# Semi-Truths: A Large-Scale Dataset for Testing Robustness of AI-Generated Image Detectors

**Anonymous Author(s)**
Affiliation
Address
email

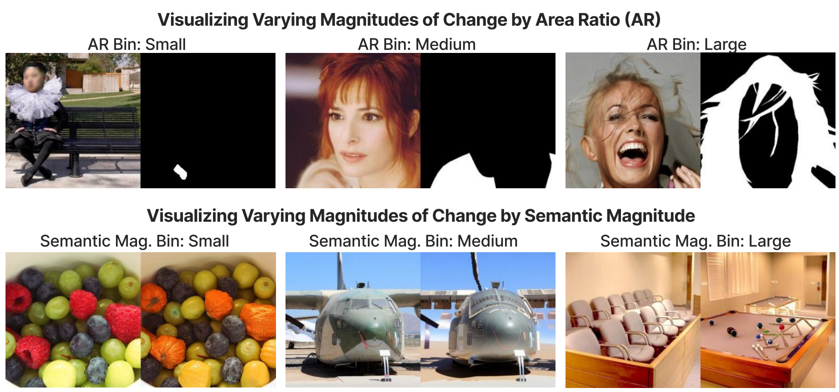


Figure 1: Various resolutions of change quantified with respect to area of augmentation. This information was computed via (1) image masks used for inpainting, or (2) post-generation methods using MSE based custom metric from cross-attention-based editing.

## Abstract

While text-to-image diffusion models have demonstrated impactful applications in art, design, and entertainment, these technologies also facilitate the spread of misinformation. Recent efforts have developed AI-generated image detectors claiming robustness against various augmentations, but their effectiveness remains unclear. Can these systems detect varying degrees of augmentation? Do they exhibit biases towards specific scenes or data distributions? To address these questions, we introduce SEMI-TRUTHS, featuring $27,635$ real images, $245,360$ masks, and $850,226$ AI-augmented images featuring varying degrees of targeted and localized edits, created using diverse augmentation methods, diffusion models, and data distributions. Each augmented image includes detailed metadata for standardized, targeted evaluation of detector robustness. Our findings suggest that state-of-the-art detectors are sensitive to different degrees of edits, data distributions, and editing techniques, providing deeper insights into their functionality.

## 1 Introduction

The rise of text-to-image generative models has democratized automated image creation for both ML practitioners and the general public. While existing architectures like Variational Autoencoders [81, 29] and GANs [4, 96, 27, 32, 35] have produced realistic images for years, diffusion models [15, 66, 13] have enhanced image quality, diversity, and ease of use, driving their rapid adoption. However, this technology is a double-edged sword. Despite its applications in art, design, marketing, and entertainment [31, 91], as it becomes increasingly pervasive, it's critical to identify and understand misuse that spreads misinformation [90, 52]. In recent events, AI-generated images have been increasingly used for harmful purposes like spreading misinformation and committing crimes such as fraud, defamation, and identity theft [23, 76]. One alarming factor associated with these models is their ability to alter small attributes of an original image, we refer to such images as semi-truths. A notable example is the spread of false propaganda during the Israel-Palestine conflict [40]. Rather than creating images from scratch, individuals often alter specific parts or attributes to evade detection. For instance, the "Sleepy Joe" [69] video circulated on Twitter in 2020, where President Joe Biden's face was edited to appear as if he fell asleep during an interview. The implications of such subtle edits and their potential to spread misinformation underscore the critical need for automated detection of such attacks.

| Dataset | Magnitude of Change | Targeted Editing | Quality Check | Data Collection | Generation | | | Data Dist. | | Scale | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | GANs | Diffusion | #Methods | Scene | #Real Bench. | Real | Fake |
| 1 DFDC [6] | ✗ | ✗ | ✗ | Generated | ✓ | ✗ | 8 | Face | 1 | $488.4k$ | ∼1.7M |
| 2 FaceForensics++ [68] | ✗ | ✗ | ✗ | Generated | ✓ | ✗ | 4 | Face | 1 | $509.9k$ | ∼1.8M |
| 3 Celeb-DF [93] | ✗ | ✗ | ✓ | Generated | ✓ | ✗ | 1 | Face | 1 | $225.4k$ | ∼2.1M |
| 4 DeepFakeFace [73] | ✗ | ✗ | ✗ | Generated | ✗ | ✓ | 3 | Face | 1 | $30k$ | $90k$ |
| 5 CIFAKE [5] | ✗ | ✗ | ✗ | Generated | ✗ | ✓ | 1 | General | 1 | $60k$ | $60k$ |
| 6 DiffusionDB [87] | ✗ | ✗ | ✓ | Sourced | ✗ | ✓ | 1 | General | 0 | 0 | 14M |
| 7 MidJourney prompts [80] | ✗ | ✗ | ✗ | Sourced | ✗ | ✓ | 1 | General | 0 | 0 | $248k$ |
| 8 TWIGMA [10] | ✗ | ✗ | ✗ | Sourced | ✗ | ✓ | unknown | General | 0 | 0 | $800k$ |
| 9 GenImage [98] | ✗ | ✗ | ✗ | Generated | ✓ | ✓ | 8 | General | 1 | 1.33M | 1.35M |
| 10 SEMI-TRUTHS | ✓ | ✓ | ✓ | Generated | ✗ | ✓ | 8 | General | 6 | 27,635 | ∼850k |

Table 1: **SEMI-TRUTHS vs other AI-generated image datasets.** We compare SEMI-TRUTHS with other AI-generated image datasets across multiple categories: (1) Magnitude of Change: provides metadata on the magnitude of perturbations; (2) Targeted Editing: performs targeted editing of images; (3) Quality Check: quality assessment of fake images; (4) Data Collection: data collection strategy, *Generated* or *Sourced* from publicly available portals; (5) Generation: generator category and number of methods used (TWIGMA's method was unknown since its images were sourced from Twitter); (6) Data Distribution: scene variation and diversity of real benchmarks; (7) Scale: number of real and fake images.

