Hallucinations and Fairness in AI Models: A Survey of Recent Developments

Zhengyang Zhu

University of Rochester

Abstract

Recent advances in vision and vision-language models have unlocked impressive zero-shot and generative capabilities, yet they also expose two critical reliability challenges: hallucination — the production of content not grounded in the input, and fairness — systematic performance disparities across demographic or contextual subgroups. Although each phenomenon has been studied in isolation, their underlying causes often intertwine: spurious correlations in training data can simultaneously induce ungrounded outputs and biased predictions. This survey provides the first integrative review of hallucination and fairness in modern AI systems. We formalize both concepts, summarize evaluation metrics, and propose a unified taxonomy spanning three dimensions: *task* (hallucination detection/mitigation, bias discovery, subgroup fairness, adversarial robustness), *model* (foundation vision-language models, standard image classifiers, adversarially trained frameworks), and *dataset* (COCO, Visual Genome, CIFAR-100, BREEDS, ImageNet variants). Drawing on more than fifty recent papers—including empirical studies such as Can CLIP Count Stars?, DIM, DBD, Bag-of-Tricks, and L2T—we analyze prevailing mitigation strategies, highlight trade-offs between factuality and equity, and identify open challenges such as incomplete ground truth and metric misalignment. We conclude by outlining future directions toward unified benchmarks and reliability-aware training protocols, aiming to foster AI systems that are both truthful and equitable in real-world deployment.

1 Introduction

Modern AI systems, especially large vision and vision-language models, have achieved unprecedented capabilities but also prompted serious concerns about their reliability and ethics. Recent multimodal foundation models – exemplified by CLIP (Radford et al. 2021[1]) – have demonstrated state-of-the-art performance on image recognition and retrieval tasks, enabling rich image descriptions and zero-shot predictions. However, alongside these successes, two phenomena have emerged as critical challenges in real-world deployment: **hallucinations** and **fairness**. Hallucinations refer to ungrounded or spurious outputs that mis-align with the input (e.g. describing objects not actually present in an image), while fairness concerns relate to biased performance across different subgroups, leading to disparate outcomes or treatment. Both issues have gained prominence as AI is increasingly used in high-stakes and diverse settings, where generating *factually correct* and *equitable* results is as important as raw accuracy.

Prior surveys and reviews have largely treated hallucination and fairness as separate topics, or focused on specific model classes. On one hand, the phenomenon of hallucination in language and vision-language models has been surveyed in depth – for example, Liu *et al.* (2024)[2]provided a comprehensive overview of LVLM hallucinations, characterizing their symptoms and causes and reviewing mitigation techniques. On the other hand, the literature on fairness in machine learning

Preprint. Under review.

is extensive, with surveys like Mehrabi *et al.* (2021) [**?**]cataloguing sources of bias and fairness definitions and examining unfair outcomes in various ML subdomains. However, these works each focus on a single dimension (either hallucination or fairness) or on a narrow set of scenarios. Notably, there has been a gap in understanding how these issues might intersect in modern AI systems, and how advances in one area (e.g. robust training) could inform the other. This survey aims to fill that gap by providing an integrative view that spans hallucinations, fairness, adversarial robustness, and subgroup performance in vision and vision-language models. By examining hallucination and fairness side by side, this survey highlights cross-cutting insights.

2 2. Preliminary

This section aims to provide the necessary background and relevant knowledge and definitions to understand the topic of hallucination and fairness of AI.

2.1 2.1 Hallucinations in Vision-Language Models

Hallucination in vision-language models are referred to as the generation of content that is not grounded in the input. In image captioning and related tasks, hallucination causes the models to produce descriptions with objects or details that do not actually appear in the image. [3]. For example, errors in caption might lead to false information, such as mentioning a "*TV in the background*" when there is no TV present. Such examples of *object hallucinations* pose problems, especially when factual correctness is crucial. (e.g., applications that assist visually impaired patients) [3] Hallucinations could greatly undermine user trust and the reliability of these models. [3]

As already mentioned above, *object hallucination*, where nonexistent objects are named in a description, is the most studied form in the issues of hallucinations in vision-language models. [3] Models may also hallucinate attributes or actions (e.g., describing a shirt as *blue* when it is actually red) Recent work highlighted and described *multi-object hallucination* as a challenge: when asked to describe multiple objects, large vision-language models (LVLMs) often invent extra objects or get “distracted,” especially if the scene is complex. [4] The major underlying causes are related to biases in the data set and the limitations of the model. The primary cause is *co-occurrence bias* in training data – models learn spurious correlations between objects that frequently appear together. [5]biten2022letFor instance, if training captions often mention *"umbrella"* whenever *"rain"* is present, a model might hallucinate an umbrella in any rainy scene. This systematic bias in co-occurrence statistics can lead models to hallucinate objects unless explicitly corrected. [5] More generally, when models rely too heavily on prior knowledge or linguistic priors instead of image evidence, hallucinations occur (analogous to *suspicious correlations* noted in robust modeling. [6]

