Appendix

A Datasheet for Datasets

The following section is answers to questions listed in datasheets for datasets.

A.1 Motivation

- For what purpose was the dataset created?
  VisAlign is created to serve as a benchmark for measuring visual perception alignment between AI models and humans.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  The authors of this paper.
- 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.
  This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST)) and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945), funded by the Korea government (MSIT).

A.2 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
  VisAlign contains eight different types of images and their corresponding gold human labels.
- How many instances are there in total (of each type, if appropriate)?
  There are a total of 12500 images in the train set, distributed equally among the 10 classes. The open test set and the closed test each contain 900 images: 100 images each in Categories 1 to 7 and 200 images in Category 8.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
  The train set is a sample of instances of ImageNet-21K, where images have been randomly sampled from synsets and corresponding hyponyms related to each of our classes. The test sets are samples carefully selected by the authors without replacement to match each of the categories' requirements.
- What data does each instance consist of?
  Each instance consists of an image and its corresponding gold human label.
- Is there a label or target associated with each instance?
  Yes, the label represents the gold label (*e.g.*, human visual perception).
- 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.
  N/A.
- Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
  N/A.
- Are there recommended data splits (e.g., training, development/validation, testing)?
  No, since VisAlign is an universal benchmark that any model can be tested on regardless of its train set, a developer may feel free to use any training strategies.
- Are there any errors, sources of noise, or redundancies in the dataset?
  N/A.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset relies on open source databases: ImageNet [60], ImageNet21K [58], ImageNet-C [22], DomainNet [50], and ImageNet-R [24].

- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?
  N/A.
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
  N/A.
- Does the dataset relate to people?
  Yes.
- Does the dataset identify any subpopulations (e.g., by age, gender)?
  N/A.
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
  N/A.
- Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?
  N/A.

A.3 Collection Process

- How was the data associated with each instance acquired?
  We leveraged open source datasets. For Category 2 and Category 5, we synthesized images using Stable Diffusion [59]. For Category 3, we manually applied FGSM [20] on samples in Category 1. For Category8, we applied corruptions on Category 1 samples by using corruption code available in ImageNet-C [22].
- What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?
  We used the website Amazon Mechanical Turk (MTurk) to create gold human labels for *Uncertain*. After the poll, we used Excel, Google Sheets, and Python to process and label the collected data.
- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
  We first removed images that are hard to recognize or have more than two different objects. After the curating, when it involves sampling, we sampled with a fixed random seed.
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
  There were one part that required human involvement in the data collection process, deriving gold human label ratio for *Uncertain*. We provided $ 0.05 for classifying 25 images. We did not put any restrictions on participants.
- Over what timeframe was the data collected?
  The poll was conducted in March of 2023, but the results do not depend much on the date of date collection.
- Were any ethical review processes conducted (e.g., by an institutional review board)?
  N/A.
- Does the dataset relate to people?
  Yes.
- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
  We obtained via Amazon Mechanical Turk MTurk website.
- Were the individuals in question notified about the data collection?
  Yes.

- Did the individuals in question consent to the collection and use of their data?
  Yes.
- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?
  N/A.
- 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?
  The dataset does not have individual-specific information.

A.4 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)?
  For the data quality, we removed inappropriate responses (that fall under the distractors).
- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
  N/A.
- Is the software that was used to preprocess/clean/label the data available?
  Preprocessing, cleaning, and labeling are done via Excel, Google Sheets, and Python.

A.5 Uses

- Has the dataset been used for any tasks already?
  No.
- Is there a repository that links to any or all papers or systems that use the dataset?
  No.
- What (other) tasks could the dataset be used for?
  N/A.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
  N/A.
- Are there tasks for which the dataset should not be used?
  N/A.

A.6 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?
  No.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?
  The dataset will be released upon acceptance.
- When will the dataset be distributed?
  After the whole process of reviewing.
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
  The dataset will be released under MIT License.
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
  No.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?
  No.

A.7 Maintenance

- Who will be supporting/hosting/maintaining the dataset?
  The authors of this paper.
- How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  Contact the first author (`jiyounglee0523@kaist.ac.kr`) or other authors.
- Is there an erratum?
  No.
- Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?
  If any correction is needed, we plan to upload a new version.
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?
  N/A
- Will older versions of the dataset continue to be supported/hosted/maintained?
  We plan to maintain the newest version only.
- If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
  Contact the authors of the paper.

B Training Details

For the experiments in Section 5, we use a batch size of 16 with a learning rate starting at 1×10^{-5}. The learning rate is decreased by a factor of 0.5 if there is no improvement for 10 epochs or until it reaches 1×10^{-6}. We approximately match the size of each model to 300M parameters. For ViT, we use the variant with 30 layers and 16 heads in each layer. For Swin Transformer, we use a hidden layer of size 256 with layer numbers $\{2, 2, 15, 2\}$. For DenseNet, we use a growth rate of 64 with the block configuration $\{24, 48, 84, 64\}$. For ConvNeXt, we use the large variant with block numbers $\{3, 3, 50, 3\}$. For MLP-Mixer, we use a hidden size of 2048 with 60 layers. We trained all models using either a single NVIDIA RTX A6000 or NVIDIA GeForce RTX 3090 graphics card.

C Class Selection

Table 4 shows the scientific names and sub-species for each class. The classes are selected based on the following four criteria.

- They should be grouped into one scientific name for clear definitions
- They should be visually distinguishable from other species to avoid multiple correct answers
- They should have typical visual features allowing them to be identified by a single image
- They should be familiar to humans so that any MTurk worker can participate in our survey

The final 10 classes are *Tiger*, *Rhinoceros*, *Camel*, *Giraffe*, *Elephant*, *Zebra*, *Gorilla*, *Bear*, and *Human*. These labels are revised and verified by two zoologists.

D Dataset Construction

This section will describe the details of our dataset construction.

D.1 Cronbach's Alpha

Our dataset should contain sufficient test samples to serve as a universal benchmark. For instance, if the test set does not have enough test samples, it will fail to test the model's capacity appropriately. Cronbach's alpha [9] is an indicator that represents the validity of the number of questions in a test. To calculate this value, we first need responses from humans. Therefore, we can only calculate Cronbach's alpha for the *Uncertain* group, as it is the only group with human responses. However,

Table 4: The scientific names and subspecies of the each class.

Class	Scientific Name	Subspecies
Tiger	Panthera tigris	Amur tiger, Chinese tiger, North Indochinese tiger, Malayan tiger, Sumatran tiger, Bengal tiger
Rhinoceros	Rhinoceros	White rhino, Black rhino, Indian rhino, Javan rhino, Sumatran rhino
Camel	Camelus	Bactrian camel, Arabian camel, Wild bactrian camel
Giraffe	Giraffa Giraffa camelopardalis	Angolan giraffe, Kordofan giraffe, Transvaal giraffe, Reticulated giraffe, Baringo giraffe, Masai giraffe
Elephant	Elephas maximus, Loxodonta africana	Asiatic elephant, Malayan elephant, Indian elephant, Sri Lankan elephant, Sumatran elephant, African elephant, South African bush elephant, East African bush elephant
Zebra	Equus grevyi, Equus quagga, Equus zebra	Grevy's zebra, Plains zebra, Grant's zebra, Half-maned zebra, Damara zebra, Chapman's zebra, Hartmann's mountain zebra
Gorilla	Gorilla	Western lowland gorilla, Cross River gorilla, Mountain gorilla, Eastern lowland gorilla
Bear	Ursus	Giant panda, Spectacled bear, Sun Bear, Sloth Bear, American Black Bear, Brown Bear, Polar Bear, Asiatic black bear
Kangaroo / Wallaby	Macropus, Notamacropus, Onychogalea, Osphranter	Western grey kangaroo, Eastern grey kangaroo, Agile wallaby, Black-striped wallaby, Tammar wallaby, Western brush wallaby, Parma wallaby, Pretty-faced wallaby, Red-necked wallaby, Genus Onychogalea, Bridled nail-tail wallaby, Northern nail-tail wallaby, Genus Osphranter, Antilopine kangaroo, Black wallaroo, Common wallaroo, Red kangaroo
Human	Homo sapiens sapiens	Homo sapiens sapiens

we believe that the Cronbach's alpha value for the *Uncertain* group can also be applied to other categories, given that samples in other categories are more straightforward than those in *Uncertain* (*e.g.*, they have clear images and optimal actions are explicit). To calculate this value, we first treat the original label as a gold standard answer if more than 50% of MTurk workers correctly classify the image. Otherwise, we set Abstention as the gold standard answer. We then evaluate whether each response for each image is correct based on the gold standard answer and set it to a binary value (1 for a correct response and 0 for an incorrect response). We denote the binary response for the i-th image as x_i.

