch, sundial etc.

tool: a tool (or a person using a tool), like e.g. hammer, power drill, chainsaw etc.

Since you are an API, you always respond with minimal messages that contain exactly one word, which is the category name.

User Prompt:

What is the main object in this image? Categories are: primate, dog, cat, bird, fish, snake, butterfly, fruit, boat, vehicle, chair, ball, bottle, instrument, timekeeper, tool.

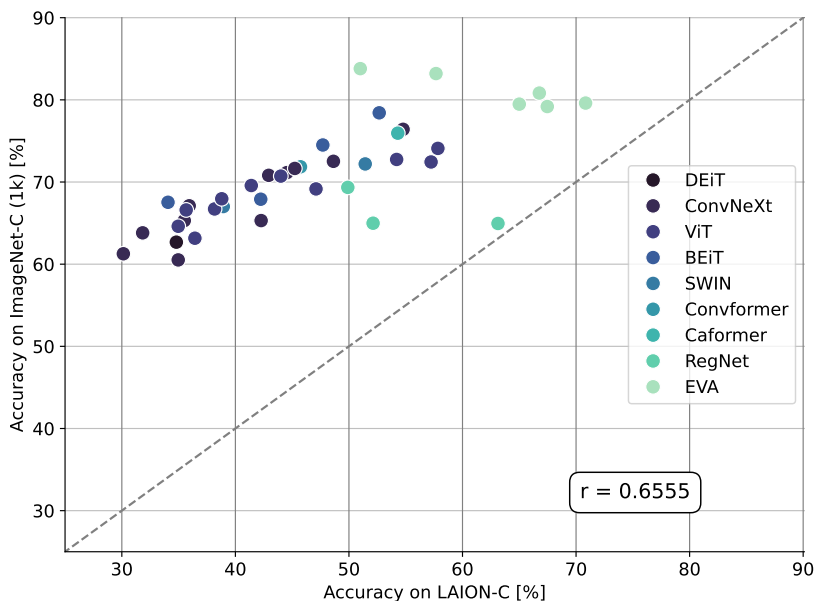


Figure 9: **Performance Divergence of Models on LAION-C and ImageNet-C (1k classes).** The figure illustrates the scattered performance of models across the ImageNet-C and LAION-C dataset, where a Kendall’s tau coefficient (τ) of 0.66 and the shallow slope indicate a dispersed performance on LAION-C. To provide a clearer trend and to better visualize the dispersion, we supplement the suite of models with additional top-performing models sourced from the timm leaderboard (Wightman, 2024), bringing the total number of models to 40 (see Tab. 6 for a complete list).

Table 5: **Model performance on LAION-C correlates with other OOD benchmarks.** We evaluated a suite of 18 models (ViT and ConvNeXt variants trained on either LAION-2B or ImageNet) on IN-C, IN-A, IN-R, IN-Sketch and IN-Val. Evidently, the correlations between all of these OOD benchmarks are high, indicating that they measure related quantities.

	IN-C	LAION-C	IN-A	IN-R	IN-Sketch	IN-val
IN-C	1.00	0.86	0.88	0.91	0.86	0.90
LAION-C	0.86	1.00	0.69	0.70	0.81	0.72
IN-A	0.88	0.69	1.00	0.99	0.94	1.00
IN-R	0.91	0.70	0.99	1.00	0.93	0.99
IN-Sketch	0.86	0.81	0.94	0.93	1.00	0.95
IN-val	0.90	0.72	1.00	0.99	0.95	1.00

