DO YOU SEE WHAT I SEE? PROBING VISUAL PERCEPTION IN ARTIFICIAL INTELLIGENCE WITH A GESTALT-THEORY DRIVEN OPTICAL ILLUSION DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning algorithms are increasingly deployed in safety-critical domains such as autonomous driving and medical diagnostics, where ambiguous data poses a fundamental challenge. Optical illusions offer a compelling window into human perceptual processing, yet remain largely absent from machine learning benchmarks. We introduce *Ambivision*, a dataset of optical illusions featuring intermingled animal pairs grounded in Gestalt psychological principles such as proximity, similarity, closure, etc. Wagemans et al. (2012), designed to evoke genuine perceptual ambiguity rather than superficial visual difficulty. Using Ambivision as a diagnostic probe, we show that widely used attribution methods (Grad-CAM, Integrated Gradients, PipNet, ACE) cannot produce *discriminative, human interpretable* explanations when both readings rest on the same pixels. We further find, across seven architectures, that supplying an explicit cue to the discriminative region (gaze direction or the eye) improves classification on ambiguous data over both no-cue and random-location baselines, whereas appending the same information as coordinates before the classifier head does not, which is an observation about *where* the usable signal lives, not a general-purpose training recipe. We release the dataset and the generation and annotation scripts. Ambivision offers a replicable, theoretically grounded methodology for studying where computational and human perception diverge. The dataset is accessible via a kaggle link which will be provided upon acceptance.

1 INTRODUCTION

We address a novel problem in the area of optical illusion: The rabbit and duck image (Figure 1) remains one of the most studied examples of perceptual ambiguities. What makes such illusions scientifically valuable is precisely what makes them computationally challenging: No single pixel or region determines the interpretation of the image. Instead, interpretation depends on the relational structure of the image: how parts are grouped, which region is assigned as figure versus ground, and how cues like gaze direction drive parsing of the whole. Take a look at Figure 1: Depending on the viewer's perception of the eye's direction, the image may be interpreted as either a rabbit or a duck. This makes such stimuli not only a natural benchmark for vision models but also an illuminating case for understanding how models represent perceptual structure more broadly. Another interesting aspect is if one were to apply current attribution-based methods Lundberg & Lee (2017b); Sundararajan et al. (2017) to explain why a model classifies the image as a rabbit or a duck, the resulting saliency maps could look identical, because the pixels relevant to both interpretations overlap entirely. The actual disambiguating signal remains hidden. The need for human-interpretable explanations grounded in real perceptual concepts was already noted in early work on LIME Ribeiro et al. (2016), and has been repeatedly adressed in works such as Fel et al. (2023a), but remains largely under-addressed in the computer science community.

Human differentiation between the two figures relies on the interpreted direction of the gaze. The novel aspect of this work is that we focus on the gaze direction of the object depicted in the image, not where the observer is gazing such as in datasets such as PET Gilani et al. (2015) . While gaze is not the only perceptual handle available, we will highlight the usefulness of adding this singular feature for the classification task in our evaluations. Our contributions are as follows:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

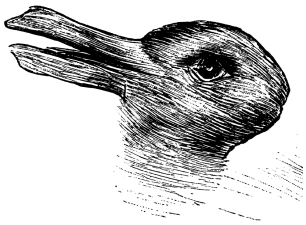

Figure 1: In this image, two interpretations are simultaneously valid: a rabbit and a duck. What determines which interpretation dominates is not any single region of the image but it can be resolved for humans using a relational cue - the perspective or direction of gaze. Shift your gaze such that you either focus on the left or the right end of the drawing and you systematically shift the perceived animal. This is precisely what makes such stimuli a meaningful challenge for vision models: the disambiguating information is not concentrated in a salient local feature but distributed across the relational structure of the image, the same kind of structure Gestalt psychology has long identified as central to human perceptual organization Koffka (2013).

- **A Gestalt-grounded diagnostic dataset.** We introduce *Ambivision*, an open-source dataset of 200+ animal optical illusions in which two animals share the same salient image region, constructed systematically from Gestalt principles rather than opportunistic collection. Each image carries fine-grained concept annotations (eye coordinates, a normalized gaze-direction vector, and per-animal bounding boxes), and we release the generation and annotation scripts. Machine generation is not merely a practical shortcut: the difficulty of producing convincing optical illusions at scale makes large human-drawn datasets infeasible, and a generative pipeline offers a valuable lens on the gap between human and AI perceptual learning.
- **A structural failure mode of attribution-based XAI.** Using Ambivision as a probe, we show that widely used explanation methods (Grad-CAM, Integrated Gradients, PipNet, ACE) cannot produce *discriminative* explanations when competing interpretations are supported by the same pixels: their attributions for the two classes coincide, so the disambiguating signal is never surfaced. We argue this is not a tuning artifact but a consequence of the shared assumption that different predictions arise from different salient features—an assumption perceptual ambiguity systematically violates.
- **Evidence that disambiguation requires concept-level, not purely spatial, information.** Across seven architectures we find that supplying an explicit cue to the discriminative region (gaze direction or the eye) measurably improves classification on ambiguous data over both no-cue and random-location baselines, whereas feeding the same information as appended coordinates before the classifier head does not. We report this as a diagnostic finding about *where* the usable signal lives, and discuss its limits (Section 5) rather than as a general-purpose training recipe.

Because the dataset is AI-generated rather than hand-drawn, we explicitly address potential biases and mitigation strategies in Section 3.1.

2 Related Work

The ability of Vision Language Models (VLMs) to reason about images in natural language raises a question that earlier discriminative architectures could not meaningfully address: how does a model respond when the image it is asked to interpret is inherently ambiguous, when no single coherent reading of the scene exists? Related work has approached the generation and evaluation of optical illusions from several directions. IllusionVQA Shahgir et al. (2024) constructs a task-comprehension benchmark of 435 instances scraped from online repositories, but its illusions are designed to be challenging while remaining unambiguous. Collections of hand-drawn illusions Williams & Yampolskiy (2018) similarly rely on existing material Boring (1930) rather than principled generation. On the generative side, adversarial illusions Bagdasaryan et al. (2024) target model robustness; Burgert et al. Burgert et al. (2024) embed hidden images as overlays; and Oliva et al. Oliva et al.

(2006) produce hybrid images by combining the low spatial frequencies of one image with the high spatial frequencies of another to conceal facial expressions. Most closely related to our approach, Geng et al. Geng et al. (2024b) generate multi-view illusions by estimating noise from multiple geometrically transformed views during the reverse diffusion process, producing images whose appearance changes under rotation or flipping. Further diffusion-based illusion works include motion hybrids with noise decomposition Geng et al. (2024a), or consider the perception under rotation of the image Sela & Elber (2007).

Our work differs from all of the above in two respects. First, the ambiguity we target is perceptual rather than geometric: our illusions do not reveal a second interpretation when physically transformed, but when the viewer's reading of a relational cue, aka the gaze direction, shifts. Second, and most critically, our generation pipeline is grounded in explicit Gestalt psychological principles rather than image-level manipulation heuristics. This grounding is what distinguishes our dataset as a diagnostic benchmark: it allows the gap between human perceptual organization and machine vision to be studied in a theoretically interpretable, reproducible way.

The perceptual ambiguity that distinguishes our dataset from prior work also exposes a structural limitation of the most widely used tools for interpreting model behavior. This distinction has direct consequences for explainability: Attribution-based methods fundamentally assume that different predictions arise from different salient features. The broader landscape of explainable AI encompasses countless methods Lundberg & Lee (2017a) from counterfactuals Mothilal et al. (2020); Antorán et al. (2020), prototypes Nauta et al. (2023); Chen et al. (2018), to global explainability methods Setzu et al. (2021); Morichetta et al. (2019); Newen & Müller (2022). We discuss and evaluate experiments with a collection of other representative methods in Section 5 to our dataset as a prominent representative of the field and show with examples that it also has problems extracting useful concepts.