However, existing datasets for training and evaluating AI-generated image detectors primarily consist of fully synthesized images, often limited to human faces [6, 68, 93, 36, 14]. This narrow focus fails to capture the diversity of real-world augmentations and does not reveal model biases toward different degrees of augmentation. To address this, we introduce SEMI-TRUTHS, which includes AI-augmented images with varying levels of perturbation (detailed comparison in Table. 1, enabling the evaluation of detectors against more realistic and diverse attacks like the "Sleepy Joe" video [69].

We categorize the magnitude of change in SEMI-TRUTHS using two criteria: (1) the size of the augmented region, and (2) the semantic change achieved. Quantitative metrics are used to quantify the degree of semantic change and their efficacy is validated by evaluating their correlations with human judgment. Each original and altered image pair is annotated with descriptive features representing these changes. Synthetic images in SEMI-TRUTHS are created using diffusion inpainting and prompt-based-editing editing [25, 51] for 5 different diffusion algorithms [60, 71, 58, 67]. To avoid data distribution bias, the original images are sourced from 6 existing semantic segmentation benchmarks. Our approach to curating SEMI-TRUTHS employs a flexible, plug-and-play method for human-guidance-free image editing followed by model sensitivity analysis. This ensures reusability and applicability to new data distributions, large language models for prompt perturbation, and various image synthesis methods.

Finally, we demonstrate how the knowledge abstractions in SEMI-TRUTHS can be used to identify the sensitivities of existing detectors. By stress-testing 6 models, we reveal unique sensitivities to different data distributions, diffusion models, and perturbation degrees. Our goal is to offer a resource for targeted, interpretable, and standardized evaluation of AI-Generated image detection systems, and to provide a customizable evaluation pipeline for the community.

## 2 Related Work

**AI Generated Image dataset** The field of AI-based image generation and editing has rapidly evolved from autoencoders [18] and graphics-based techniques [78] to GANs [97, 55, 2, 46, 7] and, more recently, diffusion models [54, 67, 58, 21]. These advancements have heightened ethical concerns regarding identity theft and misinformation, [3, 24, 28] necessitating robust datasets for AI-generated image detection. While most research has focused on GAN-generated human faces [6, 68, 93, 36, 14], there is a growing emphasis on diffusion-based techniques for detection of deepfakes [73], digital forgery [72] and generic AI-generated content [98, 5, 80, 87]. However, existing datasets face several limitations that restrict their applicability as a benchmark for developing robust detection systems. They often come from a single model [80, 87] or source data distribution [98, 5], lack detailed generation and image metadata [10], and provide limited control over degree and quality of edits [80, 87, 98, 5, 73, 10, 63]. Furthermore, they do not offer scalable pipelines for integrating future image generation and editing techniques and are limited in their analysis of detection methods. Recognizing these gaps, we introduce SEMI-TRUTHS that incorporates multiple model variations, editing techniques, and source data distributions, provides comprehensive metadata, and offers fine-grained control over the quality and degree of edits (Table. 1 summarizes SEMI-TRUTHS's contributions).

**Image editing pipelines** With the advent of diffusion models, the field of image editing has seen tremendous advancements [30]. Recent developments in image inpainting, both in text-conditioned [88, 89, 84, 92] and unconditioned [48] settings, have enabled fine-grained control over image editing significantly enhancing precision and quality. While image inpainting requires the use of masks, prompt-based image editing [25, 51] performs targeted edits conditioned solely on text prompts. Existing frameworks like LANCE [59] and InstructPix2Pix [8] leverage this capability to develop automated image editing pipelines. LANCE [59], leveraging large language models (LLMs)[79] and image captioning[43], enables human-supervision-free image edits across diverse perturbations. Building on this, we extend LANCE [59] to handle a broader range of perturbation magnitudes, guided by semantic change definitions [9, 33]. Our approach integrates LlaVA [47] and LLAMA [79] models, combining inpainting and prompt-based techniques for precise, contextually informed edits.

**Stress Testing Pipelines** Stress testing pipelines, crucial in software engineering, remain underutilized in machine learning. While various metrics exist for performance assessment and model comparison [64], they often lack the depth to fully capture model robustness and explain failure cases adequately. While initiatives like Stress Test NLI [53] focus on generating adversarial examples to evaluate models' inferential capabilities across six tasks, DynaBench [37] and CheckList [65] take a different approach by employing human-in-the-loop systems to dynamically benchmark and assess the robustness of natural language models in real-world scenarios. Simultaneously, in the vision community, Li et al. [44] utilize diffusion models to create ImageNet-E, honing in on assessing classifier robustness through object attributes, while Luo et al [49]. explore model sensitivity to user-defined text attributes using StyleGAN [2]. Building upon these endeavors, LANCE [59] advances the field by extracting insights from failures via a targeted editing algorithm, enabling stress testing across diverse attributes. Our work extends this paradigm to AI-generated image detection, presenting a versatile pipeline capable of performing image edits with varying magnitudes of perturbations across any diffusion model for a given set of image data points, facilitating evaluation and bias discovery in detector architectures through a comprehensive range of stress tests.