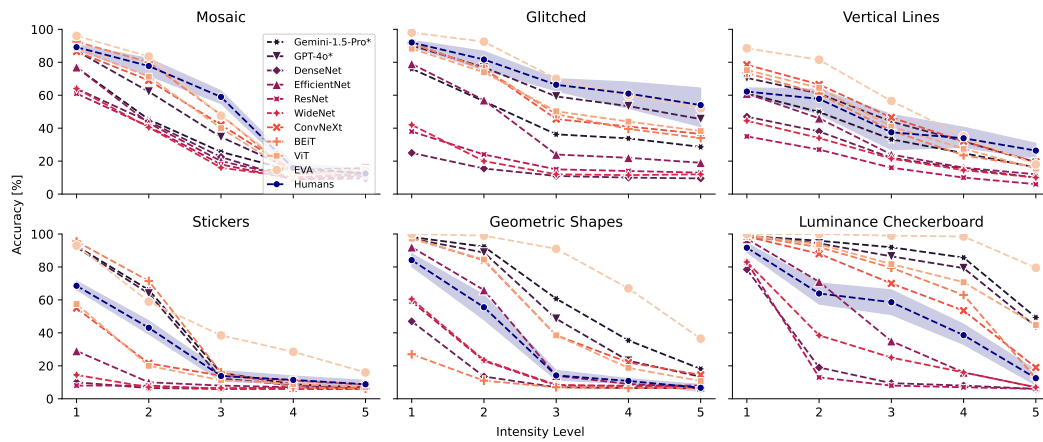


Figure 10: **Model performance on LAION-C.** Analogous to Fig. 8, we relate distortion intensity level to classification accuracy for the different distortions, showing the different models individually. The shaded region around human performance corresponds to the 95% confidence interval, which we omit for the models for better visibility.

A.8 MODELS

Table 6: **Model overview.** For each model used in our evaluation, we show the full model names, as used in timm, an abbreviated name used in the main text and a description of the model. While the first 16 models were used in all analyses and figures, the rest was only used for selective analyses such as Fig. 7.