3 AMBIVISION: ANIMAL OPTICAL ILLUSIONS- OUR DATASET

We introduce the dataset Ambivision, a novel dataset of animal-based optical illusions constructed through iterative prompt engineering using ChatGPT-4 and ChatGPT-4o OpenAI (2024). Designing convincing optical illusions is inherently challenging; therefore, we explicitly incorporated established psychological principles into the prompt engineering process, most notably concepts from Gestalt theory Koffka (2013). We elaborate these principles together with an example prompt in our discussion of the bias mitigation strategies in 3.1. Each image in Ambivision depicts one animal hidden inside the body of another animal, creating an intentionally ambiguous perceptual boundary. Example images can be found in Figure 2. The dataset consists of over 200 images annotated with

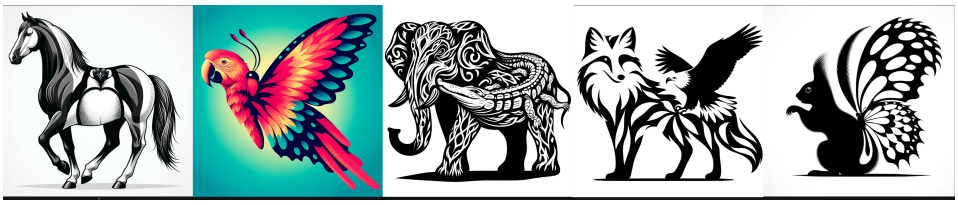

Figure 2: Example images of the novel dataset Ambivision.

the class label, eye coordinates, gaze vector and bounding boxes for both animals- providing rich, concept-level labels beyond standard object annotations. We focused on incorporating more ground truth psychological concepts such as the mentioned Gestalt theory principles Koffka (2013) rather than optimizing the illusion generation process- allowing for very high-quality illusions with regard to hidden shared features. Prioritizing psychological grounding over generation throughput resulted in high-quality illusions with deeply shared features: a number of users did not immediately identify the second hidden animal, which we take as informal evidence of genuine perceptual ambiguity. Of the resulting dataset, 41 of those images are RGB, while the majority are black and white. This design choice reflects the increased difficulty in generating convincing color illusions. Ambivision is, to the best of our knowledge, the first dataset of its kind to systematically encode perceptual ambiguity with fine-grained concept annotations, making it a valuable baseline for evaluating both

classification performance, as well as explainability in ambiguous visual settings, where the image itself is genuinely underdetermined.

We provide four different versions of the dataset. For the convenience of the user, every dataset is annotated with the label of the animals, followed by the eye coordinates (e_x, e_y), the normalized direction vector (d_x, d_y) and the bounding boxes (x_1, y_1), (x_2, y_2). We provide the same dataset with direction arrows drawn directly into the image as a baseline for this work. An additional baseline is the data set with random markings in the image and one baseline where just the eye is encircled.

3.1 BIAS MITIGATION STRATEGIES

Because this dataset was generated with ChatGPT OpenAI (2024), we had to ensure diverse results and mitigate biases. For each animal class, we confirmed that the gaze and eye position vary. This applies equally to both animals in the image to avoid introducing unintentional class-specific cues. (In the Appendix A, we include one exemplary overview of all eye positions and the spread of the eye positions for the bird class as plots)We varied artistic representation by specifically prompting for more or less realistic art styles. We specifically prompted the animals to be in alternating positions and poses, for example moving, sitting and flying. Another principle we applied for prompting is a principle borrowed from the psychological domain. The design rules for optical illusions go back to fundamental problems of psychology, such as Gestalt theory introduced as early as 1935 Koffka (2013). These laws of conceptual organization provide guidelines of what design concepts can trick our minds or make it difficult for our brains to correctly organize and interpret visual data.

One example of bias mitigation involved the owl class, where initial generations consistently produced forward-facing poses. This likely stems from owls' distinctive ability to rotate their heads 270 degrees, therefore causing the forward-facing owl to be overrepresented in the training data. We addressed this by explicitly specifying alternative poses (e.g., "flying owl") and requesting diverse orientations in our prompts, successfully increasing the diversity in the resulting examples.

Prompt Example: *Generate an artistic black-and-white image featuring a fusion of a horse and a penguin. The design blends the penguin's fur with the horse's body and therefore seemingly vanishes within. This creates an optical illusion. The image emphasizes the shared features of the two animals by incorporating concepts such as proximity, closure, figure-ground and continuation.*

4 METHODOLOGY

In this paper, we show that supplying an explicit cue to the discriminative region—the eye coordinate or the gaze direction—during training measurably improves recognition on ambiguous data, even for small sample sizes. We report this as a diagnostic finding about where the usable signal lives rather than a general-purpose training recipe: as our experiments show (Section 5), marking the eye alone is competitive with the gaze arrow, indicating that the gain stems from cueing the informative region rather than from gaze direction per se. The direction of the gaze is a problem statement often considered in various different settings in literature Mukherjee & Robertson (2015); Liu et al. (2019). Previously, it has already been recognized as an important factor in social interactions Wang et al. (2021) or even in object detection itself Bâce et al. (2017). Our work takes a novel angle by bridging both fields and explicitly distinguishing the depicted gaze direction from the observer's gaze. The goal of this new dataset is to provide an ambiguous dataset baseline that helps evaluate future XAI algorithms on their ability to provide meaningful explanations on ambiguous data, and also to have a "hard to solve" optical illusion benchmark for the evaluation of vision models.

A key distinction for this proposition of including the gaze, in contrast to existing ones, is that we track where the animal or human is looking, not what the person looking at. We consider gaze as: The gaze direction starting from the eye coordinate and then considering the head tilt, the direction of the gaze. The mathematical definition consists of the following components:

- **The Eye Position (e)**: A 2D point representing the eye's position in the plane, denoted as a vector $[e_x, e_y] \in \mathbb{R}^2$.

- **Gaze Looking Direction (d)**: A unit vector representing the normalized direction in which the head is oriented, given as $[d_x, d_y] \in \mathbb{R}^2$.

We can then define the gaze direction as $g = e + \alpha \cdot d$, with $\alpha \in \mathbb{R}$ a scaling constant to ensure a set length. For explanation purposes, we make it intuitively long enough to be well visible to the human eye. α is given a length based on the given image size. We normalize the gaze, ensuring a unified vector notation. For the sake of this evaluation methodology, we define that looking straight ahead is annotated as (0.0, 0.0). We always annotated the position of the right eye if two were present.

Our contribution extends this capability by systematically incorporating psychological principles into the generation process. Rather than reproducing the entire diffusion architecture, we leverage the state-of-the-art DALL-E 3 model Betker et al. (2023) to benefit from its extensive hyperparameter optimization and generation quality. Through carefully designed prompts informed by perceptual psychology, we create Ambivision: a dataset of optical illusions with controlled ambiguity and structured annotations for gaze-based disambiguation.

5 Experiments: Benchmarking the concept of gaze direction and the eye on Ambivision- our dataset

We evaluate learning improvements when including gaze direction and eye annotations, keeping all other conditions identical. Importantly, we focus on the *depicted* gaze direction of the object in the image (e.g., which way an animal is looking), not the observer's gaze as in eye-tracking literature Saab et al. (2019); Wang et al. (2017). We fine-tune several pre-trained ImageNet classifiers—ResNet18, ResNet34, ResNet50, VGG13, VGG16 He et al. (2015), ViT-b-16 Dosovitskiy et al. (2021), and Swin-v2-t Liu et al. (2022)—on learning rates $\{0.0001, 0.00001, 0.000005\}$ over varying epochs using the ADAM optimizer Kingma & Ba (2017) on a NVIDIA A100-SXM4-80GB GPU.