# 3 SEMI-TRUTHS

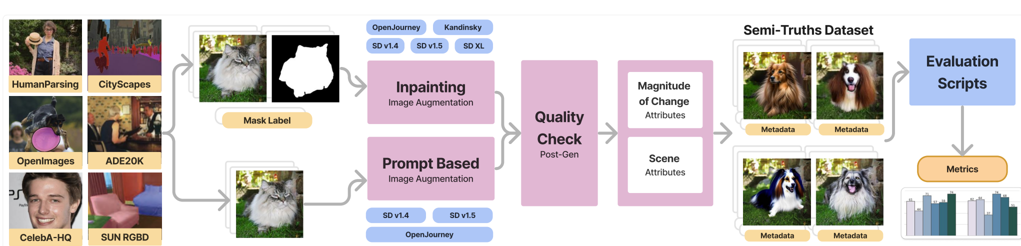


Figure 2: **End-to-end pipeline for SEMI-TRUTHS curation and detector stress testing.** The SEMI-TRUTHS pipeline sources data from 6 benchmarks and uses two editing techniques to perturb images. These images undergo quality checks, metric analysis, and stress testing of detectors across our curated tests.

To precisely evaluate a detector's ability to distinguish between AI-generated and real images, we curate SEMI-TRUTHS, consisting of over $27,635$ real images and $850,226$ fake images. We consider several crucial factors: strategies for targeted editing at varying magnitudes of augmentation, diversification of scene distributions, generation techniques, perturbation methods, and the quality of generated images. This section outlines the methods used to guide and quantify the magnitudes of augmentation, followed by a description of our generation and quality check pipeline. Finally, we detail the various aspects of the curated dataset.

| ***Small Changes:*** | ***Medium Changes:*** | ***Large Changes:*** |
|---|---|---|
| Do not significantly alter the overall meaning or context of the image. This could include changing the color of a specific object, adding or removing a minor detail, adjusting the composition or perspective of the image, or slightly adjusting the color distribution of the image. | Slightly alter the viewer's perception of the image and its subject. They could involve minor changes to an object or its setting, like altering a background element, moving an object or person to another location within the frame, or changing the emotions of the people in the frame. | Involve substantial modifications to the image that fundamentally transform its interpretation or message. It may even appear surprising or strange to an audience. This could include altering, adding or removing major elements of the image background and making changes to the subject of the image. |

Table 2: **Semantic Taxonomy** Magnitudes of semantic change, used to guide the perturbation of image captions and mask labels for targeted image generation.

### 3.1 Magnitudes of Augmentation

The alteration made to an image can be quantified in two ways: (1) the proportion of the image area that has been altered (area ratio of change), and (2) the degree to which the semantics of the image were altered (semantic change). To control the degree of alteration along these axioms, we start with an initial description of the image. This description is obtained by selecting a segmentation mask and the corresponding class label for local understanding, or, in the absence of mask information, by generating a caption for the image using BLIP [43].

**Introducing Perturbations** Motivated by the categorization of semantic and abstract content from visual semantics research [9], we create a taxonomy for small, medium, and large semantic changes (see Table 2). This taxonomy is used to guide the perturbation of an image caption or mask label using LLaVA-Mistral-7B [47] or LLAMA-7B [79].[1] As shown in Figure 3, the model is provided with a semantic magnitude category, its definition, a caption to perturb, and the image (if using LLaVA-Mistral-7b). For prompt-based-editing, a diffusion model edits images based on perturbed captions, introducing semantic changes. In inpainting, the mask and perturbed label restrict the area of change based on mask size, allowing precise control over alterations in the image area and semantics.

**Measuring Surface Area Change** While segmentation masks help localize perturbations to an image, providing a ratio for measuring Surface Area Change, diffusion model imprecision can compromise their accuracy. Dong et al. [16] demonstrate diffusion models can "color outside the box" during inpainting. Furthermore, the lack of mask guidance in prompt-based-editing necessitates the use of post-editing metrics. Therefore we employ SSIM [86], MSE, and a custom metric[1] which collectively assess the extent to which the structural components and the number of pixels differ between the original and perturbed images. Our custom metric, derived from MSE, uses thresholding to remove noisy components followed by connected component analysis to generate masks indicating areas of change. Similar to the area ratio computed using the mask and the image, we compute a ratio using the generated mask to quantify the surface area of change. Each of these metrics is normalized between 0 and 1 and categorized into small, medium, and large changes based on percentiles: the bottom $25^{th}$ percentile for small, the $25^{th}$ to $75^{th}$ percentile for medium and anything beyond the $75^{th}$ percentile for large.

**Measuring Semantic Change** As mentioned previously, the pre-editing semantic change metric is defined according to the taxonomy presented in Table 2. However, the stochasticity of large language models (LLMs) and diffusion models necessitates the implementation of post-editing metrics that provide a quantitative measure of semantic change. We use three different scores: LPIPS [94] and DreamSim [19], computed between the original and perturbed images, and Sentence Similarity [75],

[1]Additional details provided in the supplementary

calculated between the original and perturbed captions/mask labels.[1]. These metrics are normalized and categorized like Surface Area Change metrics, indicating small, medium, and large changes.