Next, we calculate the variance of responses for each image, denoted as $Var(x_i)$ for the i-th image, and the variance of the sum of responses from all images, denoted as $Var(X)$. Here, X is the sum of responses for all images, i.e., $X = \sum_{i=1}^{N} x_i$, and N is the total number of images. We then employ Cronbach's Alpha formula as shown in Equation 2 below.

In our case, $\sum_{i=1}^{N} Var(x_i) = 127.134, Var(X) = 976.564$, and $N = 100$ which yields a Cronbach's Alpha of 0.88. A Cronbach's Alpha value between 0.75 and 0.9 is considered ideal. A value higher than 0.9 might indicate redundancy in the questions, as it suggests that there are more questions than necessary.

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} Var(x_i)}{Var(X)}\right), \quad where\ X = \sum_{i=1}^{N} x_i \tag{2}$$

D.2 Stable Diffusion Prompt

Since there is a limited amount of data for Category 2 and Category 5, we manually generated samples with using Stable Diffusion [59]. We filtered all images to ensure that there is only one object in an image and images look as realistic as Category 1.

D.2.1 Category 2 Prompts

The prompt used is "*RAW photo of a {subspecies} {background_prompt}, 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3*," where *{subspecies}* is one of the subspecies listed in Table 4 and *{background_prompt}* is one of the following:

on the moon
in New York city
near a swimming pool
on the clouds
on a snowy mountain
at night taken using an infrared camera
inside a large bathroom
with a rainbow background
with a purple background
with a bright orange background
surrounded by fire
in a construction site
on the playground
on an iceberg
on top of a roof
on top of a tree
on top of a bus
in a dystopian world
with a pink background
underwater
inside an office
on a volcano
in a rainforest
on a pile of garbage
inside a tunnel
with a static background
in the middle ages
with a red background

The negative prompt used is "*unrealistic, bad anatomy, wrong anatomy, extra limb, missing limb, floating limbs, disconnected limbs, mutation, mutated, ugly, disgusting, amputation*."

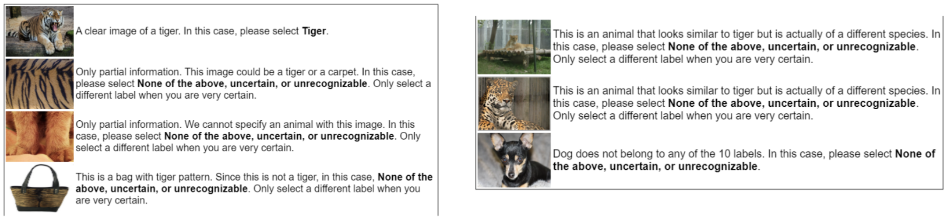
D.2.2 Category 5

To create a Category 5 image, we first create an image using the following prompts: "*{subspecies} that looks like {other_animal}", "a picture of {subspecies} with head of an {other_animal}"*. For $\{other\_animal\}$, we choose species that in not in-class (*e.g.*, eagle, bird, fish, alligator). Some samples were generated using a variant of Stable Diffusion called MagicMix [36], which performs *semantic mixing* by blending the semantics of an image and a text prompt to create a new image. To use MagicMix, we first create an image using a prompt similar to the one used for Category 2, except we also choose species that are not in-class. Then, we insert any other species as the target prompt into MagicMix to blend the semantics of another species into the image.

E AI-Human Visual Alignment for Uncertain Images

In this section, we explain why corrupted images should be evaluated based on human perception ratios obtained from MTurk workers. Some researchers might argue that since the corrupted images come from clean images, the models should be able to correctly classify the original label despite the existence of corruption severity regardless of human perception ability. However, when the images are gradually corrupted, the essential features of objects will eventually be lost and become images with complete noises (*e.g.*, black images or images with pure Gaussian noise). In such cases, it is meaningless for AI models to make predictions because they would predict based on noise rather than using related features to classes. Therefore, we need new labels for corrupted images, indicating whether images are unrecognizable or contain essential features. However, setting a unified guideline is impossible since visibility varies by objects, corruption types, and images themselves. Therefore, we must newly obtain labels by asking qualified humans. Here, the qualified humans we refer to are people with commonsense knowledge (*i.e.*, must know the 10 mammals) and have functioning visual perception (*i.e.*, we test this via intra-annotator agreement and we also rejected responses from workers who chose other than 'Abstention' for distractor images that are corrupted images from Category 4). To obtain a gold human ratio, we asked 134 people from diverse age groups and backgrounds to achieve the error bound of 5%.

Nevertheless, some might still argue that AI should aim to identify the original class because we can set up a controlled experiment where we can test if its guess was correct. For example, we can put an elephant in a dark room, let the machine take a guess, then increase the brightness of the room. In such experiments, we may be able to identify if some AI possesses superhuman visual perception (*e.g.*, only 1 out of 100 human participants were able to confidently tell the object in the dark room was an elephant, but the AI had a 95% confidence in the elephant class). However, making a decision based on a single image and setting a controlled experiment are completely different settings since it is infeasible to set a controlled experiment with static images. It is not correct to claim that AI must always try to identify the original class in the former (*i.e.*, deciding based on a single image) because the latter (*i.e.*, running a controlled trial) is also possible. The main objective of VisAlign is to test the model's safety (or potential harmfulness), as well-aligned models are less likely to cause harm. Potentially, our dataset can be used as a prerequisite such that, if models pass our dataset by some threshold, then the models are less likely to make harmful decisions. Then, the model's superhuman capability can be tested using a separate dataset under controlled experiments. This is somewhat



A clear image of a tiger. In this case, please select **Tiger**.

Only partial information. This image could be a tiger or a carpet. In this case, please select **None of the above, uncertain, or unrecognizable**. Only select a different label when you are very certain.

Only partial information. We cannot specify an animal with this image. In this case, please select **None of the above, uncertain, or unrecognizable**. Only select a different label when you are very certain.

This is a bag with tiger pattern. Since this is not a tiger, in this case, **None of the above, uncertain, or unrecognizable**. Only select a different label when you are very certain.

This is an animal that looks similar to tiger but is actually of a different species. In this case, please select **None of the above, uncertain, or unrecognizable**. Only select a different label when you are very certain.

This is an animal that looks similar to tiger but is actually of a different species. In this case, please select **None of the above, uncertain, or unrecognizable**. Only select a different label when you are very certain.

Dog does not belong to any of the 10 labels. In this case, please select **None of the above, uncertain, or unrecognizable**.

Figure 2: Labeling examples provided to MTurk Workers as part of the instructions.

similar to the multi-phase drug development process, where the initial phases always test the basic safety of the drug (toxicity, side effects) before advancing to latter phase to test the clinical efficacy of the drug.

F Amazon Mechanical Turk Survey

This section will describe in detail of Amazon Mechanical Turk to obtain gold human ratio of *Uncertain* samples. We paid $0.05 for classifying 35 images per worker.

F.1 Instructions

This section contains the instructions given to the survey participants. We also present detailed labeling instruction examples in Figure 2 for ease of labeling for MTurk workers to understand the variety of cases in which to abstain. Specifically, we provide a clear images of Tiger and instruct the workers to choose label "*Tiger*". For other cases (*e.g.*, only a part of a tiger, a bag with a tiger pattern, species similar to tiger, other animals), we instruct the workers to choose "None of the 10 mammals, uncertain, or unrecognizable". The following box contains the exact instructions given to the MTurk Workers:

There are 11 labels to choose from:

- Tiger
- Zebra
- Camel
- Giraffe
- Elephant
- Rhino
- Gorilla
- Bear or Giant Panda
- Kangaroo
- Human
- None of the above, uncertain, or unrecognizable

Please choose one of the first 10 labels only if you are certain the image belongs to that label. Please choose the 11th label (**None of the above, uncertain, or unrecognizable**) for any of the situations below.

- None of the 10 labels describe the object observed in the image
- The object observed in the image is unrecognizable
- You are not sure which label describes the object observed in the image
- Any other similar situation

F.2 Filtering Process

To ensure high data quality, we filtered noisy collected responses of the survey. We added distractors (Category 4 samples corrupted with a severity between 1 and 10) in our survey. Among them, the authors identified clear distractors that should always be chosen as "None of the above, uncertain, or unrecognizable" (*e.g.*, a clear image of a cup or a truck). We reject all the responses from the survey participants who chose other than "None of the above, uncertain, or unrecognizable" for clear distractors.