For the direction, we included arrows in the image that annotated the gaze direction in the training set, and validated it allowing either animal to be classified. Figure 3 shows the accuracies when allowing both animal classes to be correct. Despite allowing both classes as correct, in all cases, including the direction leads to significantly higher accuracies with otherwise the same hyperparameters. We include graphs with all tested learning rates in the Appendix A. As baseline check, we tested the same by annotating random other areas in the image, to double check that the concept incorporated was meaningful. We also include one experiment where we tested how eye annotation alone performs for vision models on our dataset. Our experiments show that both gaze direction and eye annotation alone lead to meaningful improvements, raising accuracy rates by over 20% in a setting allowing up to 1000 classes. Our results shown here were generated using the ADAM optimizer Kingma & Ba (2017), because initial experiments revealed to us that this was the optimizer that produced the best accuracy results in practice, and it is known to be state-of-the-artChoi (2019). All experiments were performed using a NVIDIA A100-SXM4-80GB GPU. Calculating lower epochs took seconds, the overall plotting of all accuracies up to 1.000 training epochs took several hours.

Marking the eye outperformed all other tests, however it should be noted that while the eye is a meaningful concept, this dataset did not necessarily ensure uniqueness of looking direction, but always a unique combination of eye coordinate and gaze vector. The interesting part of this experiment was that this eye annotation did not have to be human-level accurate. An overview of those results can be seen in Figure 3. We evaluated two different approaches for the improvement of the optimization problem $\arg\min_{g \in G}[L(f, g, \pi_{x'}, e, d) + \Omega(g)]$: (1) concatenating gaze vector and eye coordinates before the final softmax layer and learning via an MLP Popescu et al. (2009), optimizing:

$$L(f, g, \pi_{x'}, e, d) = L\left(L_{task}(f, g, \pi_{x'})|(e, d)\right) \tag{1}$$

and (2) incorporating gaze and eye annotation directly into the image by highlighting areas around the eye and in gaze direction. Strategy (1) yielded no detectable accuracy gain; strategy (2) produced significant improvements. As can be seen in Figure 3, both eye coordinates and gaze direction led to significant improvements over the baseline of no annotation and random annotations. All other results for other learning rates can be found in the Appendix A, but all show the same general results.

We preprocessed our dataset such that the shape of the images was adjusted to $3 \times 224 \times 224$ and normalized using the mean. As a further test, we took a look at what exactly our network learns if

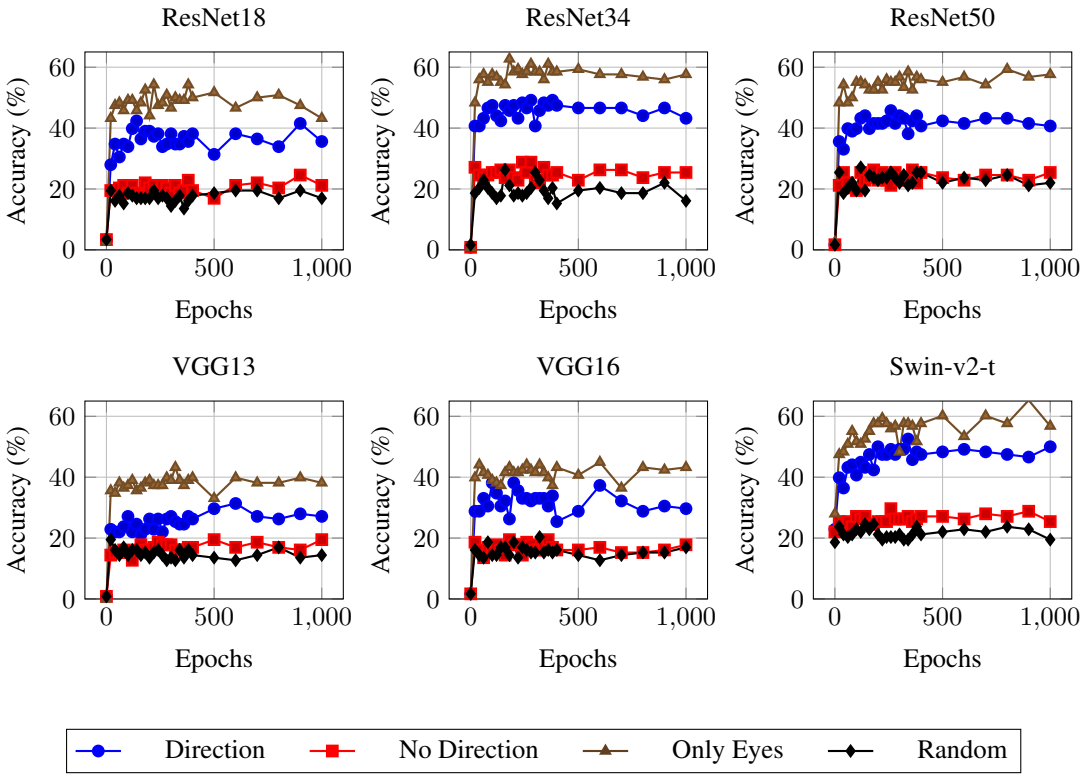

Figure 3: Accuracy vs. epochs at LR=10^{-4} for ResNet, VGG, and Swin-v2-t. Adding an explicit cue to the discriminative region, either the gaze arrow (*Direction*) or the eye (*Only Eyes*), yields higher accuracy than the no-cue (*No Direction*) and *Random*-location baselines across models. Notably, marking the eye alone is competitive with or stronger than the gaze arrow (see text), indicating the gain comes from cueing the informative region rather than from gaze direction *per se*. Remaining architectures and learning rates are in Appendix A.

we directly incorporate the direction arrow as well using LIME Ribeiro et al. (2016): As you can see in Figure 4, the network then highlights normal features but also the gaze vector as a feature, compared to worse features that overlap with the rivaling class and no highlighting of the eyes. We include another example of this as an interesting additional observation for GradCAM Selvaraju et al. (2017) as well as for PipNet Nauta et al. (2023) in the Appendix A.

6 Ambivision as a diagnostic: observations on explanation models

Attribution-based explanations rest on an implicit assumption: that different predictions are driven by different salient features, so highlighting the influential region tells the user *why* one class was chosen over another. Ambivision is constructed to violate exactly this assumption. When two animals occupy the same pixels, the two competing predictions are supported by the same region, and any purely spatial explanation of one class is, by construction, almost indistinguishable from the explanation of the other. We use this property as a diagnostic: rather than asking whether an attribution method is *faithful*, we ask whether it is *discriminative*—whether its explanation for "class A" differs meaningfully from its explanation for "class B" on the same image. We examine four representative methods spanning the main families—Grad-CAM (Selvaraju et al., 2017) and Integrated Gradients (Sundararajan et al., 2017) (saliency-based), PipNet (Nauta et al., 2023) (prototype-based), and ACE (Ghorbani et al., 2019) (automatic concept-based)—each applied without modification. We stress that this is a qualitative, illustrative analysis on selected examples; we do not claim a quantitative failure rate, but rather surface a structural difficulty that the shared salient region makes visible.