Early studies by Rohrbach et al. (2018)[7] drew attention to object hallucination in captioning models, analyzing how model architectures and training objectives contributed to the issue. They introduced the ***CHAIR*** metric (Caption Hallucination Assessment with Image Relevance) to quantify how many generated words refer to objects absent from the image. This work showed that despite the high overall captioning performance, models often *’hallucinate’* objects, indicating a gap between the maximization of BLEU scores and the accuracy of the factual.[5] Subsequent research tried to mitigate hallucinations via improved attention mechanisms, adversarial training, and better visual grounding. Notably, some studies hypothesized that *more detailed captions* inherently cause more hallucinations (since a model trying to say more might guess extra content). [8] However, Feng *et al.* (2024)[8] challenge this assumption; they identified flaws in the way hallucinations were evaluated, which had led to potentially incorrect conclusions. By proposing more reliable metrics (discussed below), they showed that it is possible to generate very detailed descriptions while keeping hallucinations low. This view suggests that the prevalence of hallucinations is not just about *how much* detail a model includes, but *how* that detail is generated and measured.

2.2 2.2 Fairness in AI and Image Classification

Fairness in AI systems refers to the principle that models should make decisions or predictions without unwarranted bias toward or against particular groups. In practice, this often means ensuring consistent performance across different subsets of data (e.g. demographic groups or subcategories) and avoiding the propagation of harmful stereotypes. In computer vision, fairness issues have been vividly demonstrated in tasks like face recognition: Buolamwini and Gebru (2018)[9] showed that commercial gender classification systems had drastically lower accuracy for darker-skinned female faces than for lighter-skinned male faces. In their study, the worst-group error rate was 22.4% – nearly 7× higher than the error rate on the best-performing group. [9]Such *accuracy disparities* illustrate how seemingly high overall accuracy can mask severe inequities, underscoring the need for fairness audits in vision models.

In image classification, a key fairness concern is that models may perform unevenly across *subgroups* of data. A classifier might achieve good average accuracy but still severely under-perform on certain categories or demographic subsets. This phenomenon is often termed subgroup bias or *subpopulation shift*. For example, Zhang *et al.* (2024) found that an ImageNet-trained ResNet-18 achieved 54.6% accuracy on the broad *"aquatic mammals"* category of CIFAR-100, yet within that category its accuracy ranged from 72% on "otter" down to only 34% on "beaver".[10] This large difference indicated that the model learned features that favor some sub-classes and disadvantage others, revealing a hidden bias. Such subgroup biases can arise from imbalanced training data or spurious correlations: if certain species or object subgroups are under-represented or consistently associated with specific contexts, the model's predictions will be skewed.[10] In the example above, "beaver" images might differ in background or frequency, leading the classifier to systematically misidentify them – a fairness issue since the model is less reliable for that subgroup. Researchers have started addressing *multiple biased subgroups* by identifying latent factors in the model's feature space that correspond to these performance gaps and then mitigating them. This ensures that previously *unknown* biases (those not labeled in the data) can be discovered and corrected, moving beyond fairness only on pre-defined attributes.

Another emerging fairness concern in vision models is **quantity bias**. This term, introduced by Zhang *et al.* (2024) in the context of CLIP, refers to systematic errors in handling numeric quantities. Their study *"Can CLIP Count Stars?"* showed that CLIP's vision-language embeddings are biased in how they represent quantity, often causing downstream applications to consistently misestimate counts.[6] For instance, when using a CLIP-based image generation (Stable Diffusion) to draw "five pandas", the model repeatedly produced seven pandas. This suggests the model's joint image-text representation encoded a bias toward higher counts (over-counting by default). [6]Such a bias is a fairness issue in the sense of *functional bias*: certain inputs (prompts asking for a specific number) yield systematically distorted outputs. It stems from the training data – CLIP was trained on image-text pairs where numeric phrases may not be evenly distributed, making it less "trustworthy" for exact counts. More broadly, *spurious correlations* and *uneven data distributions* (as noted by Sagawa *et al.*, 2020 and others) can lead to unexpected biases in vision models' behavior. Ensuring fairness thus requires not only attention to protected attributes (like race or gender), but also to any latent factors (like quantity, background context, etc.) that cause performance to degrade for certain subsets. In summary, fairness in image classification means striving for equitable performance and treatment across all subgroups of data – whether those are defined by demographics, object subcategories, or other attributes – and guarding against systematic errors that impact one group more than others.

3 3. Taxonomy

This section presents a comprehensive taxonomy of recent research addressing hallucination and fairness in AI models. Building on the preliminaries, we now categorize the research landscape of hallucinations and fairness in AI models along three complementary dimensions: **task**, **model**, and **dataset**. Each perspective provides a different organizing axis for survey works and methods in this field. Table 1 summarizes the taxonomy, and the following subsections detail each category, highlighting major trends, notable papers, and persistent challenges.