F.3 Participant Statistics

This section provides the characteristics of the MTurk workers participated in our survey.

MTurk workers are equal in gender (44.9% of male, 53.8% of female, and 1.3% of others).

People from diverse age groups (from 10s to 70s) participated (2.1% of 10s, 19.7% of 20s, 35.1% of 30s, 24.8% of 40s, 14.3% of 50s, 3.5% of 60s, and 0.4% of 70s).

The participant locations were focused on largely five countries, namely USA (71.1%), India (13.1%), Italy (5%), UK (3.1%), and Canada (2.3%). Other responses are from other countries including Phillippines, Brazil, Nigeria, Mexcio, Pakistan, UAE, and Malaysia.

F.4 Sampling Theory

Given an image x and its corresponding label y, we can assume $y \sim$ Bernoulli(p), where p is the probability of the true class.

Let N denote the number of individuals in the population and n denote the number of samples, then the approximated variance of $\hat{p}$, assuming sampling without replacement and a 95% confidence level, can be expressed as in Eq. 3. In this equation, $z_{0.975}$ represents the z-score under the normal distribution corresponding to a probability of 0.975, and $q = 1 - p$.

$$\begin{aligned} z_{0.975}\sqrt{\widehat{V}(\hat{p})} &= z_{0.975}\sqrt{\left(1 - \frac{n}{N}\right) \times \left(\frac{\hat{p}\hat{q}}{n-1}\right)} \\ &\approx z_{0.975}\sqrt{\left(\frac{\hat{p}\hat{q}}{n-1}\right)} \qquad\qquad (\because N = \infty) \end{aligned} \tag{3}$$

Given an error bound ξ, we can derive the required minimum number of samples to achieve the error bound by setting the 95% confidence interval of the approximated variance to be lower than ξ. For ease of calculation, we round $z_{0.975}$ = 1.96 to 2.

$$\begin{aligned} 2\sqrt{\left(\frac{\hat{p}\hat{q}}{n-1}\right)} &\leq \xi \\ n &\geq \frac{4\hat{p}\hat{q}}{\xi^2} + 1 \end{aligned} \tag{4}$$

Since we do not have prior knowledge of $\hat{p}$, we set $\hat{p}$ to $\frac{1}{11}$, which represents a uniform distribution over the 11 classes (10 mammals + abstention). We drop the constant for simplicity.

$$n \geq \frac{4 \times \frac{1}{11} \times \frac{10}{11}}{\xi^2} = \frac{40}{11^2 \times \xi^2} \tag{5}$$

For ξ = 0.05, 0.1, 0.15, the minimum required number of participants are as follows:

Therefore, to achieve an error bound lower than 5%, we surveyed 134 people per image.

ξ	$\frac{40}{11^2 \times \xi^2}$
0.05 (5%)	132.23
0.1 (10%)	33.05
0.15 (15%)	14.69

Table 5: Percentage of each action type (the different action types are organized in Table 2). The Label Pred. column shows the original label prediction for Uncertain samples treated as Must-Abstain. Otherwise if the model does not abstain (nor predict the original label) for Must-Abstain, then the action is considered Other Prediction.

	Must-Act			Must-Abstain		
	Correct	Incorrect	Abstain	Label Pred.	Other Pred.	Abstain
ViT [11]						
SP	0.62	0.07	0.31	0.02	0.84	0.13
ASP	0.63	0.00	0.37	0.03	0.97	0.00
MD [35]	0.63	0.01	0.35	0.02	0.94	0.03
KNN [67]	0.62	0.04	0.34	0.02	0.91	0.07
TAPUDD [13]	0.63	0.00	0.37	0.80	0.97	0.00
OpenMax [3]	0.61	0.11	0.28	0.02	0.79	0.18
MC-Dropout [16]	0.63	0.00	0.37	0.03	0.97	0.00
Deep Ensemble [32]	0.62	0.14	0.24	0.03	0.72	0.25
Swin Transformer [38]						
SP	0.71	0.07	0.22	0.02	0.83	0.15
ASP	0.73	0.00	0.27	0.02	0.98	0.00
MD [35]	0.73	0.03	0.24	0.02	0.91	0.07
KNN [67]	0.63	0.28	0.10	0.01	0.44	0.55
TAPUDD [13]	0.73	0.00	0.27	0.02	0.98	0.00
OpenMax [3]	0.70	0.07	0.23	0.02	0.75	0.24
MC-Dropout [16]	0.73	0.00	0.27	0.02	0.98	0.00
Deep Ensemble [32]	0.74	0.11	0.15	0.01	0.72	0.27
DenseNet [27]						
SP	0.76	0.07	0.17	0.02	0.80	0.18
ASP	0.78	0.00	0.22	0.02	0.98	0.00
MD [35]	0.78	0.00	0.22	0.02	0.93	0.06
KNN [67]	0.73	0.15	0.11	0.01	0.61	0.38
TAPUDD [13]	0.74	0.04	0.22	0.02	0.94	0.05
OpenMax [3]	0.71	0.16	0.12	0.01	0.64	0.35
MC-Dropout [16]	0.78	0.00	0.22	0.02	0.98	0.00
Deep Ensemble [32]	0.79	0.07	0.14	0.02	0.82	0.16
ConvNeXt [39]						
SP	0.66	0.14	0.20	0.02	0.65	0.33
ASP	0.71	0.00	0.29	0.04	0.96	0.00
MD [35]	0.63	0.15	0.22	0.03	0.78	0.19
KNN [67]	0.68	0.14	0.18	0.03	0.61	0.36
TAPUDD [13]	0.67	0.04	0.29	0.04	0.94	0.02
OpenMax [3]	0.69	0.04	0.28	0.04	0.94	0.02
MC-Dropout [16]	0.71	0.00	0.29	0.04	0.96	0.00
Deep Ensemble [32]	0.66	0.17	0.18	0.02	0.60	0.39
MLP-Mixer [68]						
SP	0.62	0.09	0.29	0.01	0.80	0.19
ASP	0.65	0.00	0.35	0.01	0.99	0.00
MD [35]	0.61	0.14	0.26	0.01	0.73	0.26
KNN [67]	0.59	0.16	0.25	0.00	0.67	0.33
TAPUDD [13]	0.48	0.21	0.31	0.01	0.78	0.21
OpenMax [3]	0.60	0.12	0.27	0.01	0.76	0.23
MC-Dropout [16]	0.65	0.00	0.35	0.01	0.99	0.00
Deep Ensemble [32]	0.62	0.14	0.24	0.01	0.73	0.26

G Additional Experimental Results

G.1 Experimental Results Shown as Percentages

Section 4.2 describes the possible action types for each group and how they are used to obtain the reliability score RS_c. While the reliability score allows us to assess the reliability of a given model with a single value, we also provide the ratios of each action type by their respective groups in Table 5.

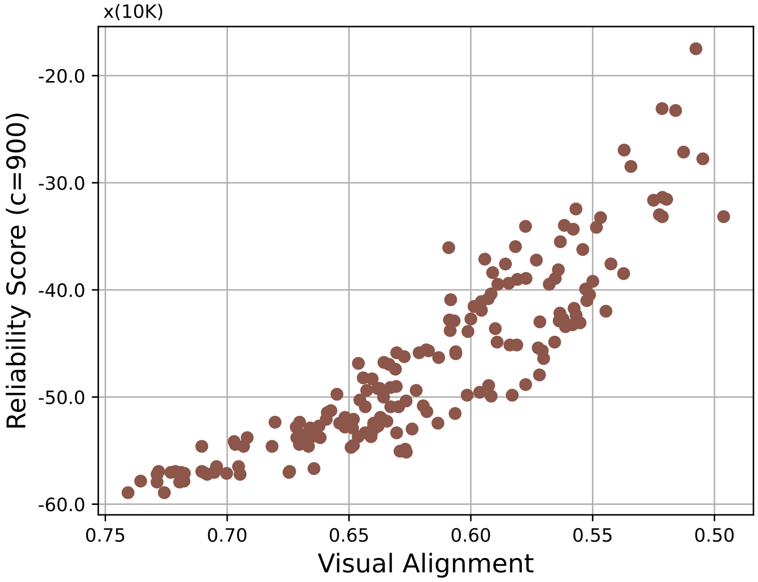


Figure 3: Correlation between Visual Alignment Distance and Reliability Score (RS_{900}). There exists a strong correlation between visual alignment distance and reliability score. This proves that visual alignment can be used as a proxy method for reliability.