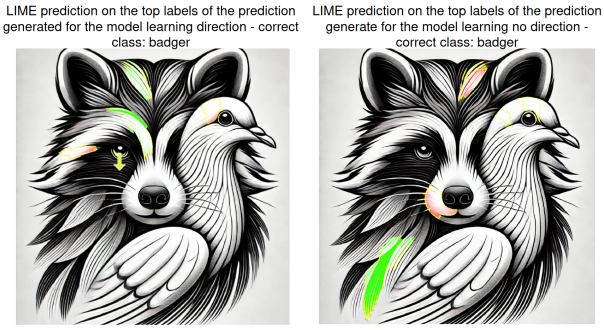

Figure 4: We see here that the LIME Ribeiro et al. (2016) explanation of the badger trained exactly the same as the other model with the exception of the direction vector is able to include fewer features shared by badger and pigeon, such as more area around the pigeon's eyes and where the fur of pigeon and badger overlap. The model also directly accounts for the eye gaze as well, which shows the usefulness of the feature in the learning process. We also made sure the gaze of the animal varied during the training. So, the animal cannot be classified simply because the arrow always points in the same direction.

Saliency methods produce near-identical maps for competing classes. For Grad-CAM (Figure 5), the attributions for the two animals overlap heavily: the method shrinks the important region for one class but continues to highlight features belonging to both, so the map does not reveal which cue separates the interpretations. Integrated Gradients (Figure 6) shows the same pattern, with the two class explanations marking largely the same pixels—in several cases the "bear" explanation attributes importance to regions that belong to the other animal. In both cases the disambiguating signal a human would use is not surfaced, consistent with the shared-region assumption above.

Prototype and concept methods inherit the same limitation. PipNet (Figure 7) retrieves prototypes that conflate visually similar textures of the two animals (e.g. cheetah fur and eagle feathers), because its prototypes are local patches rather than object-level concepts. ACE (Figure 8) produces segment clusters that do not correspond to interpretable concepts and, crucially, cannot recover abstract, relational cues such as gaze direction or eye location, since its pipeline groups pixel-level segments—a limitation shared by related concept-extraction methods (Fel et al., 2023a; Zhang et al., 2021; Fel et al., 2023b). Across all four methods, the common failure is not a lack of faithfulness but a lack of *discriminability*: the explanations do not distinguish the two valid readings, which is precisely what a human needs on ambiguous input.

The effect is not a dataset shortcut. A natural worry is that these observations reflect a spurious shortcut in the data rather than a genuine property of the methods. To check this, we run a RemOve-And-Retrain (ROAR) analysis (Hooker et al., 2019): we rank pixels by Integrated-Gradients importance, remove up to 60%, retrain, and measure test accuracy (Figure 9). If a single localized region were driving classification, removing the top-ranked features would collapse accuracy early. Instead, accuracy remains stable and only declines once more than half of the ranked features are removed, indicating that no small salient region acts as a shortcut—the discriminative signal is distributed, consistent with the perceptual, relational nature of the task.

7 Conclusion

Using classical optical illusions such as the rabbit-duck figure as motivation, we introduced *Ambivision*, a Gestalt-grounded dataset of visually merged animal optical illusions in which two valid interpretations rest on the same salient region. Each image is annotated with (1) animal class labels, (2) eye coordinates, (3) normalized gaze-direction vectors, and (4) per-animal bounding boxes.

Using the dataset as a diagnostic, we make two observations. First, widely used attribution methods (Grad-CAM, Integrated Gradients, PipNet, ACE) fail to produce discriminative human interpretable explanations on ambiguous data: because both readings are supported by the same pixels, the at-

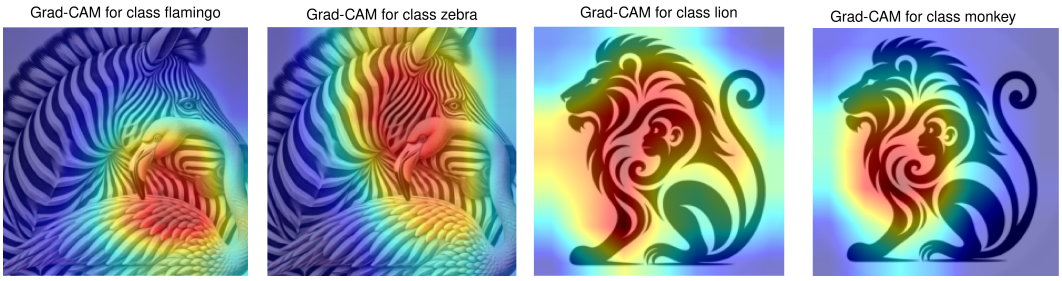

Figure 5: Pixel-based attribution like Grad-CAM struggle to separate the two animals: although the important region shrinks for one class, it still covers features belonging to both, so the map does not reveal which cue distinguishes the interpretations.

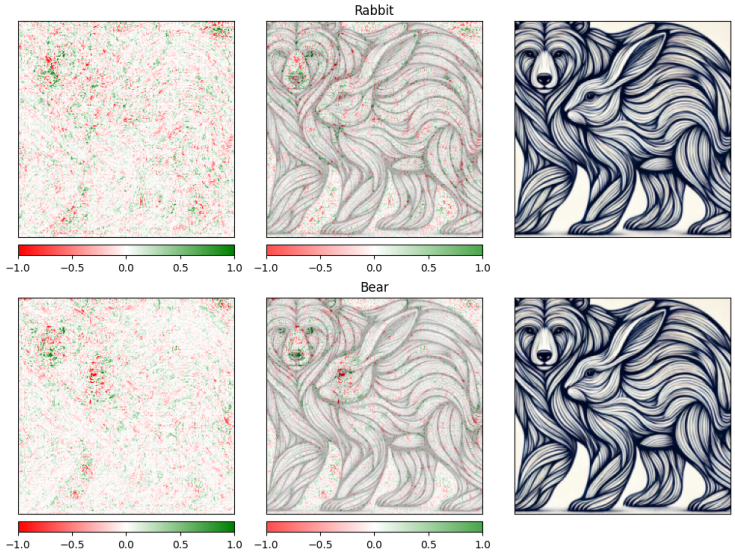

Figure 6: Integrated Gradients also struggle in marking human interpretable features: The attributions for the classes are very similar, and often not very sensible. The explanation for bear clearly marks the rabbits' head in the picture. Both the rabbit and bear explanation include markings on the bear face area. Clearly, the model struggles to distinguish the two animals, and the explanations are limited in their meaningfulness and clarity.

tributions for the competing classes coincide, and the disambiguating signal a human would use is never surfaced. Second, across seven architectures, supplying an explicit cue to the discriminative region—the eye or the gaze arrow—improves accuracy on ambiguous data over no-cue and random-location baselines, whereas appending the same information as coordinates before the classifier head does not. We report this as a diagnostic finding about *where* the usable signal lives.

Taken together, these results argue that meaningful explanations in ambiguous settings must move beyond pixel-level saliency toward abstract, relational concepts such as gaze direction and eye location—concepts that current automatic concept-discovery methods, constrained by segmentation-based clustering, do not recover. We see Ambivision as a step toward studying where computational and human perception diverge, and hope it prompts further inquiry: which abstract concepts should vision models learn, and can the psychological study of perception offer guidelines for building models that perceive and explain visual information in more human-aligned ways?

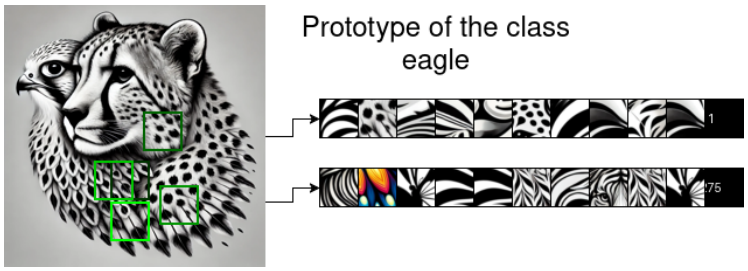

Figure 7: In this image, we see prototypes extracted via Pipnet Nauta et al. (2023) for the eagle class. Pipnet also struggles to distinguish the cheetah fur and the eagle feathers. The darker the box, the more important it is for the classification. Again, we argue that this is due to the area-based and not concept-based explanations, a clear limitation for ambiguous data. Next to the original image, we see example concepts extracted that should show similar features.