- *Task Perspective*: group studies by the problem tasks they target – including detecting or mitigating hallucinations, discovering biases, ensuring subgroup fairness, and improving adversarial robustness.
- *Model Perspective*: categorize approaches by model type, distinguishing between large foundation vision-language models (e.g. CLIP, LVLMs), standard image classifiers (e.g. ResNet), and specialized adversarial training frameworks.
- *Dataset Perspective*: classify research by the benchmark datasets used for evaluation, such as *MS-COCO*, *Visual Genome*, *CIFAR-100*, *BREEDS*, and various *ImageNet* derivatives

3.1 3.1 Task Perspective

3.1.1 Hallucination Detection and Mitigation

Detection of hallucinations requires metrics and benchmarks to identify such false content, while mitigation involves methods to prevent or correct them. As already mentioned in the preliminary section, Early work by Rohrbach *et al.*[7] demonstrated that image captioning models often hallucinate objects not in the scene, and standard captioning metrics failed to capture this problem. For this issue, they proposed an image relevance metric ground-truth object labels to quantify "object hallucination" rates. Recent large vision-language models (LVLMs) like MiniGPT-4, LLaVA, and others have highlighted the issue at scale: despite high overall performance, these models sometimes describe details that are absent, due to biases in their training or decoding process. [2] Common hallucination types include *object hallucinations*, *attribute errors*, and *modal conflicts*, as categorized by Liu *et al.* (2024). [2]

To detect hallucinations, researchers have devised automated checks often using cross-modal agreement. One line of work uses a secondary model (e.g. CLIP or a visual question answering model) to verify whether the generated description is consistent with the image.[11][12]For instance, *visual verification* steps can compare image–text alignment post-hoc, flagging content that the vision encoder cannot find evidence for. [8]Another approach is to design benchmarks with *counterfactual prompts*: Park *et al.* (2024) found that adversarially generated false dialogue to an image query exacerbates hallucinations, helping measure model brittleness as well.[13]

On the mitigation side, several strategies have emerged. *Decoding-time methods* can reduce ungrounded content without retraining. Reinforced grounding techniques bias the generation process toward the visual input: Favero *et al.* (2024)[**?**] propose increasing the mutual information between the image and generated text (their M3ID decoding), which significantly lowers hallucination rates in image descriptions. Moreover, the *differentiated decoding (DBD)*, developed by Chang *et al.* (2024) [8]generates multiple partial descriptions in parallel and selects a consensus, coupled with new CLIP-based precision/recall metrics to ensure added details are image-accurate.

Another mitigation techniques are *training-time interventions*. These include adversarial training and data augmentation to reduce over-reliance on language priors, which Biten *et al.* (2022) [5]showed that augmenting caption training data to disrupt spurious word-object co-occurrences can lessen object hallucinations. In his study, by reducing biases (e.g. tuning "surfboard" always on a "beach"), the model is less tempted to hallucinate frequent, but absent objects.

However, current challenges in this field still exists, and it has to do with how the hallucination metrics are, yet, to reach perfect level. In recent researches, the models still may falsely penalize *novel but correct details* due to incomplete ground truth,[8] and mitigation often must be balanced with preserving fluency and coverage. As Liu *et al.* (2024)[2] noted, no single method eliminates all root causes of LVLM hallucinations, indicating open problems in aligning these models to reality.

3.1.2 Uncover Biases

This task focuses on identifying latent biases or spurious correlations that cause models to perform inconsistently. Unlike hallucinations (an overt output error), bias issues often lurk in the model's decision-making process and may not be immediately visible without targeted analysis. A classic example is a classifier that learns a shortcut (like associating a background or an attribute with a class label) – it may achieve high overall accuracy but fail in rare cases that break the shortcut. Defining and discovering such biases is crucial for the reliability of the AI models, because otherwise there exists unknown or unlabeled biased subgroups in the data, which affect the performance of the model output.

Several methodologies have been proposed for bias discovery. One of them is the *Domino* approach (Cross-model embedding analysis) introduced by Eyuboglu *et al.* (2022), which uses image–text embeddings to cluster the dataset and find where the model's error rate is abnormally high. [14] This helps pinpoint, for example, a subset of images (clustered by a certain attribute combination) that the classifier struggles with. Another strategy is to utilize explainable feature attribution: Singla and Feizi (2022) [15]generate *Salient ImageNet* by annotating which parts of an image are "core" vs. "spurious" features for the label. By visualizing model attention or gradient maps, they identify when a model attends to spurious cues (like background textures) instead of core objects, thereby revealing bias in

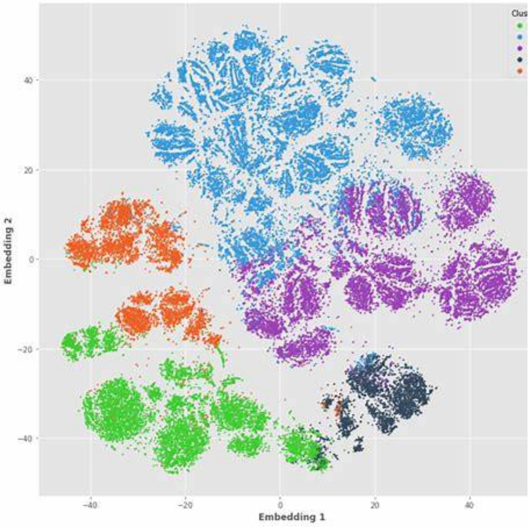


Figure 1: Domino slice-discovery visualization. Two-dimensional t-SNE projection of CLIP image–text embeddings for 10k validation images. Domino automatically clusters the embeddings into semantic “slices”; each point is colored by its assigned slice and sized by the slice’s error-rate (larger, warmer points=higher misclassification rate for that cluster). Dense red clusters (e.g., snow-covered scenes, low-light interiors) reveal previously hidden sub-populations where the classifier fails disproportionately, while blue clusters indicate slices on which the model performs well.