G.2 Correlation between Visual Alignment and Reliability Score

Figure 3 shows the correlation between visual alignment distance and reliability score measured in Table 3. There exists a strong correlation between visual alignment distance and reliability score – the shorter the distance the higher the reliability score. This indicates that visual alignment score can be used as a proxy method for reliability, underscoring the importance of visual alignment.

H Experiment Results from Pre-training and Self-supervised Learning

Previous studies [1, 75, 23, 44] suggest that training on larger data and pre-training by self-supervised learning (SSL) methods help improve robustness and Out-of-Distribution (OOD) detection. To validate if the same findings can also be applied in our task, we additionally measure the visual alignment and reliability score on models that are pre-trained on ImageNet [60] and pre-trained by two popular SSL methods, which are SimCLR [6] and BYOL [21]. For models that are pre-trained on ImageNet, after pre-training, we initialize the top classification layer and train on our train set while freezing the pre-trained parameters during fine-tuning. For models that are pre-trained by SSL methods, we do not freeze any layers after pre-training.

The results are shown in Table 6 and Table 7. The results in Table 6 can be compared to the results in Table 3. For ImageNet pre-trained models, Transformer-based models show improved performance, whereas MLP-based and CNN-based models show similar or decreased visual alignment scores, especially when evaluated with SP. This indicates that the effect of pre-training on larger datasets is dependent on model architecture. Interestingly, distance-based abstention functions display higher visual alignment scores. We suspect that the improved output embeddings from pre-training enable distance-based abstention functions to capture more precise features. Deep Ensemble has better visual alignment when met with Transformer-based and MLP-based. Notably, Transformer-based models combined with KNN have the best visual alignment score. We conjecture the reason comes from both the model architecture and the abstention function. Contrary to CNN-based models, Transformer-based models are able to capture global features of images instead of only local features. Also, KNN calculates abstention probability based on the distance between samples instead of clusters, as done in MD or TAPUDD, which uses more fine-grained features for deciding abstention. Therefore, deciding abstention using fine-grained details on global features gets boosted when trained on a larger set,

Table 6: Average and standard deviation of ImageNet pre-trained models distance-based visual alignment and reliability score across 5 seeds. Bold indicates the best performance in each category and underline is the second best. Deep Ensemble does not have standard deviation since it is the output of 5 different seeds. For comparison, please refer to Table 3 for results without pre-training.