Abbreviation	Full Model Name	Description
EVA-G-P14-560-M30M-IN22K	eva_giant_patch14_560.m30m_ft_in22k_in1k	EVA giant model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-30M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2023).
EVA02-L-P14-448-MIM-M38M-IN22K	eva02_large_patch14_448.mim_m38m_ft_in22k_in1k	EVA02 large model, patch size 14, pre-trained with masked image modeling (MIM) on a Merged-38M dataset, fine-tuned on ImageNet-22k and ImageNet-1k (Fang et al., 2024).
VIT-H-P14-336-CLIP-LAION-IN12K	vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k	Vision Transformer (ViT) huge model, patch size 14, pre-trained on LAION-2B dataset using OpenCLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-L-P14-224-CLIP-OPENAI-IN12K	vit_large_patch14_clip_224.openai_ft_in12k_in1k	Vision Transformer large model, patch size 14, pre-trained on WIT-400M using CLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-B-P32-384-CLIP-LAION-IN12K	vit_base_patch32_clip_384.laion2b_ft_in12k_in1k	Vision Transformer base model, patch size 32, pretrained on LAION-2B using OpenCLIP, fine-tuned on ImageNet-12k and ImageNet-1k (Dosovitskiy et al., 2021).
VIT-B-P16-224-AUGREG-IN21K	vit_base_patch16_224.augreg2_in21k_ft_in1k	Vision Transformer base model, patch size 16, trained on ImageNet-21k and fine-tuned on ImageNet-1k (Dosovitskiy et al., 2021).
BEITV2-L-P16-224-IN1K	beitv2_large_patch16_224_in1k_ft_in1k	BEiTV2 large model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022).
BEITV2-B-P16-224-IN1K	beitv2_base_patch16_224_in1k_ft_in1k	BEiTV2 base model, patch size 16, trained on ImageNet-1k, fine-tuned on ImageNet-22k and ImageNet-1k (Bao et al., 2022; Peng et al., 2022).
CONV-XXL-CLIP-LAION-IN1K	convnext_xlarge_clip_laion2b_soup_ft_in1k	ConvNeXt xlarge model, pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-1k (Liu et al., 2022).
CONV-B-CLIP-LAION-AUGREG-IN12K	convnext_base_clip_laion2b_augreg_ft_in12k_in1k_384	ConvNeXt base model, pre-trained using OpenCLIP on LAION-2B, fine-tuned on ImageNet-12k and ImageNet-1k (Liu et al., 2022).
WRN101-2-TV-IN1K	wide_resnet101_2_tv_in1k	Wide ResNet-101 model, trained on ImageNet-1k, with original torchvision model weight (He et al., 2016; Zagoruyko & Komodakis, 2016).
WRN50-2-RACM-IN1K	wide_resnet50_2_racm_in1k	Wide ResNet-50 model, trained with RandAugment RACM recipe on ImageNet-1k (He et al., 2016; Zagoruyko & Komodakis, 2016).
RN50-A1-IN1K	resnet50_a1_in1k	ResNet-50 model trained on ImageNet-1k (He et al., 2016; Wightman et al., 2021).
EFF-B3-RA2-IN1K	efficientnet_b3_ra2_in1k	EfficientNet-B3 model, trained with RandAugment RA2 recipe on ImageNet-1k (Tan & Le, 2019).
DN201-TV-IN1K	densenet201_tv_in1k	DenseNet-201, DenseNet pre-trained on ImageNet-1k (Huang et al., 2017).
DN161-TV-IN1K	densenet161_tv_in1k	DenseNet-161, DenseNet model pre-trained on ImageNet-1k (Huang et al., 2017).
GPT-4o	gpt-4o-2024-08-06	At the time of writing, the most recent snapshot of OpenAI's flagship model (OpenAI, 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples.
Gemini-1.5-Pro	gemini-1.5-pro-002	At the time of writing, the most recent stable version of Google's Gemini model (Team et al., 2024). Only evaluated on 48,000 LAION-C samples and 12,000 ImageNet-C samples.
	convnextv2_pico_fcmae_ft_in1k	
	convnextv2_tiny_fcmae_ft_in22k_in1k	
	convnext_base_fb_in22k_ft_in1k	
	convnext_large_mlp_clip_laion2b_augreg_ft_in1k_384	
	convnext_large_mlp_clip_laion2b_soup_ft_in12k_in1k_384	
	convnext_tiny_in12k_ft_in1k	
	convnext_small_fb_in22k_ft_in1k_384	
	convnext_xlarge_fb_in22k_ft_in1k	
	convnext_small_in12k_ft_in1k_384	
	convnextv2_large_fcmae_ft_in22k_in1k_384	
	vit_betwixt_patch16_reg4_gap_256_sbb2_e200_in12k_ft_in1k	
	vit_medium_patch16_rope_reg1_gap_256_sbb_in1k	
	vit_wee_patch16_reg1_gap_256_sbb_in1k	
	vit_medium_patch16_reg4_gap_256_sbb2_e200_in12k_ft_in1k	
	vit_medium_patch16_reg4_gap_256_sbb_in12k	
	vit_pwee_patch16_reg1_gap_256_sbb_in1k	
	vit_betwixt_patch16_rope_reg4_gap_256_sbb_in1k	
	vit_betwixt_patch16_reg4_gap_256_sbb_in12k_ft_in1k	
	maxxvitv2_rmlp_base_rw_384_sw_in12k_ft_in1k	
	vgg19_bn_tv_in1k	
	regnety_1280_swag_lc_in1k	
	regnety_1280_swag_ft_in1k	
	regnety_320_swag_ft_in1k	
	inception_v3_ifadv_in1k	
	beit_base_patch16_224_in22k_ft_in22k_in1k	
	beit_large_patch16_512_in22k_ft_in22k_in1k	
	deit3_large_patch16_384_fb_in22k_ft_in1k	
	deit_base_distilled_patch16_224_fb_in1k	
	swin_base_patch4_window7_224_ms_in22k_ft_in1k	
	swinv2_base_window12to24_192to384_ms_in22k_ft_in1k	
	swinv2_large_window12to24_192to384_ms_in22k_ft_in1k	
	eva_large_patch14_336_in22k_ft_in1k	
	convformer_b36_sail_in22k_ft_in1k_384	
	caformer_b36_sail_in22k_ft_in1k_384	
	efficientformerv2_s2_snap_dist_in1k	

1134 A.9 DATASHEET FOR LAION-C

1135

1136 As proposed by one of our anonymous reviewers, we here include a Datasheet for LAION-C fol-
 1137 lowing the template proposed by [Gebu et al. \(2021\)](#).

