AI Vision Systems Often Aren’t Really ‘Looking’ at All

Originally published at https://www.unite.ai/ai-vision-systems-often-arent-really-looking-at-all/

AI vision systems may be explaining images with stereotypes instead of what they actually see. Remove the label, and their supposed visual explanations collapse.

Since frontier AI platforms are constantly bleeding money in hope of an AI event horizon, little economies in their provisioning and services equal huge savings: for instance, every time an LLM can provide an answer from its own trained matrix instead of making a RAG-based journey to the web to obtain *current* information that will need distilling and refining, it answers more quickly, and saves the provider money.

Of course, it is more likely to hallucinate, without spending extra tokens and energy on contextualizing and checking its presumptions.

Likewise, even when an LLM does reach out to the web, it will very often cite vague, inappropriate, or even meaningless and unrelated URLs, because it saved energy by only evaluating the metadata for those pages, instead of actually reading them.

Similarly, if you upload a PDF to an LLM and wish to discuss it, the LLM may eventually wipe the PDF contents from its memory and use only sparse metadata about it to answer your queries, without even stating that you should upload it again – leading to errors, wrong assumptions and yet more hallucinations.

Industrial Trials

These are pragmatic, if not entirely honest economies imposed on the user for operational motives. In the upstream world of data-gathering and model training, where millions – even hundreds of millions – of images must be processed *en masse*, and cannot be hand-annotated by humans, the pressure to minimize energy and time expenditure is equally intense.

One such ‘shortcut’ is to use an intermediary Vision Language Model such as Contrastive Language Image Pretraining (CLIP), which maps both images and text into a shared semantic space, so that visual concepts can be compared using language, rather than explicit labels.

What does that mean? Well, imagine a child learning from millions of captioned pictures until they develop a rough sense of which images and words naturally go together.

Trouble is, very often, the captions were written by website owners and other self-serving entities that were more interested in good SEO than good labeling, leading to a great deal of ‘noise’ or inaccurate labeling.

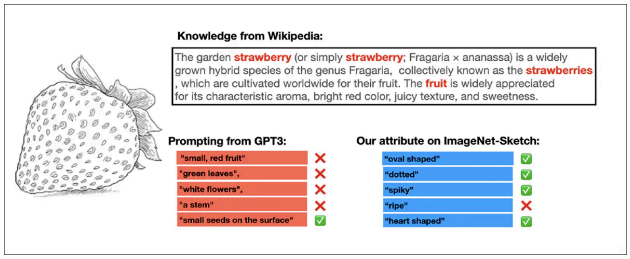
Still, the scale at which a concept and a related word recur in huge datasets usually means that the core words will get associated with the correct image, since the smaller number of ‘nonsense’ labels will normally be forced into an outlier ghetto, and will not usually gain association with the image. So it more or less works, most of the time.

Data annotation is done this way because it’s cheap and minimally functional, not because it’s good.

Off-Label

Into this troublesome scenario comes a new paper from Carnegie Mellon, illustrating starkly that many captions assigned to images – for instance by CLIP – are simply *stereotypes* based on referring to the image's (text-based) label, rather than by actively looking at the image itself.

In one striking example from the paper, CLIP is paired with descriptors generated from the ImageNet class name *strawberry*, then tested on ImageNet-Sketch, where every strawberry is a black-and-white drawing – yet the descriptors still insist the fruit is *'red'* and *'ripe'*:



Class-name descriptors fail on ImageNet-Sketch: language priors say 'red' and 'ripe', while image-grounded attributes match the black-and-white drawing. Source

Here, CLIP is paired with descriptors generated only from the class name *strawberry*, via GPT-3 or external knowledge. These never see the image, so they default to canonical traits such as *red*, and *leafy*. When applied to the monochromatic line-based illustrations in ImageNet-Sketch, the process fails.

By contrast, attributes derived from the images themselves select features that remain true under the shift, such as *dotted* or *heart-shaped*.

Thus we can see that the object identified is being advertised not for what it is, but, as it were, 'by reputation'; the adjectives 'red' and 'leafy' are accurate for the subdomain *strawberry*, but exceed the bounds of the image (instance) in question.

The authors of the new work – titled *Attributes Should Come from Images, Not Class Names: Distribution-Conditioned Attribute Selection for Vision-Language Models* – argue that this is a persistent and widespread problem, and suggest selecting attributes directly from the images themselves, so they reflect what is actually visible under distribution shift, rather than relying on class-name priors.

The authors state:

> *'A popular route to interpretable zero-shot classification asks a large language model (LLM) to describe each class name and prompts CLIP with the resulting descriptors. We show that these descriptors carry little visual evidence of their own: removing the class name from the prompt collapses ImageNet accuracy from 59.5% to 15.5%.*
>
> *'The diagnosis is that the descriptors are conditioned on the label rather than on the images, so they describe the concept in general and mislead exactly when the data shifts; an LLM insists that strawberries are red, but every strawberry in ImageNet-Sketch is a colorless line drawing.*
>
> *'We therefore select attributes from the target image collection instead: we score a large attribute pool against the images in CLIP's joint embedding space and keep the top-scoring attributes per class.'*

Their proposed remedy is that the model stays the same, but instead of relying on class-name descriptions, it checks a *pool of candidate attributes* against the images, and keeps only those that genuinely fit what is visible.

The cost of this is arguably acceptable, since it does not suggest re-invoking the saved energy of the 'cheating' methods that led to the issue in the first place. Rather, it adds a *scoring pass* over candidate attributes per class, so that latency increases slightly – but far less than retraining or full RAG-style pipelines. In theory, the energy cost would be acceptable, weighed against the improvement in alignment.