	Visual Alignment (↓)									Reliability score (↑)		
	Must-Act			Must-Abstain				Uncertain	Average	RS_0	RS_{450}	RS_{900}
	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8				
ViT [11]												
SP	$0.064_{\pm 0.001}$	$0.107_{\pm 0.001}$	$0.085_{\pm 0.001}$	$\mathbf{0.211}_{\pm \mathbf{0.006}}$	$0.760_{\pm 0.003}$	$0.439_{\pm 0.004}$	$0.650_{\pm 0.006}$	$\underline{0.262_{\pm 0.002}}$	$\underline{0.322_{\pm 0.002}}$	710	−77590	−155890
ASP	$\mathbf{0.033}_{\pm \mathbf{0.000}}$	$\mathbf{0.062}_{\pm \mathbf{0.001}}$	$\mathbf{0.044}_{\pm \mathbf{0.001}}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.564_{\pm 0.000}$	$0.587_{\pm 0.000}$	390	−226410	−453210
MD [35]	$0.218_{\pm 0.001}$	$0.341_{\pm 0.002}$	$0.236_{\pm 0.001}$	$0.609_{\pm 0.004}$	$0.764_{\pm 0.002}$	$0.694_{\pm 0.004}$	$0.613_{\pm 0.004}$	$0.402_{\pm 0.003}$	$0.485_{\pm 0.001}$	634	−109616	−219866
KNN [67]	$0.399_{\pm 0.001}$	$0.588_{\pm 0.001}$	$0.465_{\pm 0.001}$	$0.450_{\pm 0.001}$	$\mathbf{0.469}_{\pm \mathbf{0.001}}$	$0.556_{\pm 0.001}$	$\mathbf{0.300}_{\pm \mathbf{0.002}}$	$0.452_{\pm 0.000}$	$0.460_{\pm 0.000}$	639	**-29061**	**-58761**
TAPUDD [13]	$0.320_{\pm 0.017}$	$0.405_{\pm 0.021}$	$0.315_{\pm 0.017}$	$0.657_{\pm 0.021}$	$0.733_{\pm 0.014}$	$0.753_{\pm 0.014}$	$0.616_{\pm 0.029}$	$0.441_{\pm 0.008}$	$0.530_{\pm 0.004}$	587	−132163	−264913
OpenMax [3]	$0.042_{\pm 0.002}$	$0.068_{\pm 0.000}$	$0.049_{\pm 0.001}$	$0.728_{\pm 0.006}$	$0.868_{\pm 0.006}$	$0.750_{\pm 0.010}$	$0.820_{\pm 0.006}$	$0.420_{\pm 0.002}$	$0.468_{\pm 0.002}$	579	−138021	−276621
MC-Dropout [16]	$\underline{0.034_{\pm 0.000}}$	$\underline{0.064_{\pm 0.001}}$	$\underline{0.046_{\pm 0.001}}$	$0.909_{\pm 0.000}$	$0.964_{\pm 0.000}$	$0.927_{\pm 0.000}$	$0.947_{\pm 0.001}$	$0.519_{\pm 0.000}$	$0.551_{\pm 0.000}$	390	−226410	−453210
Deep Ensemble [32]	0.064	0.107	0.085	0.208	0.759	$\underline{0.437}$	0.649	**0.261**	**0.321**	708	−78042	−156792
Swin Transformer [38]												
SP	$0.149_{\pm 0.104}$	$0.179_{\pm 0.104}$	$0.168_{\pm 0.100}$	$\underline{0.212_{\pm 0.021}}$	$0.711_{\pm 0.073}$	$\mathbf{0.383}_{\pm \mathbf{0.060}}$	$0.637_{\pm 0.089}$	$0.319_{\pm 0.016}$	$0.344_{\pm 0.010}$	**737**	−44263	−89263
ASP	$0.083_{\pm 0.067}$	$0.105_{\pm 0.068}$	$0.099_{\pm 0.064}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.599_{\pm 0.028}$	$0.610_{\pm 0.029}$	383	−229567	−459517
MD [35]	$0.127_{\pm 0.053}$	$0.183_{\pm 0.048}$	$0.143_{\pm 0.051}$	$0.759_{\pm 0.004}$	$0.854_{\pm 0.003}$	$0.851_{\pm 0.002}$	$0.667_{\pm 0.006}$	$0.485_{\pm 0.026}$	$0.509_{\pm 0.022}$	537	−156963	−314463
KNN [67]	$0.293_{\pm 0.029}$	$0.371_{\pm 0.024}$	$0.344_{\pm 0.023}$	$0.280_{\pm 0.002}$	$\underline{0.573_{\pm 0.001}}$	$0.460_{\pm 0.002}$	$\underline{0.386_{\pm 0.002}}$	$0.374_{\pm 0.013}$	$0.385_{\pm 0.011}$	$\underline{732}$	−33468	−67668
TAPUDD [13]	$0.181_{\pm 0.041}$	$0.220_{\pm 0.038}$	$0.189_{\pm 0.040}$	$0.850_{\pm 0.006}$	$0.846_{\pm 0.008}$	$0.926_{\pm 0.004}$	$0.742_{\pm 0.017}$	$0.540_{\pm 0.026}$	$0.562_{\pm 0.019}$	421	−211979	−424379
OpenMax [3]	$0.092_{\pm 0.071}$	$0.116_{\pm 0.072}$	$0.107_{\pm 0.069}$	$0.762_{\pm 0.007}$	$0.815_{\pm 0.010}$	$0.727_{\pm 0.021}$	$0.800_{\pm 0.012}$	$0.476_{\pm 0.029}$	$0.487_{\pm 0.031}$	585	−135315	−271215
MC-Dropout [16]	$0.086_{\pm 0.068}$	$0.110_{\pm 0.069}$	$0.104_{\pm 0.065}$	$0.910_{\pm 0.000}$	$0.946_{\pm 0.007}$	$0.921_{\pm 0.004}$	$0.932_{\pm 0.005}$	$0.548_{\pm 0.027}$	$0.570_{\pm 0.027}$	383	−229567	−459517
Deep Ensemble [32]	0.178	0.206	0.195	0.214	0.703	**0.383**	0.634	0.322	0.354	701	−79849	−160399
DenseNet [27]												
SP	$0.535_{\pm 0.375}$	$0.553_{\pm 0.356}$	$0.561_{\pm 0.344}$	$0.673_{\pm 0.190}$	$0.746_{\pm 0.090}$	$0.735_{\pm 0.106}$	$0.733_{\pm 0.118}$	$0.609_{\pm 0.226}$	$0.643_{\pm 0.223}$	361	−135089	−270539
ASP	$0.503_{\pm 0.400}$	$0.517_{\pm 0.386}$	$0.521_{\pm 0.379}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.777_{\pm 0.131}$	$0.789_{\pm 0.162}$	172	−326528	−653228
MD [35]	$0.567_{\pm 0.316}$	$0.600_{\pm 0.281}$	$0.578_{\pm 0.305}$	$0.788_{\pm 0.123}$	$0.817_{\pm 0.116}$	$0.821_{\pm 0.109}$	$0.752_{\pm 0.099}$	$0.634_{\pm 0.174}$	$0.695_{\pm 0.187}$	209	−305791	−611791
KNN [67]	$0.575_{\pm 0.329}$	$0.604_{\pm 0.297}$	$0.597_{\pm 0.302}$	$0.697_{\pm 0.032}$	$0.723_{\pm 0.039}$	$0.716_{\pm 0.038}$	$0.710_{\pm 0.023}$	$0.606_{\pm 0.130}$	$0.654_{\pm 0.146}$	489	−47661	−95811
TAPUDD [13]	$0.655_{\pm 0.201}$	$0.660_{\pm 0.204}$	$0.636_{\pm 0.230}$	$0.853_{\pm 0.038}$	$0.840_{\pm 0.064}$	$0.855_{\pm 0.046}$	$0.791_{\pm 0.053}$	$0.696_{\pm 0.093}$	$0.748_{\pm 0.105}$	227	−286873	−573973
OpenMax [3]	$0.512_{\pm 0.394}$	$0.529_{\pm 0.378}$	$0.535_{\pm 0.367}$	$0.806_{\pm 0.100}$	$0.847_{\pm 0.059}$	$0.832_{\pm 0.059}$	$0.825_{\pm 0.054}$	$0.674_{\pm 0.154}$	$0.695_{\pm 0.194}$	216	−300384	−600984
MC-Dropout [16]	$0.512_{\pm 0.387}$	$0.543_{\pm 0.349}$	$0.547_{\pm 0.343}$	$0.961_{\pm 0.043}$	$0.963_{\pm 0.040}$	$0.962_{\pm 0.041}$	$0.963_{\pm 0.039}$	$0.755_{\pm 0.156}$	$0.776_{\pm 0.175}$	172	−326528	−653228
Deep Ensemble [32]	0.566	0.575	0.579	0.713	0.794	0.781	0.778	0.622	0.676	583	−132617	−265817
ConvNeXt [39]												
SP	$0.330_{\pm 0.393}$	$0.359_{\pm 0.376}$	$0.338_{\pm 0.384}$	$0.658_{\pm 0.197}$	$0.832_{\pm 0.055}$	$0.686_{\pm 0.172}$	$0.819_{\pm 0.064}$	$0.517_{\pm 0.246}$	$0.567_{\pm 0.235}$	369	−237681	−475731
ASP	$0.314_{\pm 0.402}$	$0.335_{\pm 0.391}$	$0.321_{\pm 0.395}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.999_{\pm 0.001}$	$0.685_{\pm 0.155}$	$0.706_{\pm 0.168}$	369	−237681	−475731
MD [35]	$0.380_{\pm 0.348}$	$0.407_{\pm 0.333}$	$0.402_{\pm 0.328}$	$0.690_{\pm 0.095}$	$0.769_{\pm 0.039}$	$0.639_{\pm 0.134}$	$0.711_{\pm 0.024}$	$0.536_{\pm 0.177}$	$0.567_{\pm 0.181}$	630	−97020	−194670
KNN [67]	$0.364_{\pm 0.369}$	$0.398_{\pm 0.352}$	$0.389_{\pm 0.348}$	$0.609_{\pm 0.123}$	$0.715_{\pm 0.039}$	$0.625_{\pm 0.082}$	$0.662_{\pm 0.099}$	$0.462_{\pm 0.127}$	$0.528_{\pm 0.107}$	716	−33934	−68584
TAPUDD [13]	$0.628_{\pm 0.168}$	$0.616_{\pm 0.176}$	$0.624_{\pm 0.167}$	$0.808_{\pm 0.073}$	$0.670_{\pm 0.050}$	$0.806_{\pm 0.083}$	$0.710_{\pm 0.032}$	$0.653_{\pm 0.104}$	$0.689_{\pm 0.101}$	235	−158165	−316565
OpenMax [3]	$0.319_{\pm 0.406}$	$0.345_{\pm 0.389}$	$0.333_{\pm 0.393}$	$0.796_{\pm 0.056}$	$0.807_{\pm 0.033}$	$0.728_{\pm 0.121}$	$0.802_{\pm 0.023}$	$0.537_{\pm 0.197}$	$0.583_{\pm 0.194}$	660	−85290	−171240
MC-Dropout [16]	$0.315_{\pm 0.402}$	$0.337_{\pm 0.390}$	$0.322_{\pm 0.394}$	$0.953_{\pm 0.032}$	$0.971_{\pm 0.017}$	$0.955_{\pm 0.030}$	$0.970_{\pm 0.018}$	$0.658_{\pm 0.173}$	$0.685_{\pm 0.182}$	369	−237681	−475731
Deep Ensemble [32]	0.432	0.448	0.438	0.651	0.827	0.681	0.812	0.532	0.603	593	−134407	−269407
MLP-Mixer [68]												
SP	$0.198_{\pm 0.294}$	$0.279_{\pm 0.269}$	$0.234_{\pm 0.282}$	$0.550_{\pm 0.101}$	$0.742_{\pm 0.005}$	$0.589_{\pm 0.080}$	$0.650_{\pm 0.051}$	$0.422_{\pm 0.160}$	$0.458_{\pm 0.155}$	608	−35842	−72292
ASP	$0.165_{\pm 0.295}$	$0.228_{\pm 0.281}$	$0.196_{\pm 0.286}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$	$0.686_{\pm 0.096}$	$0.659_{\pm 0.120}$	289	−268361	−537011
MD [35]	$0.303_{\pm 0.232}$	$0.391_{\pm 0.210}$	$0.339_{\pm 0.217}$	$0.726_{\pm 0.025}$	$0.710_{\pm 0.030}$	$0.685_{\pm 0.038}$	$0.706_{\pm 0.022}$	$0.535_{\pm 0.118}$	$0.549_{\pm 0.105}$	498	−121452	−243402
KNN [67]	$0.270_{\pm 0.249}$	$0.362_{\pm 0.227}$	$0.320_{\pm 0.231}$	$0.642_{\pm 0.020}$	$0.698_{\pm 0.010}$	$0.648_{\pm 0.037}$	$0.616_{\pm 0.014}$	$0.478_{\pm 0.112}$	$0.504_{\pm 0.110}$	639	$\underline{-29511}$	$\underline{-59661}$
TAPUDD [13]	$0.553_{\pm 0.115}$	$0.572_{\pm 0.120}$	$0.559_{\pm 0.114}$	$0.815_{\pm 0.015}$	$0.688_{\pm 0.010}$	$0.802_{\pm 0.018}$	$0.749_{\pm 0.036}$	$0.649_{\pm 0.072}$	$0.673_{\pm 0.046}$	331	−133319	−266969
OpenMax [3]	$0.170_{\pm 0.298}$	$0.245_{\pm 0.279}$	$0.203_{\pm 0.288}$	$0.830_{\pm 0.007}$	$0.857_{\pm 0.010}$	$0.838_{\pm 0.007}$	$0.811_{\pm 0.011}$	$0.563_{\pm 0.123}$	$0.565_{\pm 0.126}$	318	−249432	−499182
MC-Dropout [16]	$0.166_{\pm 0.295}$	$0.230_{\pm 0.280}$	$0.197_{\pm 0.285}$	$0.936_{\pm 0.017}$	$0.960_{\pm 0.004}$	$0.939_{\pm 0.015}$	$0.949_{\pm 0.010}$	$0.643_{\pm 0.108}$	$0.628_{\pm 0.127}$	289	−268361	−537011
Deep Ensemble [32]	0.310	0.369	0.338	0.558	0.736	0.595	0.652	0.440	0.500	647	−92503	−185653

which leads to the best visual alignment. The overall reliability score increases when pre-trained with ImageNet, and this represents that the models that are pre-trained on ImageNet are more likely to abstain.