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The LAION-C dataset was created to serve as a benchmark for evaluating the robustness and Out-of-Distribution (OOD) generalization of large-scale vision models. It can also be used to study the difference between human and model perception.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Information will be provided upon publication.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Information will be provided upon publication.

Any other comments?

None.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the LAION-C dataset represent images grouped into 16 superclasses with various synthetic distortions applied to them at 5 severity levels. Each superclass contains 273 images, and the distortions include mosaic effects, glitched images, vertical lines, geometric shapes, stickers, and luminance checkerboard patterns.

How many instances are there in total (of each type, if appropriate)?

In total, LAION-C consists of 131.040 images. (16 classes \times 273 images \times 6 corruptions \times 5 severity levels.)

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated / verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of the ImageNet validation set and only contains 4,368 of the 50,000 images. As such, LAION-C is not representative of ImageNet, because it only consists of coarse superclasses. This decision was made to facilitate measuring human classification performance on LAION-C, which would not be possible with the fine-grained classes of ImageNet.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of an RGB image, as well as metadata about the ground-truth class, corruption type, and severity level, which are simply part of the filename.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is labeled with its superclass (one of 16) and can be traced back to its original ImageNet class label.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing from individual instances as each image in the dataset is synthetically altered and labeled with the type of dis-

1188 tortion and its severity, ensuring comprehensive
1189 data for evaluation purposes.

1191 **Are relationships between individual in-**
1192 **stances made explicit (e.g., users' movie**
1193 **ratings, social network links)?** If so,
1194 please describe how these relationships are
1195 made explicit.

1196 The dataset does not contain explicit relation-
1197 ships between individual instances such as so-
1198 cial links or ratings since it primarily focuses on
1199 image recognition and distortion type evaluation
1200 without any relational context between the im-
1201 ages.

1203 **Are there recommended data splits (e.g.,**
1204 **training, development / validation, test-**
1205 **ing)?** If so, please provide a description of
1206 these splits, explaining the rationale behind
1207 them.

1208 Since the dataset is primarily used for bench-
1209 marking purposes, splitting specifics are not pro-
1210 vided. Essentially, the entire dataset is a valida-
1211 tion set.

1213 **Are there any errors, sources of noise, or**
1214 **redundancies in the dataset?** If so, please
1215 provide a description.

1216 The dataset is designed to introduce controlled
1217 noise through synthetic distortions to test model
1218 robustness. There are no unintentional errors
1219 or redundancies; all modifications serve the pur-
1220 pose of benchmark evaluation.

1222 **Is the dataset self-contained, or does**
1223 **it link to or otherwise rely on external**
1224 **resources (e.g., websites, tweets, other**
1225 **datasets)?** If it links to or relies on exter-
1226 nal resources, a) are there guarantees that
1227 they will exist, and remain constant, over
1228 time; b) are there official archival versions of
1229 the complete dataset (i.e., including the ex-
1230 ternal resources as they existed at the time
1231 the dataset was created); c) are there any
1232 restrictions (e.g., licenses, fees) associated
1233 with any of the external resources that might
1234 apply to a future user? Please provide de-
1235 scriptions of all external resources and any
1236 restrictions associated with them, as well as
links or other access points, as appropriate.

1237 The dataset is entirely self-contained.

1239 **Does the dataset contain data that might**
1240 **be considered confidential (e.g., data**
1241 **that is protected by legal privilege or by**
doctor-patient confidentiality, data that

includes the content of individuals non-
public communications)? If so, please
provide a description.

The dataset does not contain confidential data as
it is based on publicly available ImageNet data.

Does the dataset contain data that, if
viewed directly, might be offensive, in-
sulting, threatening, or might otherwise
cause anxiety? If so, please describe why.

The dataset does not contain offensive or dis-
turbating content as it focuses on visual distortions
applied to non-sensitive images. Additionally,
the images sourced from ImageNet are manually
filtered to exclude any content that could be con-
sidered disturbing.

Does the dataset relate to people? If not,
you may skip the remaining questions in this
section.