Figure 8: Example concepts extracted by ACE Ghorbani et al. (2019) for the bird class. We argue that ACE cannot find abstract concepts, such as gaze direction, because it clusters segmentations on a pixel-based level. We argue that we are currently missing concept-based XAI that goes beyond the grouping of pixels.

8 Reproducibility Statement:

The Ambivision dataset is available under a kaggle link provided upon acceptance and left out due to blind review. Our code will be published after acceptance, but is included in the submission.

AI use statement

In this work, we used generative AI tools for grammatical/ spelling checks. Additionally, we used generative AI tools for the generation of the dataset Ambivision. We have reviewed all AI-assisted work, including discussing potential biases in the dataset. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

Ambivision contains only AI-generated images of animals; it involves no human subjects, no personally identifiable information, and no sensitive or private data, and the classification task carries no direct risk of harm. The main ethical consideration is that the images are produced by generative models (ChatGPT and DALL-E 3) and therefore inherit whatever biases exist in those models' training data—for example, over-represented poses or viewpoints for particular animal classes. We treat this as a genuine limitation rather than a non-issue: Section 3.1 details the mitigation strategies we applied (varying pose, orientation, and artistic style, and balancing gaze and eye position across both animals), and we release the generation prompts and annotation scripts so that these biases can be independently audited and the dataset extended or corrected. Finally, the dataset is intended as a diagnostic benchmark for studying perception and explanation methods; it is not designed for, and should not be deployed in, safety-critical decision-making without further validation.

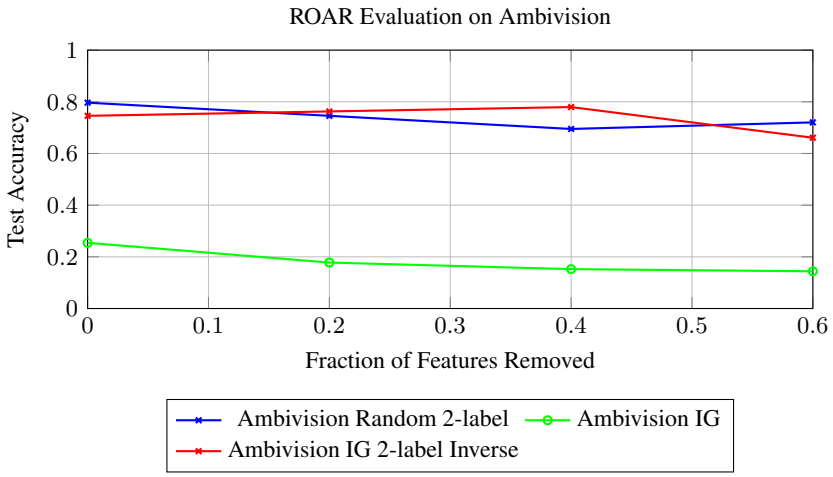

Figure 9: ROAR Evaluation comparing different feature removal strategies, IG = Integrated Gradients Sundararajan et al. (2017), 2-label= both animal classes allowed as correct

Reproducibility statement

For questions regarding the reproducibility and all experiments, especially check the Appendix. We included some more examples, discuss biases and provide all experiments.

References

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a clue: A method for explaining uncertainty estimates. *arXiv preprint arXiv:2006.06848*, 2020.

Mihai Bâce, Philippe Schlattner, Vincent Becker, and Gábor Sörös. Facilitating object detection and recognition through eye gaze. In *19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2017)*. ETH Zurich, 2017.

Eugene Bagdasaryan, Rishi Jha, Vitaly Shmatikov, and Tingwei Zhang. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 3009–3025, 2024.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023. OpenAI.

Edwin G Boring. A new ambiguous figure. *The American Journal of Psychology*, 1930.

Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *CoRR*, abs/1806.10574, 2018. URL `http://arxiv.org/abs/1806.10574`.

D Choi. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.

Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36:54805–54818, 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023b.

Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*, pp. 255–258, 2014.

Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024a.

Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24154–24163, 2024b.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6912–6921, 2019.

Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. Animalweb: A large-scale hierarchical dataset of annotated animal faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6939–6948, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013.

Gang Liu, Yu Yu, Kenneth A Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):1092–1099, 2019.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. URL https://arxiv.org/abs/2111.09883.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017a. URL https://arxiv.org/abs/1705.07874.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017b.

Andrea Morichetta, Pedro Casas, and Marco Mellia. Explain-it: Towards explainable ai for unsupervised network traffic analysis. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pp. 22–28, 2019.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.

Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753, 2023.

Carina Newen and Emmanuel Müller. Unsupervised deepview: Global explainability of uncertainties for high dimensional data. In *2022 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 196–202. IEEE, 2022.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics (TOG)*, 25(3):527–532, 2006.

OpenAI. Chatgpt, 2024. URL `https://www.openai.com/chatgpt`. Retrieved from OpenAI.

Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7): 579–588, 2009.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

Khaled Saab, Jared Dunnmon, Alexander Ratner, Daniel Rubin, and Christopher Ré. Improving sample complexity with observational supervision. 2019.

Guy Sela and Gershon Elber. Generation of view dependent models using free form deformation. *The Visual Computer*, 23(3):219–229, 2007.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx-from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021.

Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. Illusionvqa: A challenging optical illusion dataset for vision language models. *arXiv preprint arXiv:2403.15952*, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

Xin Wang, Nicolas Thome, and Matthieu Cord. Gaze latent support vector machine for image classification improved by weakly supervised region selection. *Pattern Recognition*, 72:59–71, 2017.

Xinming Wang, Jianhua Zhang, Hanlin Zhang, Shuwen Zhao, and Honghai Liu. Vision-based gaze estimation: A review. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):316–332, 2021.

Robert Max Williams and Roman V Yampolskiy. Optical illusions images dataset. *arXiv preprint arXiv:1810.00415*, 2:2, 2018.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11682–11690, 2021.

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 365–381. Springer, 2020.

A Appendix

We promised to deliver additional material for various points in the paper for whomever it may interest. First, we show an exemplary overview of existing types of gaze annotation datasets and their purposes for a literature overview in Table 1. In the following, we show the spread of the eye coordinates of the class "bird" and then a scatter plot of all eye coordinates in total.

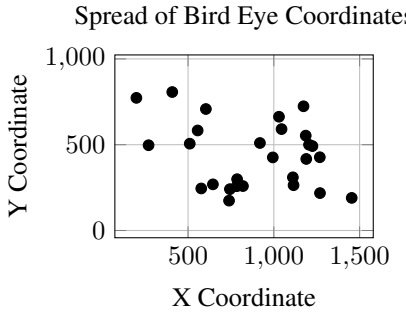

Figure 10: Scatter plot showing the distribution of bird eye coordinates.

To show that the eye coordinates in general are not placed in one area, we show here the distribution of all eye coordinates in one collected plot in Figure 11. The only tendency visible is that more of them are in the center areas than the edges, which makes sense because the body of an animal is always around the eye.