As shown in Table 7, the results from SSL are highly dependent on both the model architecture and whether the abstention method is distance-based or not. For example, distance-based methods perform better on *Must-Abstain* categories when paired with Swin Transformer. Unlike other abstention methods, Deep Ensemble generally performs better in all groups regardless of the model architecture. Note that even if the same abstention method is used, the effects on the performance are reversed depending on the model architecture used. As an example, when TAPUDD is combined with Swin Transformer, the performance increases on all Must-Abstain categories and decreases on all Must-Act categories, but the performance difference is reversed when TAPUDD is combined with DenseNet instead.

Overall, Deep Ensemble helps increase visual alignment performance in both ImageNet pre-training and SSL. However, other abstention functions did not show noticeable performance increases in both cases. In short, the same findings in previous studies on robustness and OOD detection can not be directly applied to visual alignment. This implies visual alignment has its unique challenges that differentiate from robustness and OOD detection tasks, and there is much room for developing new methods for better visual alignment. In general, KNN shows the best visual alignment score in all three tables (Table 3, Table 6, Table 7). This may be due to using detailed features when calculating abstention probability. However, it is hard to find a consistency for optimal model architecture. For example, in Table 3, Swin Transformer and DenseNet, which have different architectures, have the best performance on average across all seven abstention functions. Therefore, more research on finding the optimal model architecture in visual alignment is needed.

I Discussion on Uncertainty

I.1 Continuity of Uncertainty

In this section, we will discuss a critical aspect of uncertainty which is continuity. Uncertainty is continuous and it is challenging to draw clear distinctions among classes (*i.e.*, as we mentioned in the main paper that it is hard to distinguish between "car" and "truck") and between clear and uncertain images (*i.e.*, if at least one person claims an image as "uncertain", then it becomes an uncertain image). However, as our ultimate goal is to construct a universal testing benchmark that quantitatively measures visual perception alignment between models and humans, our classes should have clear defintions so that model developers can easily prepare their models and training strategy. Therefore, after careful consideration, we used the taxonomy classification in biology which is the meticulous product of decades of efforts by countless domain experts to hierarchically distinguish each species as accurately as possible with clear definitions. Also, in order to comprehensively measure the visual perception alignment between models and humans, the models should be tested under various conditions including clear in-class images (Must-Act), clear out-of-class images (Must-Abstain) and confusing images (Uncertain). As there is no clear boundary between clear and uncertain images, the best scenario would be to survey all images in our dataset to 134 people per image to obtain numerically reliable annotations. However, surveying all images is not always feasible as it requires tremendous amount of time and money considering that there are 1800 images (900 each in the open and closed test sets) in our dataset. Therefore, due to the realistic reasons, we put significant effort to include only clear images that anyone can agree on in Must-Act and Must-Abstain and obtained human annotations on Uncertain images. Nevertheless, we also recognize that continuity is an essential characteristic of uncertainty that should be carefully considered and there is always a possibility of corner cases that may be disagreeable by at least one person. We have done our best to remove those corner case samples and cross-validated our final selection. Further detailed analysis and benchmark dataset on the continuity of uncertainty is highly needed and we will leave this as a future work.

I.2 Coverage of Uncertainty

"Uncertainty" is a broad concept and it is hard to define with one clear line and list all possible cases. In this paper, we chose 15 different types of corruptions to generate uncertainty in various ways following a concrete previous work [22]. Furthermore, to better represent the continuity of uncertainty explained in Appendix I.1, we apply the corruptions varying severity ranging from 1 to 10. Many types of corruption we used resemble the reality in their own way. For example, adjusting the brightness of the image is certainly realistic, and changing its resolution is similar to viewing an object beyond a filter (*e.g.*, semi-transparent glass), and weather changes are also certainly realistic. These corruptions result in some of realistic uncertain images, precisely 8.5% in the case of the open test set, where MTurk survey participants were struggling with differentiating between two or more animals (rather than being confused between one animal and abstention). Despite our meticulous effort, we are well aware that those corruptions certainly do not cover all possible uncertainties that arise in the real world. However, "uncertainty" is too broad to specify and diffcult to collect or generate, and hence for now we use corruptions (but sufficiently divserse types of corruptions).

J Detailed Comparisons with Previous Works

In this section, we will explain in detail how our work differs from related previous works. Our ultimate goal is to create a rigorous test (similar to tests that humans take such as college entrance exams) to quantitatively measure the visual perception gap between the models and humans across various categories. Our main interest does not lie in training but on rigorously testing the visual perception alignment. For that purpose, a dataset should satisfy the four requirements we mentioned in Section 3 and use a proper metric that reflects the visual perception alignment.

Peterson et al. [52] and Schmarje et al. [65] utilized their datasets mainly for training and did not thoroughly verify whether the model actually achieved visual alignment. Peterson et al. [52] only tested their models on in-class samples (in our case, Category 1) and out-of-class samples (in our case, Category 4 and Category 6) and they showed only accuracy and cross entropy, which is analogous to KL divergence. Therefore, they did not test their models on various possible scenarios and did not use proper measurement, as KL divergence is not an optimal choice for visual perception alignment as we described in Section 4.1. Schmarje et al. [65] only tested their models on in-class samples (in our case, Category 1) and showed accuracy and KL divergence. Therefore, although previous works trained their models with the goal of achieving visual perception alignment, none of the works have thoroughly verified how much the models have actually achieved visual perception alignment under diverse situations with an appropriate measurement.

Zhang et al. [76] and Bomatter et al. [5] are similar to our work since they show that both models and humans have better object recognition when given more contextual information, but it is difficult to say that they have comprehensively evaluated visual perception alignment. These two works only tested their models on partial aspects (in our case, Category 1, Category 2, and Category 8). Thus, these works did not test on Must-Abstain samples, which makes it difficult to claim that they "comprehensively" evaluated visual perception alignment. Zhang et al. [76] and Bomatter et al. [5] simply showed that both models and humans exhibit similar performance trends based on context (*i.e.*, when given more context, both human's and model's visual recognition performance increases), and they provided human-model correlations to describe their relative trends across conditions. However, our study on visual perception alignment is not about following human trends, but about measuring how well the model replicates human perception sample-wise. Hence, considering our research scope and criteria, it's challenging to assert that Zhang et al. [76] and Bomatter et al. [5] rigorously measured visual perception alignment.

In contrast, we quantitatively measured visual perception alignment across various scenarios with multiple human annotations on uncertain images. In addition, we borrowed Hellinger distance to precisely calculate the visual perception alignment after careful consideration of other distance-based metrics such as KL divergence and Total Variation distance. Furthermore, we incorporated specialized elements (sampling theory, statistical theories related to survey design, and experts in the related fields) in creating our dataset.

There are three key differences that distinguish our dataset compared to existing datasets that also focus on uncertainty in object recognition. First, we applied corruption and cropping with different intensities ranging from 1 to 10 to reflect the continuity of uncertainty mentioned in Appendix I.1. Uncertainty is continuous and it is critical to test models on samples where uncertainty may increase in stages. In this sense, we tested models visual perception alignment on varying degrees of uncertainty. Second, we obtained 134 human annotations per image to accurately estimate the ground truth visual perception distribution. We borrowed statistical sampling theory to achieve an error bound of lower than 5%, of which the details are in Section 3.3. Third, while our uncertain samples include uncertainty that confuses between classes, refer refer to as "inter-class uncertainty" (soft labels distributed only among target classes), we also include "recognizability uncertainty" (soft labels distributed among classes + abstention), namely whether an image itself is recognizable or not. If an image is moderately brightened (*i.e.*, intermediate phase between a clear image and a complete white image), then the object itself may or may not be recognizable. Visual perception includes not only object identification (predicting that it is an elephant) but also object recognizability (the object itself is recognizable). In this sense, we cover broader scenarios compared to previous works as we include object recognizability uncertainty in our uncertain category.

We also want to highlight that VisAlign does only contain Uncertain but also Must-Act and Must-Abstain to cover diverse scenarios as possible. In order to evaluate a model's visual perception alignment, a model should be tested under Must-Act (whether it predicts a correct class with high confidence), Must-Abstain (whether it abstains out-of-class samples with high confidence), and Uncertain (whether it reflects the human uncertainty). However, previous works are limited in that they test their model on partial cases (Category 1 and Category 4 in Peterson et al. [52], and Category 1 in Schmarje et al. [65]) which does not truly reflect visual perception alignment on various situations. It is especially important to test models on samples from out of distributions (*i.e.*, Category 5 and Category 7), but previous works overlook these samples thus did not quantitately evalute from visual perception alignment. Therefore, their dataset cannot be utilized as a benchmark to evaluate visual perception alignment. While previous papers and our work have in common with handling uncertainty, in our case, uncertain samples are a subset of our final dataset and we cover more diverse necessary situations, which previous works do not, as possible to measure the visual perception alignment.