### 3.2 Image Editing Pipeline

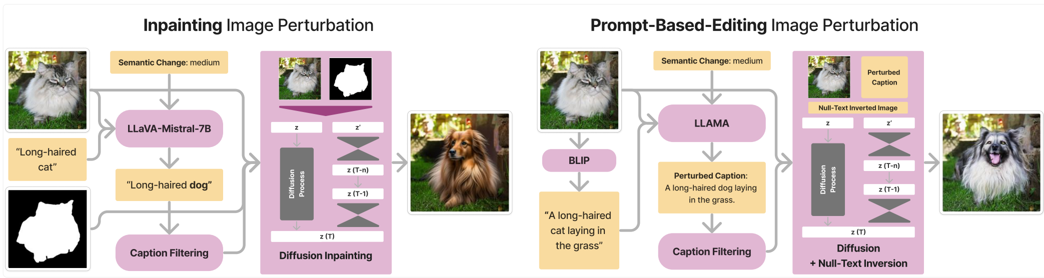


Figure 3: **Image Editing Pipeline.** Components of the image perturbation process for SEMI-TRUTHS curation using inpainting and prompt-based-editing methods.

Our image editing pipeline, delineated in Fig. 2, expands upon the work of LANCE [59] by integrating two distinct image editing techniques: (1) inpainting and (2) prompt-based-editing. Additionally, we tailor the pipeline for inpainting by leveraging LlaVA-Mistral-7B [47] to generate zero-shot mask label perturbations across various augmentation levels (detailed in Sec. 3.1) and diffusion models.[1] Furthermore, the multiple components of this pipeline demand comprehensive quality checks at each stage to ensure that the resulting images maintain structural integrity and align with the specified directions of change. To this end, we implement two rounds of data pruning within our image editing pipeline to eliminate instances of poor-quality text and image generation. Our multi-stage quality check pipeline is detailed below.

**Caption Filtering** The caption filtering stage initiates the quality check pipeline, ensuring two key aspects: (1) accuracy of generated BLIP [43] captions for prompt-based-editing in representing relevant image information, and (2) coherence and desirability of image edits produced by perturbed captions/labels, ensuring semantic alignment with original content. For the former, CLIPScore [26] measures the difference between embeddings of the original image and its generated caption, filtering out the lowest 5th percentile values. For the latter, cosine similarity between CLIP [62] text embeddings of the perturbed caption/mask label and the original is calculated, removing values above the $95^{th}$ percentile (negligible change) and below the $5^{th}$ percentile (semantic incoherence).[1]

**Post Image Edit Quality Check** In the second stage of the quality check pipeline, we aim to (1) evaluate the overall quality of generated images, ensuring semantic coherence and accurate augmentation while retaining resemblance to the original, and (2) filter out instances where diffusion models fail to incorporate desired edits. Since our images represent augmentations, conventional metrics like PSNR and SSIM [86] aren't applicable as they require a reference image. We use BRISQUE [50], a reference-free metric, discarding images with a score over 70 (top 0.3 percentile). Similarly, to ensure that the desired edits are accurately reflected in the image, we use CLIP similarity [62] between original and perturbed images, ensuring the diffusion model performed edits on the original. We also employ CLIP directional similarity [20] to confirm changes in images align with changes in captions/labels. Images between the $20^{th}$ and $80^{th}$ percentile are considered. [1]

### 3.3 SEMI-TRUTHS Details

**Data Distribution** We collect data from 6 semantic segmentation benchmarks representing various data distributions: CityScapes [12] for urban outdoor scenes, SUN RGBD [74] for indoor room scenes, CelebA HQ for human faces [34], Human Parsing for full-body portraits [45], and ADE20K [95] and OpenImages [41] for diverse themes. This combined real dataset comprises $27,635$ real images and $245,360$ masks. Using inpainting and prompt-based-editing techniques across 6 [58, 60, 71, 67] diffusion models for inpainting and 3 [60, 67] diffusion models for prompt-based-editing, with LlaVA-Mistral-7B [47] and LLAMA-7B [79] for prompt perturbation, we create $367,862$ prompt-based-editing datapoints and $1,087,865$ inpainting datapoints. After post-edit quality checks, and filtering out poor-quality generations, we retain $688,914$ inpainting augmented images and $161,312$ in prompt-based-editing augmented images, totaling $850,226$ images.[1]

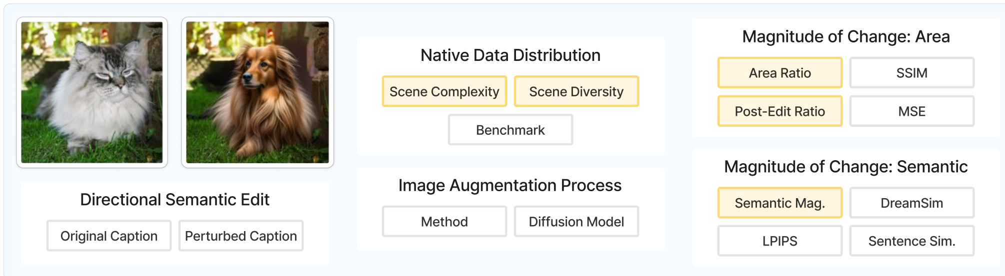


Figure 4: **SEMI-TRUTHS details and metadata.** Metadata used to describe every generated image in SEMI-TRUTHS. Attributes highlighted in yellow are novel contributions presented in this work.