Yes, the LAION-C dataset relates to people to
some extent as it includes images from Im-
ageNet, some of which feature human faces and
figures. While the primary focus of the dataset is
not on the individuals depicted or on analyzing
human-specific data, the presence of human im-
ages means that the dataset does relate to people
indirectly.

Does the dataset identify any subpop-
ulations (e.g., by age, gender)? If so,
please describe how these subpopulations
are identified and provide a description
of their respective distributions within the
dataset.

The LAION-C dataset itself does not explic-
itly identify subpopulations by age, gender, or
other demographic characteristics as part of its
core design. However, since it includes im-
ages from ImageNet, which may contain human
faces, there is an implicit presence of such de-
mographic data.

Is it possible to identify individuals (i.e.,
one or more natural persons), either di-
rectly or indirectly (i.e., in combination
with other data) from the dataset? If so,
please describe how.

While the primary intention of the LAION-C
dataset is not to facilitate the identification of in-
dividuals, it incorporates images from ImageNet,
which may include human faces.

Does the dataset contain data that might
be considered sensitive in any way (e.g.,

1242 **data that reveals racial or ethnic origins,**
 1243 **sexual orientations, religious beliefs, po-**
 1244 **litical opinions or union memberships, or**
 1245 **locations; financial or health data; bio-**
 1246 **metric or genetic data; forms of govern-**
 1247 **ment identification, such as social secu-**
 1248 **rity numbers; criminal history)?** If so,
 1249 please provide a description.

1250 While the LAION-C dataset primarily features
 1251 synthetic distortions applied to images for tech-
 1252 nical analysis, it includes images sourced from
 1253 ImageNet that may contain human faces. These
 1254 images can indirectly reveal racial or ethnic ori-
 1255 gins due to the diversity of individuals depicted.
 1256 However, there is no explicit focus on collecting
 1257 or analyzing data related to sexual orientations,
 1258 religious beliefs, political opinions, union mem-
 1259 berships, specific locations, financial or health
 1260 data, biometric or genetic data, government iden-
 1261 tification numbers, or criminal history. The in-
 1262 clusion of human images is incidental and not
 1263 intended for any analysis related to these sensi-
 1264 tive aspects.

1265 **Any other comments?**

1266 None.

Collection Process

1271 **How was the data associated with each**
 1272 **instance acquired?** Was the data di-
 1273 rectly observable (e.g., raw text, movie rat-
 1274 ings), reported by subjects (e.g., survey re-
 1275 sponses), or indirectly inferred / derived from
 1276 other data (e.g., part-of-speech tags, model-
 1277 based guesses for age or language)? If
 1278 data was reported by subjects or indirectly
 1279 inferred / derived from other data, was the
 1280 data validated / verified? If so, please de-
 1281 scribe how.

1282 The data for each instance in the LAION-C
 1283 dataset is derived from ImageNet, where images
 1284 are directly observable and not reported by sub-
 1285 jects or inferred.

1286 **What mechanisms or procedures were**
 1287 **used to collect the data (e.g., hardware**
 1288 **apparatus or sensor, manual human cu-**
 1289 **ration, software program, software API)?**
 1290 How were these mechanisms or procedures
 1291 validated?

1292 First, 16 sensible high-level classes were se-
 1293 lected that the authors deemed suitable for hu-
 1294 mans to recognize in psychophysical experi-
 1295 ments. These classes are: ball, bird, boat, bot-
 tle, butterfly, car & truck, cat, chair, dog, fish,

fruit, instrument, primate, snake, timekeeping,
 and tool. Then, 200 classes from the original
 ImageNet-1k set were selected that can consti-
 tute these high-level classes. From the pools
 of validation set images, 500 images were ran-
 domly selected per superclass. These images
 were then manually filtered to include only im-
 ages that fall clearly into one of the 16 super-
 classes (i.e. an image showing both a ball and a
 dog would have been filtered out to ensure clean
 class labels).

**If the dataset is a sample from a larger
 set, what was the sampling strategy (e.g.,
 deterministic, probabilistic with specific
 sampling probabilities)?**

See previous question. Candidate images from
 the constituent classes were sampled randomly
 with uniform probability.