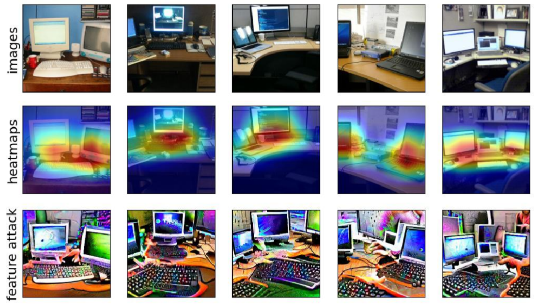
feature use. Generative modeling has also been leveraged for bias discovery. Some works train a generative model or GAN conditioned on the classifier’s outputs to see what *implicit features* the classifier has learned. For instance, Li and Xu (2022) and Lang *et al.* (2023) each propose methods to generate synthetic images that represent potential biased subgroups.[10] [16][17] If the classifier performs poorly on these generated samples (or if they highlight a coherent concept), it suggests an unknown bias. Jain *et al.* (2022) take a simpler approach: they analyze the *linear classifier weights* of an image model to extract a direction corresponding to a failure mode, then use CLIP to automatically caption images from that mode – providing a human-interpretable description of the biases (e.g., although "watermarks" or "dark backgrounds" could interfere with the results). These techniques underscore a common trend, which is: tools like CLIP ad GPT-4 are increasingly used to characterize model mistakes in natural language, facilitating bias discovery.

Yet, one prominent challenge in bias uncovering is scaling to multiple concurrent biases. Early works often assumed a single dominant bias per dataset or model. However, real-world models can fail due to many factors at once. Addressing this, Zhang *et al.* (2024) [10]proposed the DIM framework (Decomposition, Interpretation, Mitigation) which explicitly aims to discover multiple biased subgroups simultaneously. By decomposing the feature space of a trained classifier via Partial Least Squares, they isolate several candidate subgroup directions, interpret each with a text description (using a VLM), and then measure the model’s accuracy on each subgroup. Their experiments on CIFAR-100 and BREEDS confirm that models often harbor more than one bias (e.g., a CIFAR-100 animal class might have multiple overlooked subpopulations) and highlight the need for methods that can untangle this complexity.

3.1.3 Subgroup Fairness

3.1.4 Adversarial Robustness

Instead of inadvertent biases or hallucinations under normal conditions, robustness studies how models can be intentionally fooled by worst-case *adversarial perturbations*. However, robustness is deeply related to fairness and reliability – a non-robust model might be exploited in ways that cause disproportionate harm (e.g. if an attack selectively impacts a certain subgroup), and conversely, techniques to improve robustness can sometimes mitigate biases by forcing the model to focus on



Top: original ImageNet validation images correctly labeled by humans.
Middle: Grad-CAM saliency maps for a pretrained ResNet-50; the model's highest-attention regions (red) fall mainly on background elements such as keyboards and monitors rather than on the core object.
Bottom: human-annotated masks (green=core, magenta=spurious) from the Salient ImageNet benchmark. The heavy overlap between the model's saliency and the spurious mask reveals that the classifier relies on background shortcuts, exposing a feature-level bias that can cause misclassification when the shortcut is absent.

Top: original ImageNet validation images correctly labeled by humans.

Middle: Grad-CAM saliency maps for a pretrained ResNet-50; the model's highest-attention regions (red) fall mainly on background elements such as keyboards and monitors rather than on the core object.

Bottom: human-annotated masks (green=core, magenta=spurious) from the Salient ImageNet benchmark. The heavy overlap between the model's saliency and the spurious mask reveals that the classifier relies on background shortcuts, exposing a feature-level bias that can cause misclassification when the shortcut is absent.

Figure 2: **Salient ImageNet heat-map.**
Top: original ImageNet validation images correctly labeled by humans.
Middle: Grad-CAM saliency maps for a pretrained ResNet-50; the model's highest-attention regions (red) fall mainly on background elements such as keyboards and monitors rather than on the core object.
Bottom: human-annotated masks (green=core, magenta=spurious) from the Salient ImageNet benchmark. The heavy overlap between the model's saliency and the spurious mask reveals that the classifier relies on background shortcuts, exposing a feature-level bias that can cause misclassification when the shortcut is absent.

more essential features. This category includes both attack methods (to test robustness) and defenses (to make models more robust).

On the attack side, Goodfellow *et al.* (2015) [18]demonstrated that tiny perturbations to an image can cause a classifier to mislabel it with high confidence. Such *adversarial examples* revealed that despite high accuracy, models were overly sensitive to input details humans would consider insignificant. Subsequent work created a zoo of adversarial attacks (FGSM, PGD, DeepFool, etc.) and also extended to multi-modal scenarios – for instance, recent research showed that adding a carefully crafted false caption or question to an image (a *textual adversarial attack*) can induce hallucinations in LVLMs.[13]These efforts serve to stress-test AI models beyond normal operating conditions.