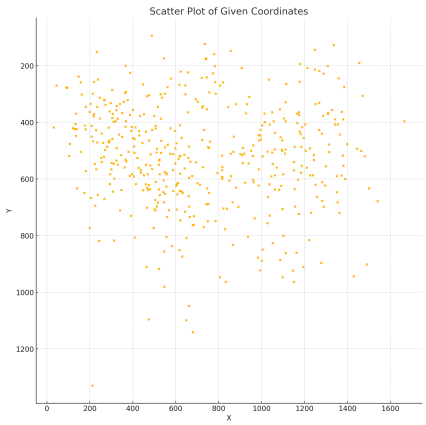

Figure 11: This scatter plot shows the spread of all eye coordinates from all classes. It is pretty diverse, considering the eyes are placed within the body of the animal and can therefore not occur on the immediate sides.

Table 1: An exemplary Overview of existing Types of Gaze Annotation Datasets and their purposes.

Dataset Name	***Type of Data***	**Type of Annotation**
Cathegory: Human Gaze Annotations		
1. MPIIGaze Zhang et al. (2015)	large-scale dataset of Human Images of 15 subjects	Focus on different lightings, variable places and day times, as natural settings as possible
2. EyeDiap Funes Mora et al. (2014)	human gaze estimation dataset of 16 test subjects changes in ambient and sensing conditions	head pose variations, gaze poses ground truth given by the 3d poses of the visual target
3. ETH-XGaze Zhang et al. (2020)	human gaze estimation dataset of 110 participants	head pose variations and different lighting
4. Gaze360 Kellnhofer et al. (2019)	human gaze estimation dataset of 238 subjects	wide range of lighting conditions
Cathegory: Animal Gaze Annotations		
5. Animal Kingdom Dataset Ng et al. (2022)	Video Dataset of Animals	annotated for relevant animal behavior, no eye gaze annotation but pose estimation
6. PET Gilani et al. (2015)	classes bird, cat, cow, dog, horse, and sheep tracking eye position for visual points of importance from 40 users	eye movements recorded of human points of interest for free vision task and visual search
7. AnimalWeb Khan et al. (2020)	animal faces collected from 350 species	annotated with 9 land-marks on key facial features
Our Dataset: Ambivision Optical Illusions)	Drawn Animal Images generated by ChatGPT4 and ChatGPT4o consisting of always two animals, one merged/hidden in the other animal	Gaze vector in 2D format, animal labels for both animals animals were distinguishable by their right eye coordinate and normalized direction vector. If it stared straight ahead, gaze vector was assigned as 0.0, 0.0. annotated bounding boxes

To go on, we continue by including all experimental results for different learning rates and architectures in larger format in this Appendix to demonstrate the validity of our results.

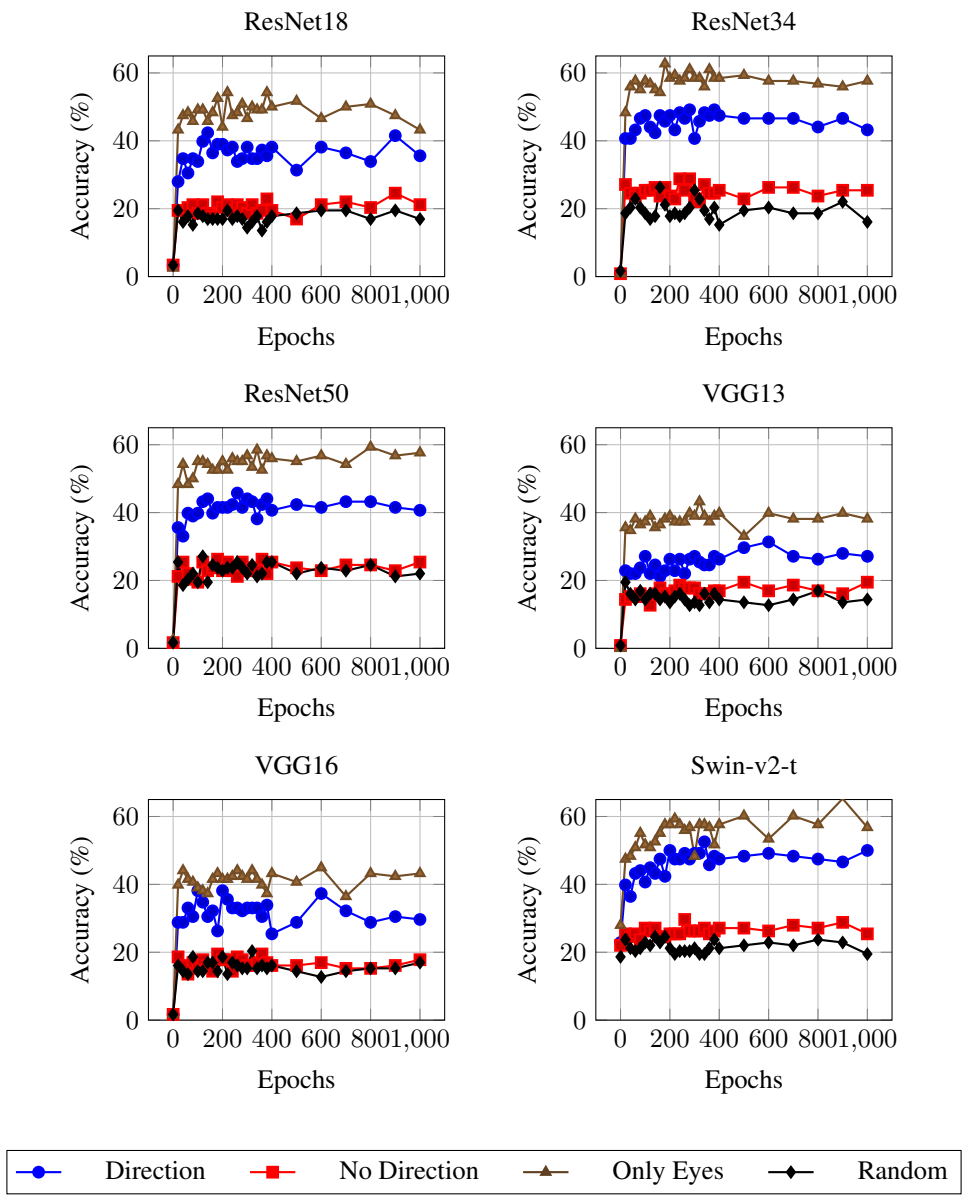

Figure 12: Accuracy vs. Epochs at LR=0.0001 for ResNet, VGG and Swin-v2-t models across different annotation conditions.

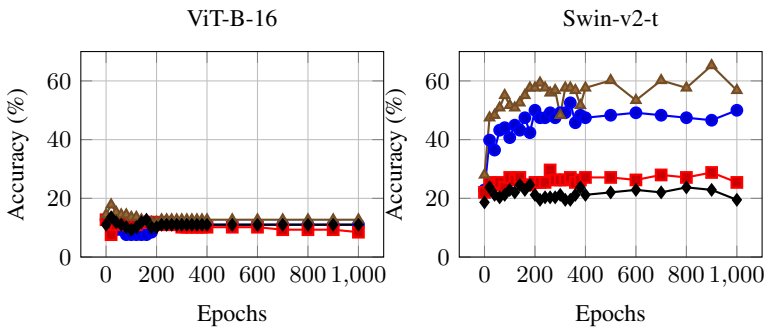

Figure 13: Accuracy vs. epochs at LR=10^{-4} for ViT-B-16 and Swin-v2-t transformer models across different annotation conditions.