**Metadata** SEMI-TRUTHS encompasses extensive metadata accompanying both real and fake image pairs and masks, offering insights into every facet of the generation process(see Fig. 4). This metadata includes details about the source data distribution, such as the original benchmark from which the image was sourced, scene complexity and diversity (defined by the number and variety of scene elements), a list of unique entities present in each image, and the ratio of mask-occupied area. Additionally, it provides information about the diffusion model, editing technique, and language model utilized for each edit, alongside the original and perturbed caption/label. Furthermore, each edited image is accompanied by quantitative and qualitative measures of change categorized across semantic and surface area-based metrics, as outlined in section 3.1. The metadata also indicates whether the change is categorized as diffuse or localized, determined using a custom algorithm detailed in the supplementary materials. All of this information is very crucial for testing the effectiveness of detectors across various axes as demonstrated in Sec. 4.

# 4 Experiments

| Detector | Backbone | Training Data Distribution | | | Precision(↑) | | | Recall(↑) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Scene | GANs | Diffusion | All | Real | Fake | All | Real | Fake |
| 1 DINOv2 [57] | ViT [17] + ResNet-50 [22] | General | ✗ | ✗ | 29.30 | 37.17 | 21.43 | 49.99 | 99.96 | 00.01 |
| 2 CNNSpot [83] | ResNet-50 [22] | General | ✓ | ✗ | 30.13 | 35.27 | 25.00 | 49.99 | 99.99 | 00.00 |
| 3 DIRE [85] | ResNet-50 [22] | General | ✗ | ✓ | 31.09 | 37.18 | 25.00 | 49.99 | 99.99 | 00.00 |
| 4 CrossEfficientViT [11] | EfficientNet-B0 [77] + ViT [17] | Face | ✓ | ✗ | 46.37 | 34.89 | 57.85 | 46.58 | 62.87 | 30.28 |
| 5 UniversalFakeDetect [56] | CLIP [62]-ViT [17] | General | ✓ | ✓ | 64.84 | 58.89 | 70.79 | 60.57 | 34.11 | 87.03 |
| 6 DE-FAKE [70] | CLIP [62] | General | ✓ | ✓ | 61.65 | 49.97 | 73.33 | 61.88 | 52.28 | 71.48 |

Table 3: **Documentation of each AI-generated Image Detection model evaluated with SEMI-TRUTHS.** We evaluated six detectors with diverse backbones and training data distributions. Models performing satisfactorily, highlighted in green, were selected for further tests.

We conduct extensive experiments with SEMI-TRUTHS to evaluate the effectiveness of AI-generated image detectors in distinguishing between real and AI-generated content (see Table. 4). In the sections below, we demonstrate how the knowledge abstraction over image augmentations in the dataset can be used to identify nuanced biases in various detectors.[1] All evaluation are conducted on a 10% sample of SEMI-TRUTHS, containing a total of 87,000 images (27,000 real and 60,000 augmented).

**Overall Detector Performance** We select a diverse set of open-source AI-generated image detectors for stress testing. As demonstrated in Table 4, each model has a unique architecture and training distribution. As a preliminary step, we assess the overall performance of these detectors evaluated in a zero-shot setting using generic quantitative metrics such as Precision, Recall, and F1-Score to identify the top-performing models for more detailed analysis. Of 6 models selected for stress-testing, half did not demonstrate performance metrics substantial enough to enable further evaluation. These include (1) DinoV2, a foundation vision model that was evaluated for zero-shot prediction of AI-generated Images, (2) CNNSpot, a ResNet-50 backbone exclusively trained on GAN-generated content, and (3) DIRE, a ResNet-50 backbone model which despite being trained on diffusion-generated content failed to demonstrate competitive metrics.

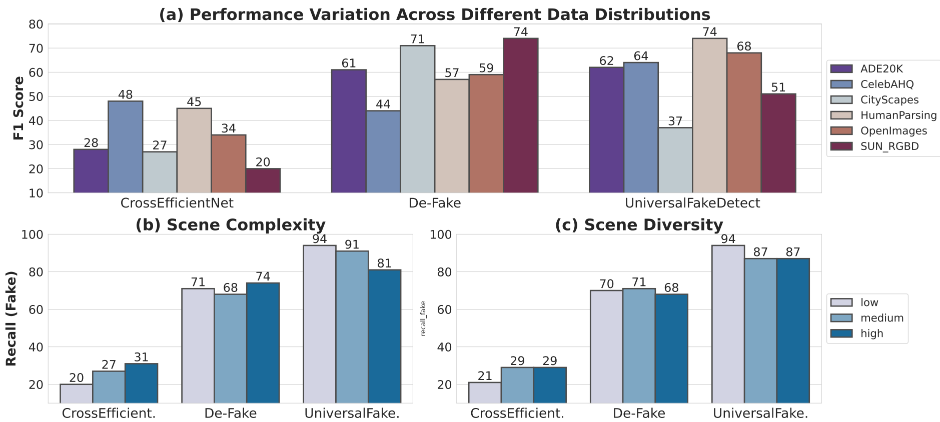


Figure 5: **Detectors are sensitive to semantic aspects of data distribution** Variation in the performance of AI-generated image detectors with respect to different benchmarks

**Sensitivity to Data Distribution** To gauge potential biases of detectors to different data distributions, we evaluate each model with respect to benchmarks present in SEMI-TRUTHS. In Figure 5 we demonstrate that each detector exhibits significant variation in performance. Notably, CrossEfficientViT [11], which is trained on GAN-generated images of human faces, exhibits a significant drop in performance on human faces sourced from benchmarks ADE20K, CityScapes [12], and SUN-RGBD [74] (CrossEfficientViT pre-emptively filters any images that do not contain a human face). In contrast, DE-FAKE [70], trained on more general scene images, exhibits the worst performance on CelebA-HQ [42] and HumanParsing [45] due to limited focus on humans and portrait-like images in its training distribution. On the other hand, UniversalFakeDetect [56], trained on indoor bedroom images as well as other generic scenes, fails to perform well with SUN RGBD and shows a remarkable drop in performance on CityScapes.