Defenses typically involve some form of adversarial training, which Madry *et al.* (2018) formulated as a robust optimization problem (minimax): train the model on adversarially perturbed inputs so that it learns to resist them.[19] This approach, while successful in substantially improving robustness to the models, often comes with a cost, that is with lower standard accuracy and increased training time. However, interestingly, there has been evidence that adversarially trained models rely on different features (e.g. more on shape than texture in images), which can align with more human-perceptual features and potentially reduce reliance on spurious cues. In other words, a robust model might inherently be less biased toward certain superficial correlations. For example, some studies found that robust models are less prone to dataset biases like background correlation, since an adversary could exploit those, the model is forced to focus on truly discriminative features. [20]Adversarial training has also been explicitly used to debias models: Zhang *et al.* (2018) introduced an *adversarial debiasing* framework where an auxiliary adversary network tries to predict the protected attribute from the model's representation, and the model is trained to fool this adversary. [21] This removes information about race or gender from the feature space, and encouraging fairer prediction outcomes.

A current challenge for adversarial robustness research is that, improving the robustness to one kind of perturbation may not generalize which to others. For instance, a model hardened against *ell-infinity*, denoting the max-norm, a method to measure the size of a perturbation in adversarial robustness research, noise might still be vulnerable to a patch attack or a more semantic change (like a different lighting condition). That's where it connects back to the dataset perspective: benchmarks like ImageNet-C, -A, and others, explicitly evaluate models under a wide range of shifts, from artificial perturbations to natural distribution shifts. The goal is often to train models that maintain performance across all these conditions, which is analogous to the goal of maintaining performance across all subgroups (fairness) – in both cases, the worst-case group or input dictates success. As a result, techniques like DRO, invariant risk minimization [10], and others straddle the line between robustness and fairness, providing a unifying view: a robust, fair model should not be overly sensitive to *who* or *what* is in the input, nor to *how* small changes are in the input.

The task perspective of this taxonomy highlights a spectrum from *hallucination detection* (identifying content errors) through *bias/fairness tasks* (identifying and correcting decision biases) to *robustness* (guarding against adversarial input). These tasks are interconnected – for example, reducing hallucinations in a captioning model may involve making it more robust to misleading language priors (a form of adversarial input), and achieving subgroup fairness may benefit from techniques developed for robustness (treating sensitive attributes as adversarial factors to neutralize).

3.2 3.2 Model Perspective

From the model-centric viewpoint, research can be grouped by the type of AI model being studied or utilized. Each class of models – foundation vision-language models, standard image classifiers, and adversarially-trained frameworks – comes with its own peculiarities in terms of hallucination and fairness issues.

3.2.1 Foundation vision-language models

A large portion of recent work focuses on multi-modal foundation models, which we define as models pretrained on massive image-text data capable of general vision-language understanding (e.g. CLIP, ALIGN, BLIP, Flamingo, and instruction-tuned LVLMs like LLaVA, mPLUG-Owl, MiniGPT-4). These models are powerful but are known to exhibit both hallucinations and biases inherited from their training corpora.[2]For example, OpenAI's CLIP (Radford *et al.*, 2021[1]) was shown to have impressive zero-shot image recognition capability by aligning image embeddings with text embeddings; yet it also reflected biases in labeling (such as occasional misclassifications along stereotypical lines) and could be vulnerable to adversarial prompts (nonsense text strings that alter its predictions). Large generative VLMs (e.g. Flamingo, BLIP-2) and recent LVLMs (which combine a frozen LLM with a vision encoder, as in LLaVA[2] similarly confront the challenge that scale alone doesn't eliminate hallucinations or bias – it sometimes even magnifies them because the models are so fluent.

Studies targeting this category often aim to audit and align foundation models. For hallucinations: as discussed, several mitigation methods like Favero et al.'s M3ID decoding [**?**]and Park et al.'s adversarial instruction tuning [13]are designed specifically for LVLMs (e.g. LLaVA or MiniGPT) to produce more truthful, grounded outputs. These foundation models also serve as *tools* within other methods – CLIP in particular is ubiquitous for evaluating image-text consistency (e.g. CLIP-score used as a hallucination metric, and for explaining classifier biases via automatic captioning. On fairness: foundation models like CLIP have been evaluated for bias, with mixed findings. Some work (e.g. DataComp analysis) suggests CLIP's training data filtering can inadvertently *exclude certain demographics*, leading to representation disparity.[22] thers have developed *debiasing techniques for VLMs* – for instance, a recent NeurIPS 2023 paper proposed a unified debiasing approach for vision-language models across tasks. [23] The challenge in this model class is their complexity: their very high capacity and multi-modal nature mean they can memorize or correlate many unwanted patterns, so aligning them with human values (truthfulness, fairness) is an ongoing effort. Surveys such as Liu *et al.* (2024) [2]provide taxonomies of hallucination types in LVLMs and call out open questions in better integrating visual grounding into these models

3.2.2 Standard Image Classifiers

By contrast, a lot of fairness and bias research has centered on simpler models like *ResNet-50* (He et al., 2016[24]) or *Vision Transformers* (Dosovitskiy et al., 2021[25]) trained on ImageNet or similar datasets. These models are often the subject of subgroup performance analyses because their behavior is easier to interpret than a giant LVLM, and biases can be studied in a controlled setting. For example, many bias discovery works (including above-mentioned: Domino, Salient ImageNet, etc.) used a standard ImageNet-trained classifier as the target to diagnose. Similarly, methods like Nam *et al.* (2020) "Learning from Failure" explicitly trained a ResNet on a biased dataset and then used its errors to guide a de-biased model.