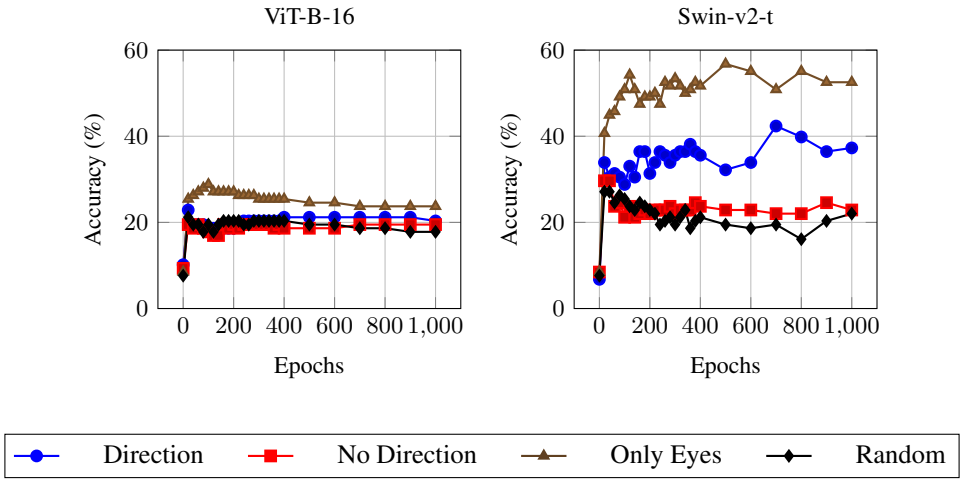

Figure 14: Accuracy vs. Epochs at LR=1×10^{-5} for ViT-B-16 and Swin-v2-t transformer models across different annotation conditions.

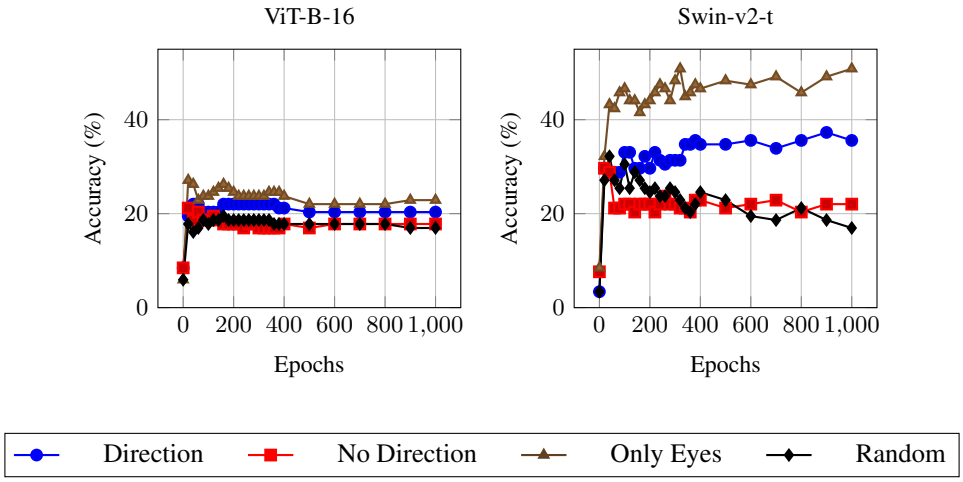

Figure 15: Accuracy vs. Epochs at LR=5×10^{-6} for ViT-B-16 and Swin-v2-t transformer models across different annotation conditions.

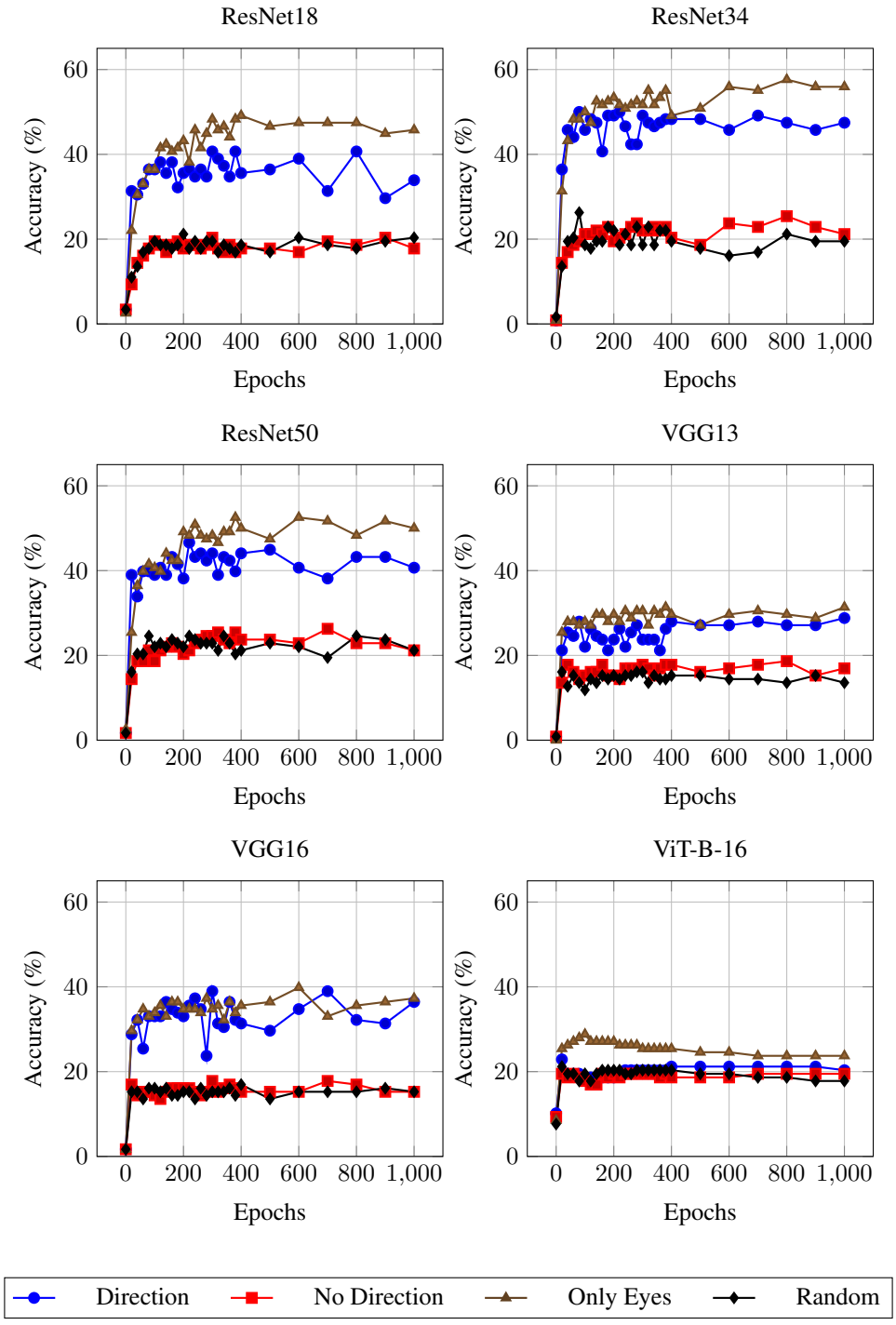

Figure 16: Accuracy vs. Epochs at learning rate 1×10^{-5} for ResNet, VGG models and ViT-B-16 under various annotation strategies.