Furthermore, we investigate the detectors' ability to handle highly complex and diverse multi-instance scenes. We evaluate performance across varying levels of Scene Diversity (number of unique class instances in the images) and Scene Complexity (number of instances in total), categorized into small, medium, and large bins.[1] We find that UniversalFakeDetect's [70] performance degrades gradually with increasing scene diversity and complexity. In contrast, DE-FAKE [70] remains fairly robust across different scene variations. Interestingly, CrossEfficientViT [11] shows improved performance with increasing scene complexity and diversity, which can be attributed to human-centered benchmarks like CelebA-HQ [42] and HumanParsing [45] segmenting distinct facial features and body parts. In this setting, lower Scene Complexity may indicate a partial image of a face. These results highlight that detectors are highly sensitive to the semantic attributes of data distributions, emphasizing the importance of stress tests to identify and address distributional weaknesses.

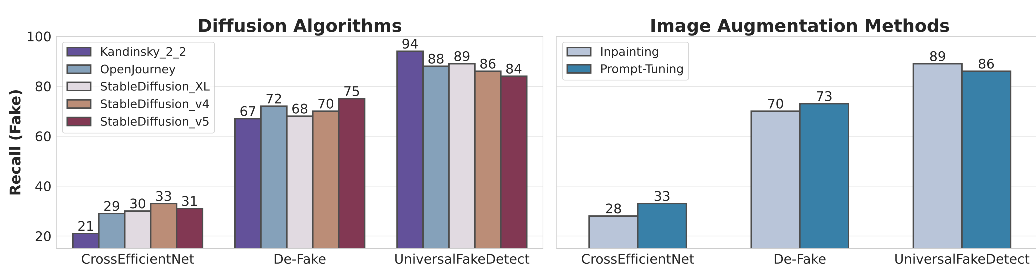


Figure 6: **Performance variation across image augmentation methods and diffusion algorithms** SEMI-TRUTHS offers data generated using various diffusion algorithms and augmentation methods facilitating detector evaluation across these aspects

**Evaluation across Editing Techniques and Models** SEMI-TRUTHS contains images generated using two different augmentation approaches - inpainting and prompt-based-editing - as well as five different diffusion algorithms - StableDiffusion v1.4, StableDiffusion v1.5, StableDiffusion XL [58], OpenJourney [60], and Kandinsky 2.2 [71]. This diversity in generated content enables

| Phrase(Original ⟶ Edited) | Counts | Recall |
|---|---|---|
| Easy cases | | |
| 1 lower lip ⟶ nose | 70 | 66.67 |
| 2left brow ⟶ left brow with slight arch | 99 | 50.0 |
| 3 car ⟶ car with shiny chrome accents | 59 | 45.16 |
| Difficult cases | | |
| 4 lower lip ⟶ lipstick | 190 | 15.79 |
| 5 skin ⟶ skin with subtle freckles | 127 | 7.14 |
| 6 left ear ⟶ earring | 177 | 6.67 |

(a) **CrossEfficientViT [11]**

| Phrase(Original ⟶ Edited) | Counts | Recall |
|---|---|---|
| Easy cases | | |
| 1 skin ⟶ leather | 74 | 98.65 |
| 2 nose ⟶ nose ring | 138 | 97.1 |
| 3 left ear ⟶ earring | 177 | 96.61 |
| Difficult cases | | |
| 4 vegetation ⟶ tree | 225 | 66.67 |
| 5 ego vehicle ⟶ mercedezbenz | 161 | 65.84 |
| 6 vegetation ⟶ building | 150 | 65.33 |

(b) **UniversalFakeDetector [56]**

| Phrase(Original ⟶ Edited) | Counts | Recall |
|---|---|---|
| Easy cases | | |
| 1 car ⟶ car with shiny silver paint | 57 | 85.96 |
| 2 vegetation ⟶ tree | 225 | 84.89 |
| 3 ego vehicle ⟶ mercedezbenz | 161 | 81.37 |
| Difficult cases | | |
| 4 skin ⟶ skin with subtle freckles | 127 | 62.99 |
| 5 nose ⟶ nose ring | 138 | 58.57 |
| 6 skin ⟶ leather | 74 | 58.11 |

(c) **De-FAKE [70]**

Table 4: **Directional Semantic Edits for investigating detector biases.** Directional Semantic Edits provide insights on which edits to a certain entity has a higher chance of fooling detectors

investigation of detector sensitivities to different augmentation procedures.[2] Figure 7 shows that UniversalFakeDetect [56] performs best on images augmented with Kandinsky 2.2 [71] and worst on images augmented with StableDiffusion v1.5 [67]. The difference in Recall score is 10%. The inverse is true for DE-FAKE [70]. CrossEfficientVit [11] performs best on images augmented with StableDiffusion v1.4 and worst with Kandinsky 2.2 [71] with a 12% drop in performance. Additionally, we see that CrossEfficientViT [11] and DE-FAKE [70] are more sensitive to inpainted images, whereas UniversalFakeDetect [56] performs worst on prompt-based-editing content.