Table 7: Average and standard deviation of self supervised distance-based visual alignment and reliability score across 5 seeds. Bold indicates the best performance in each category and underline is the second best. Deep Ensemble does not have standard deviation since it is the output of 5 different seeds. The value under each SSL performance shows its difference with the baseline's performance.

		Visual Alignment (↓)									Reliability score (↑)		
		Must-Act			Must-Abstain				Uncertain	Average	RS_0	RS_{450}	RS_{900}
		Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8				
Swin Transformer [38]													
SP	No SSL	0.106±0.004	0.362±0.014	0.221±0.017	0.793±0.016	0.828±0.043	0.800±0.022	0.829±0.028	0.625±0.031	0.571±0.015	363	−225537	−451437
	SimCLR[6]	0.125±0.009 (+0.019)	0.408±0.015 (+0.046)	0.265±0.034 (+0.044)	0.723±0.037 (-0.070)	0.782±0.026 (-0.046)	0.744±0.028 (-0.056)	0.776±0.044 (-0.053)	0.615±0.023 (-0.010)	0.555±0.015 (-0.016)	384 (+21)	−211566 (+13971)	−423516 (+27921)
	BYOL[21]	0.125±0.009 (+0.019)	0.408±0.015 (+0.046)	0.265±0.034 (+0.044)	0.723±0.037 (-0.070)	0.782±0.026 (-0.046)	0.744±0.028 (-0.056)	0.776±0.044 (-0.053)	0.615±0.023 (-0.010)	0.555±0.015 (-0.016)	384 (+21)	−211566 (+13971)	−423516 (+27921)
ASP	No SSL	0.085±0.007	0.329±0.008	0.182±0.020	0.998±0.000	0.998±0.000	0.998±0.000	0.998±0.000	0.736±0.009	0.666±0.003	294	−268356	−537006
	SimCLR[6]	0.091±0.010 (+0.006)	0.356±0.012 (+0.027)	0.214±0.028 (+0.032)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.735±0.004 (-0.001)	0.674±0.005 (+0.008)	301 (+7)	−264299 (+4057)	−528899 (+8107)
	BYOL[21]	0.091±0.010 (+0.006)	0.356±0.012 (+0.027)	0.214±0.028 (+0.032)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.999±0.000 (+0.001)	0.735±0.004 (-0.001)	0.674±0.005 (+0.008)	301 (+7)	−264299 (+4057)	−528899 (+8107)
MD [35]	No SSL	0.296±0.018	0.512±0.012	0.364±0.012	0.700±0.012	0.743±0.014	0.723±0.017	0.685±0.021	0.575±0.007	0.575±0.006	326	−248974	−498274
	SimCLR[6]	0.324±0.016 (+0.028)	0.548±0.018 (+0.036)	0.407±0.023 (+0.043)	0.638±0.017 (-0.062)	0.701±0.015 (-0.042)	0.660±0.015 (-0.063)	0.620±0.031 (-0.065)	0.558±0.011 (-0.017)	0.557±0.006 (-0.017)	373 (+47)	−216977 (+31997)	−434327 (+63947)
	BYOL[21]	0.326±0.020 (+0.030)	0.549±0.022 (+0.037)	0.408±0.026 (+0.044)	0.635±0.027 (-0.065)	0.699±0.022 (-0.044)	0.659±0.024 (-0.064)	0.615±0.039 (-0.070)	0.558±0.015 (-0.017)	0.556±0.010 (-0.018)	370 (+44)	−220580 (+28394)	−441530 (+56744)
KNN [67]	No SSL	0.370±0.017	0.580±0.008	0.456±0.018	**0.549±0.025**	0.590±0.013	**0.545±0.022**	0.554±0.035	**0.543±0.007**	0.523±0.012	**526**	**-115124**	**-230774**
	SimCLR[6]	0.375±0.019 (+0.005)	0.591±0.018 (+0.011)	0.462±0.028 (+0.006)	0.560±0.019 (+0.011)	**0.585±0.027** (-0.005)	0.558±0.021 (+0.013)	**0.537±0.022** (-0.017)	0.545±0.010 (+0.002)	0.527±0.006 (+0.003)	504 (-22)	−139446 (-24322)	−279396 (-48622)
	BYOL[21]	0.374±0.018 (+0.004)	0.591±0.018 (+0.011)	0.462±0.026 (+0.006)	0.560±0.022 (+0.011)	**0.585±0.028** (-0.005)	0.558±0.022 (+0.013)	0.538±0.023 (-0.016)	0.545±0.010 (+0.002)	0.527±0.007 (+0.003)	504 (-22)	−138996 (-23872)	−278496 (-47722)
TAPUDD [13]	No SSL	0.201±0.053	0.427±0.048	0.278±0.046	0.876±0.058	0.889±0.048	0.898±0.049	0.844±0.073	0.663±0.022	0.635±0.013	294	−268356	−537006
	SimCLR[6]	0.234±0.021 (+0.033)	0.468±0.022 (+0.041)	0.330±0.036 (+0.052)	0.834±0.026 (-0.042)	0.855±0.024 (-0.034)	0.863±0.020 (-0.035)	0.789±0.040 (-0.055)	0.644±0.010 (-0.019)	0.627±0.007 (-0.007)	301 (+7)	−264299 (+4057)	−528899 (+8107)
	BYOL[21]	0.233±0.021 (+0.032)	0.466±0.022 (+0.039)	0.329±0.037 (+0.051)	0.834±0.023 (-0.042)	0.856±0.023 (-0.033)	0.861±0.019 (-0.037)	0.791±0.038 (-0.053)	0.646±0.008 (-0.017)	0.627±0.006 (-0.008)	301 (+7)	−264299 (+4057)	−528899 (+8107)
MC-Dropout [16]	No SSL	0.099±0.008	0.358±0.013	0.225±0.029	0.831±0.037	0.810±0.023	0.817±0.032	0.724±0.084	0.656±0.030	0.565±0.011	399	−208401	−417201
	SimCLR[6]	0.107±0.008 (+0.008)	0.389±0.012 (+0.031)	0.240±0.026 (+0.015)	0.848±0.021 (+0.017)	0.825±0.026 (+0.015)	0.849±0.044 (+0.032)	0.805±0.020 (+0.081)	0.683±0.007 (+0.027)	0.593±0.006 (+0.028)	360 (-39)	−225990 (-17589)	−452340 (-35139)
	BYOL[21]	0.106±0.008 (+0.007)	0.388±0.013 (+0.030)	0.241±0.025 (+0.016)	0.856±0.015 (+0.025)	0.828±0.031 (+0.018)	0.854±0.039 (+0.037)	0.820±0.024 (+0.096)	0.683±0.007 (+0.027)	0.597±0.005 (+0.032)	355 (-44)	−228695 (-20294)	−457745 (-40544)
OpenMax [3]	No SSL	0.092±0.007	0.338±0.008	0.191±0.020	0.947±0.001	0.957±0.006	0.951±0.002	0.953±0.003	0.705±0.011	0.642±0.003	294	−268356	−537006
	SimCLR[6]	0.100±0.009 (+0.008)	0.365±0.011 (+0.027)	0.223±0.027 (+0.032)	0.940±0.005 (-0.007)	0.950±0.004 (-0.007)	0.943±0.003 (-0.008)	0.944±0.005 (-0.009)	0.701±0.003 (-0.004)	0.646±0.004 (+0.004)	301 (+7)	−264299 (+4057)	−528899 (+8107)
	BYOL[21]	0.100±0.010 (+0.008)	0.365±0.011 (+0.027)	0.223±0.027 (+0.032)	0.940±0.004 (-0.007)	0.949±0.003 (-0.008)	0.943±0.003 (-0.008)	0.945±0.006 (-0.008)	0.701±0.003 (-0.004)	0.646±0.004 (+0.004)	301 (+7)	−264299 (+4057)	−528899 (+8107)
Deep Ensemble [32]	No SSL	0.132	0.377	0.253	0.725	0.766	0.734	0.768	0.584	0.542	434	−187666	−375766