Image classifiers also remain the workhorse for evaluating robustness. Nearly all adversarial attack and defense papers in vision benchmark their methods on CIFAR-10/CIFAR-100 or ImageNet classifier models.[20] The BREEDS benchmark, for instance, takes an ImageNet-100 subset and defines coarse vs. fine labels to create subpopulation shifts, then measures how a ResNet's accuracy drops when evaluated on a shifted distribution. In Zhang *et al.* (2024)'s [10]DIM study, a plain ResNet trained on CIFAR-100 was used to demonstrate hidden subgroup biases. Such models provide a controlled testbed to validate bias mitigation strategies: if a technique cannot debias a ResNet on a known toy problem, it likely won't scale to more complex models. A major trend here is the use of alternative training methods on these classifiers: e.g. *group DRO* training that treats each class or subgroup as an "environment" to equalize, or *IRM (Invariant Risk Minimization)* which attempts to make the classifier focus on features invariant across environments. These were initially developed on simple classifiers before adaptation to larger models. The ongoing challenge for standard classifiers is that even though they are simpler, they still exhibit surprisingly complex failure modes, and techniques that work in theory (like certain invariant learning objectives) sometimes fall short in practice. Nonetheless, advances made on this model class – for example, calibration techniques to reduce overconfidence on unfamiliar inputs, or loss functions that penalize within-class feature diversity (to avoid relying on one cue) – form a foundation that is later transferred to foundation models.

3.2.3 Adversarial Training Frameworks and Robost Models

In this category, the models comprise the explicitly designed or trained for robustness. which often become a separate class due to their distinct behavior. Adversarially trained models (like the widely studied "*Madry model*" – a ResNet trained with PGD adversaries,[19] have been shown to have different feature representations and failure characteristics than standard models. For instance, adversarially trained classifiers are more immune to certain common corruptions and may exhibit less sensitivity to some spurious correlations, but they can also be less accurate on clean data or exhibit new types of biases (e.g. they might under-perform on easy samples because they focus on worst-case). [20]

Recent works at the intersection of robustness and fairness propose unified views. One perspective is treating **any error disparity as an adversarial exploit** – if a model has a higher error on subgroup A than B, one could conceive of an "adversary" that simply selects inputs from A to maximize error; a truly robust (in worst-case sense) model would then have to equalize performance. This insight connects to the use of DRO for fairness as mentioned, and to evaluation sets like ImageNet-A (natural adversarial images) which often disproportionately feature unusual depictions of objects that standard models biased by texture fail on.

The *Adversarial Instruction Tuning* method by Park *et al.* (2024)[13]is a concrete example in the multi-modal space: they pose hallucination prompts as adversarial attacks and fine-tune the LVLM to resist being led astray. This blurs the line between robust modeling and bias mitigation – the model becomes robust against a specific bias (in their case, the bias toward prior dialogue context over the visual content).

In summary, viewing research by model type, there has been a pipeline of ideas flowing from simple classifiers to large foundation models. Standard models are used to develop and benchmark core methods (for bias detection, fairness algorithms, and adversarial defenses). Foundation models are where those methods must scale and often where new problems are discovered (like novel forms of multi-modal hallucination). Adversarially trained models and frameworks represent a specialized branch that influences both other categories by offering insights into model behavior under stress. Ensuring that advances translate across these model categories is an ongoing challenge – for instance, a debiasing strategy that works for ResNet + ImageNet might not directly work for CLIP or BLIP without modification, and conversely, techniques for hallucination reduction in LVLMs might inspire new training regimes for simpler models as well.

3.3 3.3. Dataset Perspective

The choice of evaluation data has a profound impact on diagnosing and comparing approaches for hallucinations and fairness. Here, I outlined some key datasets and benchmarks that have become standard in this research area, organized by the type of problem they emphasize.

3.3.1 COCO (MS-COCO) and Visual Genome – Captioning and VQA Benchmarks

The MS-COCO dataset (Common Objects in Context) has been a cornerstone for image captioning and visual question answering (VQA) tasks. With its 123k images and 5 human-written captions per image, COCO is where the problem of *object hallucination* in captions was first systematically observed. [7] Models trained on COCO sometimes output objects that are plausible in context but not actually present (e.g. hallucinating a “clock on the wall” in a kitchen scene that has none).[5] This is partly because COCO captions are not exhaustive descriptions – they often omit small or obvious objects, so a model that “helpfully” mentions them can be penalized as hallucinating. Recognizing this, researchers created COCO-like benchmarks with extra annotations to catch hallucinations. For example, Rohrbach *et al.* [7]added object presence labels to COCO images and computed metrics like Precision@K to see how many predicted objects were truly in the image. COCO remains a primary dataset for evaluating caption fidelity, and many hallucination mitigation papers report improvements on COCO Caption metrics or the newer ***CaptionHallucination Benchmark*** (an unofficial term for evaluating object presence in captions on COCO).