Method

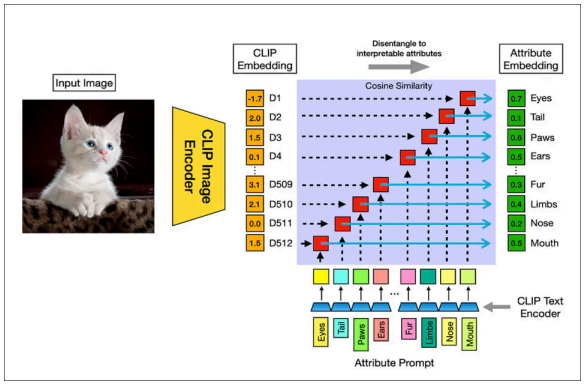
Since the methodology of the authors' tests are particularly arcane, and since minimal additional useful insight would be gained by examining them in depth, we will, untypically, only consider a shortened overview of their approach.

The work characterizes the core issue as *Class-Name Confound*: a situation where performance appears to come from meaningful descriptors, but in reality *depends on the class name itself*, with the descriptors adding little – and failing as soon as that support is removed

The paper then argues that this failure is inevitable, because descriptors generated from a class name can only describe what a thing *usually* looks like, rather than what is actually present in a particular image. As soon as the data departs from those expectations, the descriptors become not merely unhelpful but actively misleading, effectively *voting against the correct answer*.

The researchers also considered whether CLIP might simply be poor at detecting attributes without the aid of class labels, or whether GPT-generated descriptors were too generic to be useful. However, their subsequent experiments would refute both explanations.

To address this issue, a frozen CLIP model was used to compare images against a large pool of candidate descriptors drawn from VAW, LSA and GPT-3 outputs, retaining only those attributes that best matched the images, without requiring any retraining or additional image-text supervision:



An overview of the proposed system: a frozen CLIP model encodes both images and a large pool of candidate attribute texts, using cosine similarity to measure how strongly each descriptor matches the visual content. The highest-scoring attributes are then retained on a per-class basis, producing an interpretable set of prompts while keeping both the image and text encoders fixed throughout the process.

Data and Tests

The authors' experiments used CLIP with a ResNet-50 backbone and evaluated performance on ImageNet together with four distribution-shifted variants: ImageNetV2; ImageNet-Sketch; ImageNet-A; and ImageNet-R.

Attributes were selected using only the original ImageNet training images – allowing the shifted datasets to serve as an out-of-distribution test of whether image-derived attributes remained useful when visual appearances changed:

Prompt	ImageNet	-V2	-Sketch	-A	-R
Class name in the prompt					
Class name only (zero-shot CLIP)	55.31	49.39	31.58	20.92	58.10
+ LLM descriptors [13]	59.47	52.84	33.75	23.51	57.32
+ Ours (top 5)	59.53	53.33	33.12	22.79	56.30
+ Ours (top 10)	60.18	53.40	33.66	23.43	57.49
+ Ours (top 100)	**60.80**	**53.95**	**34.27**	**24.00**	**59.04**
Attribute only (no class name)					
LLM descriptors [13]	15.50	14.00	9.60	8.57	17.91
Ours, same pool as [13] (top 5)	19.40	17.00	10.57	9.80	19.49
Ours, VAW+LSA pool [15, 16] (top 5)	**23.80**	**20.80**	**14.40**	**11.83**	**28.20**

Top-1 classification accuracy across ImageNet and four distribution-shifted benchmarks. The results show that performance remains broadly similar when class names are included in prompts, but collapses when descriptors are forced to stand on their own, suggesting that much of their apparent effectiveness derives from the class label itself. By contrast, attributes selected directly from images remain substantially more informative across all datasets tested.

Re-selecting attributes from the same GPT-generated pool, but conditioning them on ImageNet images, raised class-name-free accuracy from 15.5% to 19.4%. Because the model, attribute pool and evaluation protocol remained unchanged, the gain could be attributed entirely to image-conditioned selection.

The result also indicated that CLIP can identify attributes without class labels, and that the GPT-generated pool already contained useful descriptors. The main failure instead appeared to lie in *label-conditioned generation*, which often failed to reveal the attributes most relevant to the images themselves.

Though the authors continue on a round of extensive additional experiments, we must refer the reader to the source material for further details of these, since, rather than expanding the scope of the outcomes, they only confirm the central hypotheses outlined so far – that reliance on labels can potentially undermine the potential of real image data in modern computer vision applications, with cheap reliance on 'reputational' probability a hazard to accuracy of results.

Conclusion

It's interesting to consider the debt that modern generative AI owes to the millions of hand-described images that were ingested into huge trawled datasets such as Common Crawl, back at the start of the hyperscale era.

Every time an image recognition or auto-labeling system attempts to classify or re-label an old or new image, the provenance of that knowledge harks back to some vendor writing *'Hot 1970s magazine!'* in the *alt* label of some eBay listing in 2004, or some such similar event.

Though many believe the golden age of keyword-stuffing was swept away by Google's more sophisticated PageRank algorithm nearly three decades ago, the *'wall of text, let's hope something sticks'* approach remains very much alive: only yesterday, heading the HTML of the release page for the Qwen-Image-3.0 model, was an atavistic blast (it may still be there!) of keyword-stuffing madness:

The HTML code for the release of Qwen Image 3 is a not-entirely-SFW wall of keyword text, at least, at the time of writing. Source

Whether or not, as is likely, this avalanche of often NSFW keywords in the Qwen Image 3 page is drawn systematically from target-lists, or hand-written by SEO specialists, this is a prime example of the raw origins of the modern generative AI system, with meaning often carrying a huge payload of self-interest that will later need to be stripped out or at least accounted for in some way, when such terms interfere with or hinder rational systems and applications.

First published Wednesday, July 22, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More