**Who was involved in the data collection
 process (e.g., students, crowdworkers,
 contractors) and how were they compen-
 sated (e.g., how much were crowdwork-
 ers paid)?**

Information will be provided upon publication.

**Over what timeframe was the data col-
 lected? Does this timeframe match the
 creation timeframe of the data associ-
 ated with the instances (e.g., recent crawl
 of old news articles)?** If not, please de-
 scribe the timeframe in which the data asso-
 ciated with the instances was created.

The source dataset for the creation of LAION-
 C was the 2012 ILSVRC validation set (“Ima-
 geNet”) which was collected over several years.
 The distortions applied in LAION-C were cre-
 ated specifically for benchmarking purposes at
 the time of dataset development (2023 / 2024),
 which do not coincide directly with the original
 image collection periods.

**Were any ethical review processes con-
 ducted (e.g., by an institutional review
 board)?** If so, please provide a descrip-
 tion of these review processes, including the
 outcomes, as well as a link or other access
 point to any supporting documentation.

The original ImageNet dataset underwent vari-
 ous ethical and review processes during its de-
 velopment, details are managed by the original
 collector for ImageNet.

1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Only indirectly. LAION-C includes images from ImageNet that feature human faces and figures.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Not applicable.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable.

Did the individuals in question consent to the collection and use of their data?

If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No specific data protection impact analysis has been conducted for the LAION-C dataset as its primary modifications involve applying synthetic distortions like glitches to the images for technical benchmarking purposes. These alterations do not fundamentally change the nature

of the data regarding privacy or ethical concerns beyond their original use in ImageNet.

of the data regarding privacy or ethical concerns beyond their original use in ImageNet.

Any other comments?

None.

Preprocessing / cleaning / labeling

Was any preprocessing / cleaning / labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Images were resized to 256x256 pixels and center-cropped to 224x224 pixels, as is common for ImageNet. Images were filtered manually to ensure clean labels as described above.

Was the “raw” data saved in addition to the preprocessed / cleaned / labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

No, LAION-C only consists of the modified images, but every filename can be uniquely traced back to the parent image from the ImageNet validation set, which can be found here: <https://www.image-net.org/download.php>

Is the software used to preprocess / clean / label the instances available? If so, please provide a link or other access point.

Yes, the preprocessing, cleaning, and labeling of the dataset instances were conducted using Python. The code used for these processes will be made available upon publication.

Any other comments?

None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Yes, the LAION-C dataset has been utilized to evaluate the robustness and out-of-distribution (OOD) generalization capabilities of large-scale vision models.

1350 **Is there a repository that links to any**
1351 **or all papers or systems that use the**
1352 **dataset?** If so, please provide a link or other
1353 access point.

1354 Information will be provided upon publication.
1355

1356
1357 **What (other) tasks could the dataset be**
1358 **used for?**

1359 Beyond benchmarking vision model robustness,
1360 LAION-C could be used in studies investigating
1361 the effects of image distortions on human per-
1362 ception.
1363

1364 **Is there anything about the composition**
1365 **of the dataset or the way it was col-**
1366 **lected and preprocessed / cleaned / la-**
1367 **beled that might impact future uses?** For
1368 example, is there anything that a future user
1369 might need to know to avoid uses that could
1370 result in unfair treatment of individuals or
1371 groups (e.g., stereotyping, quality of service
1372 issues) or other undesirable harms (e.g., fi-
1373 nancial harms, legal risks) If so, please pro-
1374 vide a description. Is there anything a future
1375 user could do to mitigate these undesirable
1376 harms?

1377 Given that the base images in the LAION-C
1378 dataset are sourced from ImageNet, which is al-
1379 ready publicly available, the additional risk for
1380 harm is negligible.

1381 **Are there tasks for which the dataset**
1382 **should not be used?** If so, please provide
1383 a description.

1384 We would not recommend using the LAION-C
1385 dataset for fine-tuning machine learning models,
1386 due to dataset size.