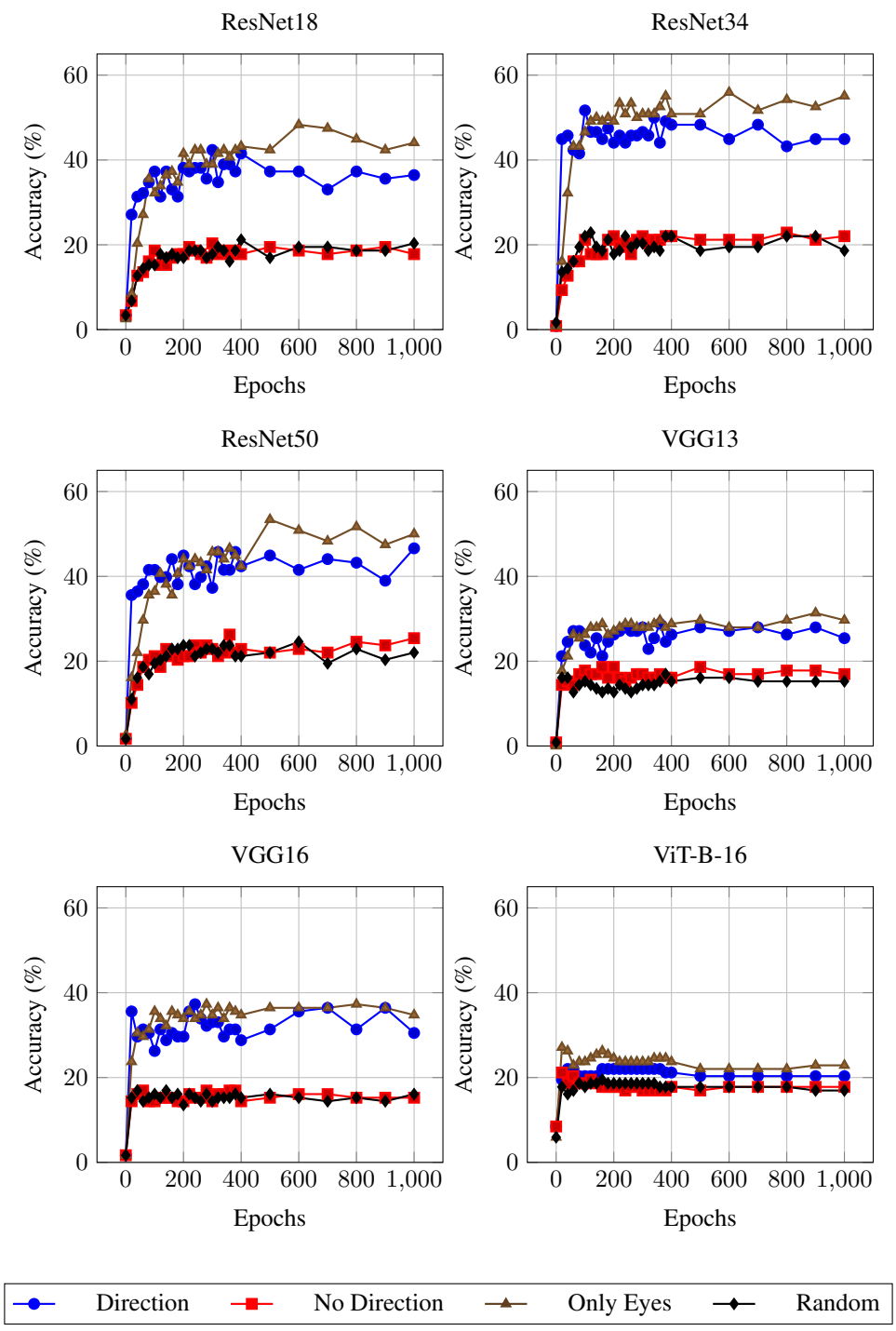

Figure 17: Accuracy vs. Epochs at learning rate 5×10^{-6} for ResNet and VGG models under various annotation strategies.

Furthermore, we discussed that the explanations generated when including direction as a concept showed more useful features for recognition. Due to page restriction, we only included one example of this phenomenon in the main paper for the algorithm LIME Ribeiro et al. (2016). However, we have other such examples, for example Gradcam Selvaraju et al. (2017) and PipNet Nauta et al. (2023), which we will show and elaborate in the following:

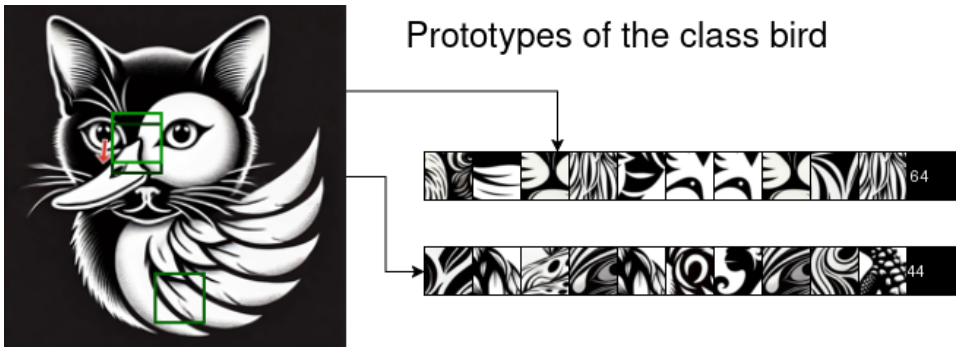

Figure 18: This image shows us the top prototypes marked by Pipnet for the class ”bird”. Again, Pipnet marks areas in the image that are important to the classification process. We can see that one of the boxes could either be the beak or the eye with the arrow, and one is focused on the feathers/wing patterns. Both are good features for the recognition of the animal. We also show some of the example patterns that Pipnet gives us for the feather pattern by which it classifies as bird. Even though this works well, this is yet another XAI method that focuses on areas instead of abstract concepts like viewing direction.

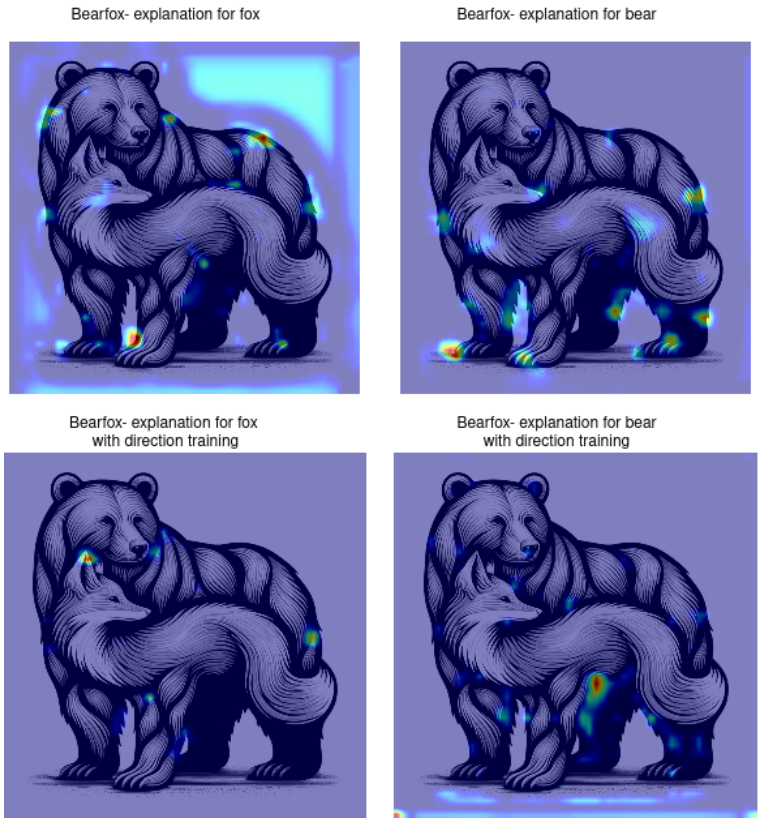

Figure 19: In this image, we can see with Gradcam on an example image how the learner was able to extract much more useful features for the model trained with the direction vector. The whiteness is removed, and for the fox explanations it focuses less on areas that are actually part of the bear. Regarding the bear explanation, less parts of the fox are highlighted.