Visual Genome (VG) is another influential dataset, containing 108k images with dense region descriptions, object annotations, and question-answer pairs. VG’s richness makes it suitable for hallucination evaluation because it provides a more complete set of “ground truth” objects per image. Liu *et al.* (2023) [2]in their LVLM hallucination survey specifically noted that using Visual Genome to compute their proposed CLIP-Precision and CLIP-Recall metrics, as VG has annotations that allow partitioning the image and caption. In the DBD work, experiments on Visual Genome showed that more detailed captions (with guidance) can be generated without increasing hallucinations, validated by checking against VG’s exhaustive labels. VG is also used in VQA, where one can test if a model’s answer contains information not supported by the image. The **VQA v2** dataset attempted to minimize language biases in questions, but models still exploit priors (e.g. always answering “yes” for “Is there…?” questions). Thus, VQA evaluations often consider whether an incorrect answer was because of a hallucinated visual detail or bias. Newer benchmarks like **VQA-CP** (which re-shuffle question-answer pairs to break correlations) force models to rely on the image, indirectly measuring hallucination-like behavior (answering from priors vs. visual evidence). In short, COCO and VG serve as *standard touchstones*: a good hallucination mitigation method should reduce false details on COCO/VG, and a fairness method for captioning should not degrade performance on these gold-standard datasets.

3.3.2 CIFAR-100 and CIFAR variants

CIFAR-100 is a small image classification dataset (32×32 images in 100 classes) that has proven useful for bias and fairness research due to its manageable size and the semantic hierarchy of its classes. Each CIFAR-100 image has a “coarse” label (superclass) and a “fine” label (subclass). This naturally lends itself to studying *within-class biases* – for example, one superclass is “vehicles” and subclasses include “bicycle”, “motorcycle”, etc. If a model learns to recognize the superclass but confuses subclasses, one can simulate a bias (e.g. only focusing on wheels for all vehicles). Zhang *et al.* (2024)[10][26] leverage CIFAR-100 in a controlled setting where certain subclasses are underrepresented to test their DIM bias discovery method. They showed that DIM could recover the known biased subgroups (the minority subclasses) and improve accuracy on them.

CIFAR is also popular for adversarial robustness evaluation – many defense papers report robust accuracy on CIFAR-10/100 against attacks. While not a fairness dataset per se, the simplicity of CIFAR allows researchers to try out new fairness-through-robustness ideas quickly. For instance, some early experiments on *adversarial noise as augmentation* were done on CIFAR to see if models become both more robust and less biased to textures (with mixed results). Additionally, extensions like **CIFAR-10S** (with sensitive attributes) or **UTK-CIFAR** (overlaying UTKFace demographic attributes on CIFAR images) have been created to benchmark fairness algorithms, since CIFAR's content is generic and one can attach synthetic "attributes" to induce biases.

3.3.3 BREEDS and Subpopulation Shift Benchmarks

BREEDS is a collection of benchmark tasks introduced by Santurkar *et al.* (2021)[20] to explicitly test models under subpopulation shifts. Derived from ImageNet, BREEDS defines hierarchies (e.g. a superclass "dog" with subclasses "poodle", "collie", etc.) and then constructs train and test sets such that the model at test time encounters new subpopulations of a class it didn't see during training. For example, in one BREEDS scenario, a classifier might train on "dog" images that are only wild canines (wolves, foxes) and then be tested on domesticated dogs – requiring generalization beyond the seen subpopulation. This setup directly evaluates robustness to distribution shift and can reveal biases: if the model implicitly assumed "dog" = "wolf-like", it will fail. BREEDS benchmarks (Entity-13, Entity-30, Living-17, Nonliving-26) cover various ImageNet categories.[20].

In fairness research, BREEDS has been used to simulate *hidden stratification*. If we treat each subpopulation in BREEDS as an "attribute" (though not a human demographic attribute, it's analogous), we can assess how different methods improve worst-subgroup accuracy. Zhang *et al.* (2024) validate their bias mitigation on BREEDS, confirming that addressing multiple unknown biases helped on the "hard" ImageNet variant they construct. Another related dataset is **ImageNet-V2** (Recht et al. 2019[**?**]), a re-collection of ImageNet test set, which although collected with the same protocol, yields slightly lower accuracy for models – indicating a mild distribution shift. Techniques that improve subpopulation robustness often also improve performance on ImageNet-V2 and other variants, so these are monitored too.

3.3.4 ImageNet Variants (ImageNet-A, -O, -R, -C, etc.)

ImageNet-A (adversarial, curated set of naturally hard images that cause mistakes), **ImageNet-O** (out-of-distribution negatives), **ImageNet-R** (renditions – art, cartoons of objects), and **ImageNet-C** (common corruptions – blurred, noised images), these are suite of stress tests for ImageNet classifiers has emerged in recent research. [20]While these are not explicitly about fairness or hallucination, they are crucial for evaluating whether improvements are *general*. For instance, a model de-biased on ImageNet might still do poorly on ImageNet-R if it relies on texture (indicating an unresolved bias toward photographic texture). Adversarially trained models, interestingly, excel on ImageNet-A (suggesting they learned more robust features). Researchers aiming for broadly reliable models will test on these variants to ensure no regressions. In our context, a fairness-enhanced model that also performs well on ImageNet-C or -R can claim it didn't trade off robustness for fairness (since sometimes re-weighting data for fairness can make model brittle, it's important to check).