1387 **Any other comments?**

1388 None.
1389
1390

1391

Distribution

1392
1393

1394 **Will the dataset be distributed to third**
1395 **parties outside of the entity (e.g., com-**
1396 **pany, institution, organization) on behalf**
1397 **of which the dataset was created?** If so,
1398 please provide a description.

1399 The LAION-C dataset will be made publicly
1400 available, allowing for distribution to third
1401 parties outside of the originating entity.
1402

1403 **How will the dataset will be distributed**
(e.g., tarball on website, API, GitHub)

Does the dataset have a digital object iden-
tifier (DOI)?

Upon publication, the dataset will be published
via Zenodo.

When will the dataset be distributed?

The dataset will be distributed upon publication.

Will the dataset be distributed under a
copyright or other intellectual property
(IP) license, and / or under applicable
terms of use (ToU)? If so, please describe
this license and / or ToU, and provide a link
or other access point to, or otherwise repro-
duce, any relevant licensing terms or ToU,
as well as any fees associated with these
restrictions.

LAION-C will be available under a CC BY-
NC 4.0 license, allowing non-commercial use
with proper attribution only, to ensure compli-
ance with the original ImageNet license.

Have any third parties imposed IP-based
or other restrictions on the data associ-
ated with the instances? If so, please de-
scribe these restrictions, and provide a link
or other access point to, or otherwise repro-
duce, any relevant licensing terms, as well
as any fees associated with these restric-
tions.

The original ImageNet data is subject to terms
of access that limit its use to non-commercial
research and educational purposes only. The full
terms of access can be found here: [https://
www.image-net.org/download.php](https://www.image-net.org/download.php)

Do any export controls or other regula-
tory restrictions apply to the dataset or
to individual instances? If so, please de-
scribe these restrictions, and provide a link
or other access point to, or otherwise repro-
duce, any supporting documentation.

Since the images are modified ImageNet im-
ages, the restrictions of the ImageNet license ap-
ply.

Any other comments?

None

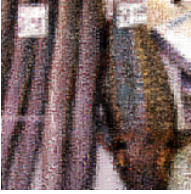


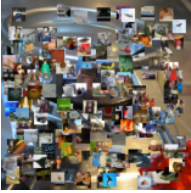
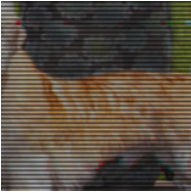


Maintenance

Who will be supporting / hosting / main-
taining the dataset?

1404 Information will be provided upon publication. Not applicable (beyond agreements made for
1405 ImageNet).
1406

1407 **How can the owner / curator / manager of the dataset be contacted (e.g., email address)?** **Will older versions of the dataset continue to be supported / hosted / maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
1408
1409
1410 Information will be provided upon publication.
1411

1412
1413 **Is there an erratum?** If so, please provide a link or other access point. Should newer versions of the dataset be created, older versions will continue to be available via Zenodo.
1414
1415
1416 There is not an explicit erratum as for now. **If others want to extend / augment / build on / contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated / verified? If so, please describe how. If not, why not? Is there a process for communicating / distributing these contributions to other users? If so, please provide a description.
1417
1418 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
1419
1420
1421 Information will be provided upon publication.
1422
1423
1424
1425
1426 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced. We encourage other researchers to build on LAION-C, for example by contributing their own corruptions. While there is no automatic mechanism (such as publicly accessible version control, e.g. via Github) for this, we encourage interested parties to reach out to the authors.
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