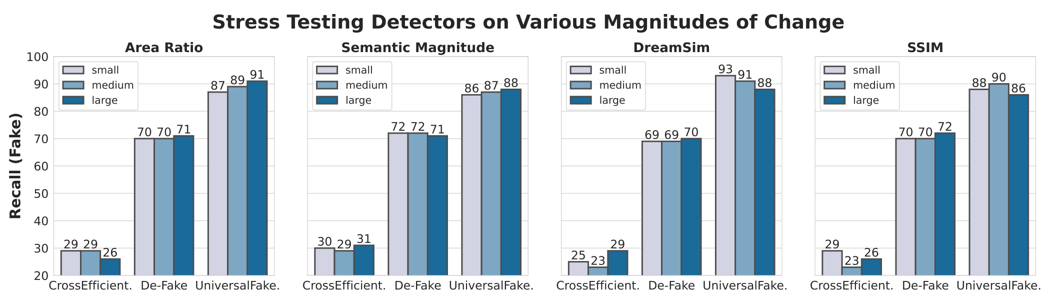


Figure 7: **Performance variation of select detectors across various magnitudes of augmentation** DE-FAKE [70] is robust across the board, Area Ratio captures the sensitivity exhibited in UniversalFakeDetect [56] and CrossEfficientViT [11]

**Evaluation across Varying Magnitudes of Augmentation** As detailed in Sec. 3.1, each image in SEMI-TRUTHS is fitted with an array of descriptive attributes that capture the magnitude of change. In Figure 7 we examine the impact of varying levels of perturbations on detector performance, focusing on both surface area and semantic changes. Note that CrossEfficientViT [11] performs better on smaller values of Area Ratio, where as UniversalFakeDetect [56] performs better on larger changes. UniversalFakeDetect's [56] performance also drops as DreamSim [19] scores increase. Even though DE-FAKE [70] is not the best performing model, it appears to be the most robust against various magnitudes of change across the board. This evaluation procedure allows us to gauge which detectors exhibit some sensitivity to different degrees of augmentation and which don't.

**Directional Semantic Edits** When describing how the semantics of an image change or how the story it portrays evolves, many quantitative metrics can be reductive. Transitioning into an embedded space to assess similarity often results in significant information loss. To address this issue, we introduce "Directional Semantic Edit" which groups generated images from SEMI-TRUTHS by original caption/mask label pairs and their perturbed versions. In the evaluation set, certain directional semantic edits occurred as frequently as 445 times. Each detector is evaluated on these groups, and metrics are sorted by Recall, as shown in Table 4. Each model exhibits distinct performance variations based on specific semantic changes. Notably, UniversalFakeDetect [56] performs best on changes to facial features but worst on changes to vegetation. Conversely, DE-FAKE [70] excels at detecting changes to cars and vegetation but struggles with changes to human faces. CrossEfficientViT [11] shows varied performance with changes to human faces, appearing in both its highest and lowest ranks, indicating sensitivity to the magnitude of the change. Furthermore, analyzing these edits can maximize the potential of these algorithms by informing decisions about the most suitable ensemble

[2]Limitations of [25], [51] restrict prompt-based-editing to StableDiffusion v1.4, StableDiffusion v1.5, [67] OpenJourney [60]

| Correlation Coeff. | Change Metrics(↑) | | |
|---|---|---|---|
| | Area Ratio | LPIPS Score | SSIM |
| 1 Pearson | 0.46 | 0.14 | −0.16 |
| 2 Kendall-Tau | 0.40 | 0.15 | −0.14 |
| 3 Spearman | 0.50 | 0.19 | −0.17 |

Table 5: **Correlation between quantitative measures of change and Human Perception** Correlation coefficients computed between human annotated magnitudes of change and quantitative metrics available in the dataset. Quantitative metrics not displayed here had coefficients 0.10.

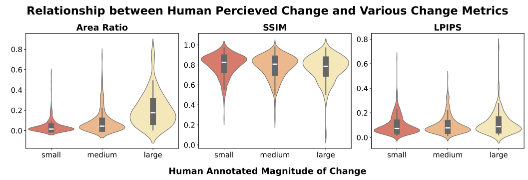


Figure 8: **Relationship between quantitative change metrics and Human Perception of change (small, medium, large) in SEMI-TRUTHS** Each violin plot shows the distribution of metric values for a change category.

techniques. For example, while UniversalFakeDetect [56] struggles with vegetation-to-tree edits, DE-FAKE [70] excels, suggesting a suitable combination for ensemble approaches. This type of analysis helps identify which directional edits are most challenging and confounding for these detector models, providing deeper insights into their function and limitations.

**Surveying Human Perception of Magnitudes of Change** In this work, we leverage several algorithms to capture the degree of visual and semantic change achieved during image augmentation. However, how do these measures compare to human perception? We aim to build an intuitive understanding of which metrics correlate with how a person may perceive the magnitude of change. We conduct a user study where annotators classify the difference between pairs of original and augmented images into "not much", "some", and "a lot", corresponding to our "small", "medium", and "large" change bins.[1] We then compute correlation coefficients (Pearson [38], Kendall Tau [61], and Spearman [1]) between human scores and quantitative measures in SEMI-TRUTHS. ( see Table.5).