	SimCLR[6]	0.144 (+0.012)	0.428 (+0.051)	0.290 (+0.037)	0.665 (-0.060)	0.734 (-0.032)	0.692 (-0.042)	0.724 (-0.044)	0.587 (+0.003)	0.533 (-0.010)	437 (+3)	−180913 (+6753)	−362263 (+13503)
	BYOL[21]	0.144 (+0.012)	0.428 (+0.051)	0.290 (+0.037)	0.665 (-0.060)	0.734 (-0.032)	0.692 (-0.042)	0.724 (-0.044)	0.587 (+0.003)	0.533 (-0.010)	437 (+3)	−180913 (+6753)	−362263 (+13503)
DenseNet [27]													
SP	No SSL	0.094±0.017	0.258±0.023	0.183±0.019	0.813±0.017	0.852±0.015	0.819±0.012	0.864±0.036	0.614±0.008	0.562±0.007	392	−211558	−423508
	SimCLR[6]	0.296±0.034 (+0.202)	0.531±0.039 (+0.273)	0.438±0.046 (+0.255)	0.604±0.025 (-0.209)	0.653±0.038 (-0.199)	0.615±0.030 (-0.204)	0.721±0.041 (-0.143)	0.592±0.022 (-0.022)	0.556±0.005 (-0.006)	478 (+86)	−128672 (+82886)	−257822 (+165686)
	BYOL[21]	0.292±0.021 (+0.198)	0.518±0.027 (+0.260)	0.446±0.008 (+0.263)	0.600±0.022 (-0.213)	0.656±0.020 (-0.196)	0.623±0.021 (-0.196)	0.705±0.029 (-0.159)	0.583±0.027 (-0.031)	0.553±0.009 (-0.009)	448 (+56)	−157502 (+54056)	−315452 (+108056)
ASP	No SSL	**0.079±0.013**	**0.224±0.023**	**0.159±0.018**	0.997±0.000	0.997±0.000	0.997±0.000	0.997±0.000	0.747±0.008	0.650±0.004	312	−260238	−520788
	SimCLR[6]	0.216±0.025 (+0.137)	0.438±0.032 (+0.214)	0.346±0.038 (+0.187)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.780±0.012 (+0.033)	0.722±0.009 (+0.072)	264 (-48)	−282786 (-22548)	−565836 (-45048)
	BYOL[21]	0.216±0.018 (+0.137)	0.419±0.022 (+0.195)	0.354±0.008 (+0.195)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.998±0.000 (+0.001)	0.775±0.015 (+0.028)	0.720±0.005 (+0.070)	276 (-36)	−276474 (-16236)	−553224 (-32436)
MD [35]	No SSL	0.170±0.016	0.323±0.022	0.250±0.025	0.873±0.014	0.866±0.019	0.854±0.009	0.825±0.032	0.620±0.022	0.598±0.006	339	−247611	−495561
	SimCLR[6]	0.266±0.024 (+0.096)	0.468±0.028 (+0.145)	0.379±0.029 (+0.129)	0.920±0.025 (+0.047)	0.920±0.027 (+0.054)	0.910±0.029 (+0.056)	0.899±0.030 (+0.074)	0.681±0.010 (+0.061)	0.680±0.015 (+0.083)	280 (-59)	−275570 (-27959)	−551420 (-55859)
	BYOL[21]	0.262±0.016 (+0.092)	0.448±0.024 (+0.125)	0.383±0.013 (+0.133)	0.923±0.020 (+0.050)	0.925±0.017 (+0.059)	0.916±0.023 (+0.062)	0.904±0.023 (+0.079)	0.674±0.013 (+0.054)	0.679±0.008 (+0.082)	291 (-48)	−269709 (-22098)	−539709 (-44148)
KNN [67]	No SSL	0.272±0.021	0.448±0.021	0.360±0.017	0.612±0.019	0.640±0.022	0.615±0.019	0.664±0.014	0.565±0.009	**0.522±0.002**	482	−157468	−315418
	SimCLR[6]	0.369±0.014 (+0.097)	0.579±0.034 (+0.131)	0.477±0.029 (+0.117)	0.630±0.023 (+0.018)	0.653±0.029 (+0.013)	0.637±0.025 (+0.022)	0.657±0.011 (-0.007)	0.576±0.014 (+0.011)	0.572±0.010 (+0.050)	449 (-33)	−161551 (-4083)	−323551 (-8133)
	BYOL[21]	0.353±0.022 (+0.081)	0.554±0.033 (+0.106)	0.471±0.018 (+0.111)	0.662±0.029 (+0.050)	0.682±0.028 (+0.042)	0.672±0.032 (+0.057)	0.672±0.032 (+0.008)	0.583±0.015 (+0.018)	0.581±0.010 (+0.059)	418 (-64)	−183182 (-25714)	−366782 (-51364)
TAPUDD [13]	No SSL	0.310±0.039	0.393±0.025	0.364±0.044	0.862±0.021	0.831±0.023	0.837±0.018	0.810±0.028	0.645±0.017	0.631±0.004	320	−249880	−500080
	SimCLR[6]	0.352±0.023 (+0.042)	0.513±0.035 (+0.120)	0.452±0.032 (+0.088)	0.875±0.028 (+0.013)	0.866±0.034 (+0.035)	0.868±0.031 (+0.031)	0.838±0.036 (+0.028)	0.665±0.015 (+0.020)	0.679±0.010 (+0.047)	280 (-40)	−275570 (-25690)	−551420 (-51340)
	BYOL[21]	0.348±0.013 (+0.038)	0.493±0.021 (+0.100)	0.452±0.020 (+0.088)	0.879±0.013 (+0.017)	0.871±0.021 (+0.040)	0.870±0.015 (+0.033)	0.846±0.013 (+0.036)	0.664±0.017 (+0.019)	0.678±0.002 (+0.047)	292 (-28)	−269258 (-19378)	−538808 (-38728)
MC-Dropout [16]	No SSL	0.093±0.015	0.288±0.023	0.199±0.027	0.764±0.049	0.817±0.054	0.734±0.058	0.823±0.058	0.590±0.016	0.539±0.025	461	−165589	−331639
	SimCLR[6]	0.231±0.028 (+0.138)	0.473±0.033 (+0.185)	0.362±0.041 (+0.163)	0.842±0.015 (+0.078)	0.866±0.023 (+0.049)	0.836±0.016 (+0.102)	0.786±0.039 (-0.037)	0.623±0.006 (+0.033)	0.628±0.005 (+0.089)	325 (-136)	−250775 (-85186)	−501875 (-170236)
	BYOL[21]	0.232±0.015 (+0.139)	0.467±0.026 (+0.179)	0.370±0.007 (+0.171)	0.842±0.009 (+0.078)	0.862±0.014 (+0.045)	0.854±0.015 (+0.120)	0.792±0.040 (-0.031)	0.617±0.017 (+0.027)	0.629±0.010 (+0.091)	332 (-129)	−244918 (-79329)	−490168 (-158529)
OpenMax [3]	No SSL	0.087±0.014	0.263±0.024	0.204±0.017	0.953±0.003	0.953±0.002	0.954±0.005	0.964±0.004	0.718±0.009	0.637±0.003	312	−260238	−520788
	SimCLR[6]	0.243±0.025 (+0.156)	0.467±0.028 (+0.204)	0.385±0.035 (+0.181)	0.924±0.003 (-0.029)	0.922±0.002 (-0.031)	0.923±0.004 (-0.031)	0.926±0.003 (-0.038)	0.737±0.012 (+0.019)	0.691±0.007 (+0.054)	264 (-48)	−282786 (-22548)	−565836 (-45048)
	BYOL[21]	0.244±0.017 (+0.157)	0.448±0.019 (+0.185)	0.392±0.008 (+0.188)	0.923±0.002 (-0.030)	0.921±0.002 (-0.032)	0.923±0.002 (-0.031)	0.925±0.001 (-0.039)	0.731±0.015 (+0.013)	0.688±0.005 (+0.051)	276 (-36)	−276474 (-16236)	−553224 (-32436)
Deep Ensemble [32]	No SSL	0.109	0.276	0.209	0.767	0.814	0.775	0.825	0.581	0.545	396	−209754	−419904
	SimCLR[6]	0.326 (+0.217)	0.558 (+0.282)	0.462 (+0.253)	0.565 (-0.202)	0.617 (-0.197)	0.575 (-0.200)	0.689 (-0.136)	0.567 (-0.014)	0.545 (+0.000)	510 (+114)	−115590 (+94164)	−231690 (+188214)
	BYOL[21]	0.313 (+0.204)	0.536 (+0.260)	0.463 (+0.254)	0.582 (-0.185)	0.639 (-0.175)	0.605 (-0.170)	0.687 (-0.138)	0.567 (-0.014)	0.549 (+0.005)	508 (+112)	−123692 (+86062)	−247892 (+172012)