1458			
1459			
1460			
1461		The image is highly pixelated, but I can discern the general shape of a dog, especially the head and front legs. The color pattern also suggests a dog's fur.	True label: fish Predicted label: dog mosaic (1)
1462			
1463			
1464			
1465			
1466			
1467		The image seems to show a fruit, possibly an apple, based on the red and green colors present in the picture.	True label: chair Predicted label: fruit mosaic (1)
1468			
1469			
1470			
1471			
1472			
1473			
1474		The distorted image appears to show the outline of a bird with wings outstretched. The pointed shapes resemble wings, and the overall posture suggests a bird in flight.	True label: ball Predicted label: bird vertical_lines (2)
1475			
1476			
1477			
1478			
1479			
1480			
1481		I can see pliers in the center of the image. Pliers are a tool.	True label: tools Predicted label: tools stickers (2)
1482			
1483			
1484			
1485			
1486			
1487		The image appears to show a bird, potentially a chicken, due to the presence of orange feathers and what looks like a wattle. The distortion makes it difficult to be certain, but the color and shape suggest a bird more than any other category.	True label: dog Predicted label: bird glitched (1)
1488			
1489			
1490			
1491			
1492			
1493			
1494		The round shape and the presence of what appears to be a drumstick suggest that this image represents a drum, which is a musical instrument.	True label: tools Predicted label: instruments glitched (1)
1495			
1496			
1497			
1498			
1499			
1500			
1501		Although the image is heavily distorted, it appears to show the face of a primate. The discernible features, such as the eyes and brow ridge, suggest a primate's facial structure.	True label: cat Predicted label: primate glitched (2)
1502			
1503			
1504			
1505			
1506			

1507 **Figure 11: Visual Reasoning in Gemini.** We provide examples of visual reasoning in Gemini-
 1508 1.5-Pro, consisting of a LAION-C sample, the reasons for classification that Gemini provided and
 1509 meta-information (like the final label, the ground-truth label and corruption details). In line with our
 1510 findings about Error Consistency (see Fig. 7), the model is frequently confused by visual impressions
 1511 that are trivial for humans to decode. Instead, it hallucinates unreasonable image properties.





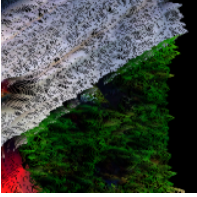
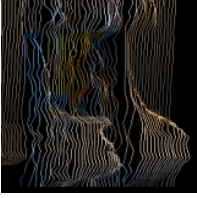

1512			
1513			
1514			
1515		The image shows a shape that resembles the outline of a bottle, with a neck and body typical of bottles, despite the heavy distortion and colorful overlay.	True label: ball Predicted label: bottle geometric_shapes (3)
1516			
1517			
1518			
1519			
1520			
1521		The image shows a metallic object with a handle, resembling a hammer, which is a common tool. The shape and context suggest it is used for hitting or driving nails.	True label: chair Predicted label: tools glitched (1)
1522			
1523			
1524			
1525			
1526			
1527			
1528		The image appears to be a mosaic or collage that forms the shape of a primate, likely a monkey or ape. The overall structure and features resemble the face and body of a primate, which is why this category is chosen.	True label: fish Predicted label: primate mosaic (2)
1529			
1530			
1531			
1532			
1533			
1534			
1535		The image appears to depict a stylized or pixelated representation of a primate, likely a monkey or ape, based on the visible facial features and overall shape.	True label: snakes Predicted label: primate mosaic (2)
1536			
1537			
1538			
1539			
1540			
1541			
1542		The image shows a shape and texture that resembles a fish, with a streamlined body and scales visible. The colors and patterns are typical of fish, making it the most likely category.	True label: chair Predicted label: fish vertical_lines (1)
1543			
1544			
1545			
1546			
1547			
1548		The image appears to have a distorted outline that resembles the shape of a primate, such as a monkey or ape. The contours suggest facial features and a posture typical of primates.	True label: bottle Predicted label: primate vertical_lines (5)
1549			
1550			
1551			
1552			
1553			
1554			
1555		The image contains multiple small images, many of which depict butterflies. The presence of wings and colorful patterns typical of butterflies is noticeable, making 'butterfly' the most fitting category.	True label: fruit Predicted label: butterfly mosaic (4)
1556			
1557			
1558			
1559			
1560			
1561			
1562			
1563			
1564			
1565			

Figure 12: **Visual Reasoning in GPT.** Figure analogous to Fig. 11. Like Gemini, GPT has a tendency to hallucinate visual impressions. Notably, when the model makes mistakes in the sticker-corruption, it is typically led astray by the contents of the sticker-images, ignoring the background completely and failing to perceive the global structure of the image, a behavior that Gemini exhibits as well.