Moreover, some ImageNet variants have been crafted to expose specific biases: **Salient ImageNet** as mentioned provides masks of spurious features, [10]and **ObjectNet** is another external test set where objects are shown in odd orientations or backgrounds to break context biases. If a model hallucinates or mislabels objects due to context, ObjectNet will reveal such mistakes (e.g. a chair upside-down might be unrecognized because the model expects chairs upright). These datasets are used less frequently in classical fairness papers, but they are gaining attention in the robustness-fairness intersection literature.

In conclusion, the dataset perspective emphasizes that progress in this field is often driven by better benchmarks. By challenging models with new test sets (be it a captioning dataset that penalizes hallucinations, or a classification benchmark that demands fairness across groups or robustness across shifts), researchers reveal weaknesses that were previously hidden by overly narrow evaluation. The community has responded with an abundance of datasets: each designed to tax a particular aspect of model performance. A survey of 50+ papers in this space consistently shows that top-performing approaches are validated on a *combination* of these benchmarks – for example, a new hallucination mitigation method might report improvements in COCO Captioning (overall quality) *and* on a targeted metric like object hallucination rate on NoCaps or VG, while a new fairness algorithm may show that it closes the gap between subgroup accuracies on CelebA and also holds up on ImageNet-R (no trade-off in robustness). By structuring the research landscape along tasks, models, and datasets, we can better understand how a given contribution fits in: e.g., is it introducing a new evaluation (dataset perspective), or a new algorithm (task perspective) tested on existing benchmarks, or perhaps a new insight about a particular model type (model perspective)? The next sections of this survey will delve deeper into each line of work, following the taxonomy outlined, and discuss cross-cutting trends and future directions.

4 Discussion

Mitigating hallucinations entails practical trade-offs that can impact fairness. For instance, methods that aggressively avoid ungrounded content may over-regularize model outputs and inadvertently under-represent rare or minority subgroup features. This challenge is compounded by several methodological limitations. First, incomplete ground-truth annotations make it difficult to distinguish hallucinations from legitimate but unannotated details, leading to ambiguity in evaluation. Second, biases in AI systems are often interdependent, so addressing one form of bias can inadvertently exacerbate others. Finally, current evaluation metrics for factuality and fairness are frequently misaligned, complicating attempts to optimize both objectives simultaneously. These factors highlight open challenges in developing models that remain both truthful and equitable under diverse conditions.

5 Conclusion

This survey presented a unified taxonomy and overview of research on hallucinations and fairness in modern AI systems, with a particular focus on vision-language models. It underscored the intertwined nature of ungrounded content generation and subgroup biases, emphasizing their joint importance for trustworthy AI. Looking ahead, a key priority is the development of unified benchmarks and evaluation metrics that jointly assess both hallucination and fairness, enabling more holistic model evaluation. It is also crucial to deepen our understanding of how multi-modal foundation models internalize and potentially amplify biases from data, as these biases underpin both untruthful outputs and disparate performance. Ultimately, we advocate for holistic, reliability-aware training and evaluation protocols to ensure that next-generation AI systems are both truthful and equitable across diverse real-world conditions.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[2] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[3] Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In Minsu Cho, Ivan Laptev, Du Tran, Angela Yao, and Hongbin Zha, editors, *Computer Vision – ACCV 2024*, pages 166–182, Singapore, 2025. Springer Nature Singapore.

[4] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Jianing Yang, David F. Fouhey, Joyce Chai, and Shengyi Qian. Multi-object hallucination in vision language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 44393–44418. Curran Associates, Inc., 2024.

[5] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390, 2022.

[6] Zeliang Zhang, Zhuo Liu, Mingqian Feng, and Chenliang Xu. Can clip count stars? an empirical study on quantity bias in clip. *arXiv preprint arXiv:2409.15035*, 2024.

[7] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019.

[8] Mingqian Feng, Yunlong Tang, Zeliang Zhang, and Chenliang Xu. Do more details always introduce more hallucinations in lvlm-based image captioning? *arXiv preprint arXiv:2406.12663*, 2024.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[10] Zeliang Zhang, Mingqian Feng, Zhiheng Li, and Chenliang Xu. Discover and mitigate multiple biased subgroups in image classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10906–10915, 2024.

[11] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. pages 14303–14312, 06 2024.

[12] Yue Chang, Liqiang Jing, Xiaopeng Zhang, and Yue Zhang. A unified hallucination mitigation framework for large vision-language models. *arXiv preprint arXiv:2409.16494*, 2024.

[13] Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. Mitigating dialogue hallucination for large vision language models via adversarial instruction tuning, 2024.

[14] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. *arXiv preprint arXiv:2203.14960*, 2022.

[15] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

[16] Zhiheng Li and Chenliang Xu. Discover the unknown biased attribute of an image classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14970–14979, 2021.

[17] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[19] Xiaosen Wang, Bhavya Kailkhura, Krishnaram Kenthapadi, and Bo Li. I-pgd-at: Efficient adversarial training via imitating iterative pgd attack. 2021.

[20] Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*, 2020.

[21] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[22] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17, 2024.

[23] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2024.

[24] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019.

[25] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.