# 5 Discussion

**Limitations and Future Work** Our in-painting pipeline currently relies on manual semantic mask input, limiting usability. To improve, we'll integrate automatic mask generation methods like SAM [39] similar to InstructEdit [82]. Additionally, using LLAMA-7B [79] and LlaVA [47] models for zero-shot editing led to many poor-quality outputs, requiring filtering. Future iterations will involve fine-tuning these models. We are also aware of potential biases in metrics like LPIPS [94], Sentence Similarity [75], and DreamSim [19], which may impact evaluations.

**Ethical Issues** While our project aims to generate specific perturbations to improve detectors, it could be used to create sophisticated fake images capable of deceiving fake image detectors, potentially facilitating misinformation and deepfakes. Additionally, despite our efforts at diversification of data and models, inherent biases from these modules may persist which can perpetuate or exacerbate existing inequalities, resulting in uneven performance across different contexts and types of images.

# 6 Conclusion

To tackle the growing risk of misinformation from AI-generated images, it is crucial that detectors are robust against perturbations. Hence, we introduce SEMI-TRUTHS, housing 850, 226 AI-generated images with detailed metadata on source data distribution, scene complexity, diversity, editing techniques, change magnitudes, directional edits, and both original and perturbed captions. Our plug-and-play image editing pipeline enables easy generation of additional augmentations for any image, along with a standardized platform for investigating detector robustness through a suite of curated tests. Our findings reveal that state-of-the-art detectors are sensitive to different degrees of edits, data distributions, and editing techniques, and provide deeper insights into their functionality. Moreover, we introduce a semantic taxonomy for defining semantic change and employ a rigorous quality check pipeline for ensuring image quality. Through thorough human evaluation, we ensure alignment between the magnitude of our edits and human perception.

In conclusion, we believe the user-friendly design of SEMI-TRUTHS will facilitate ongoing research into robustness against future generative models, helping to combat misinformation effectively.

# References

[1] *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space?, 2019.

[3] C. Avey. Ethical pros and cons of ai image generation. *IEEE Computer Society/Tech News/-Community Voices*, December 27 2023.

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

[5] Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images, 2023.

[6] Ben Pflaum Jikuo Lu Russ Howes Menglin Wang Cristian Canton Ferrer Brian Dolhansky, Joanna Bitton. The deepfake detection challenge dataset, 2020.

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023.

[9] B. Burford, P. Briggs, and J. P. Eakins. A taxonomy of the image: On the classification of content for image retrieval. *Visual Communication*, 2(2):123–161, 2003.

[10] Yiqun Chen and James Zou. Twigma: A dataset of ai-generated images with metadata from twitter, 2023.

[11] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. *Combining EfficientNet and Vision Transformers for Video Deepfake Detection*, page 219–229. Springer International Publishing, 2022.

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016.

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[14] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation, 2020.

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[16] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models, 2023.

[17] 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.

[18] faceswap. Faceswap, 2019.

[19] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[20] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021.

[21] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis, 2022.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Kathleen Magramo Heather Chen. Finance worker pays out $25 million after video call with deepfake 'chief financial officer' | CNN — cnn.com. `https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html`.

[24] M. Heikkilä. The algorithm: Ai-generated art raises tricky questions about ethics, copyright, and security. *MIT Technology Review*, September 20 2022.

[25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022.

[27] R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-seeking generative adversarial networks, 2017.

[28] Tiffany Hsu and Stuart A Thompson. The new york times company. *A.I. Muddies Israel-Hamas War in Unexpected Way*, Oct 2023.

[29] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems*, 31, 2018.

[30] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey, 2024.

[31] Rowan T Hughes, Liming Zhu, and Tomasz Bednarz. Generative adversarial networks–enabled human–artificial intelligence collaborative applications for creative and design industries: A systematic review of current approaches and trends. *Frontiers in artificial intelligence*, 4:604234, 2021.

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[33] Alejandro Jaimes and Shih-Fu Chang. Conceptual framework for indexing visual information at multiple levels. In *Electronic imaging*, 1999.

[34] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. 2018.

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019.

[36] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022.

[37] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp, 2021.

[38] Wilhelm Kirch, editor. *Pearson's Correlation Coefficient*, pages 1090–1091. Springer Netherlands, Dordrecht, 2008.

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[40] David Klepper. Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead — apnews.com.

[41] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[42] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[44] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing, 2023.

[45] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(12):2402–2414, Dec 2015.

[46] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing, 2021.

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[48] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022.

[49] Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, and Fernando De la Torre. Zero-shot model diagnosis, 2023.

[50] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Blind/referenceless image spatial quality evaluator. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 723–727, 2011.

[51] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022.

[52] Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35, 2024.

[53] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference, 2018.

[54] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[55] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019.

[56] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models, 2023.

[57] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[59] Viraj Prabhu, Sriram Yenamandra, Prithvijit Chattopadhyay, and Judy Hoffman. Lance: Stress-testing visual models by generating language-guided counterfactual images, 2023.

[60] prompthero. Openjourney.

[61] Llukan Puka. *Kendall's Tau*, pages 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[63] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection, 2023.

[64] O. Rainio, J. Teuho, and R. Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086, 2024.

[65] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[66] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[68] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[69] Elyse Samuels. White House social media director tweets manipulated video of Biden. 9 2020.

[70] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023.

[71] Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. Ikandinsky 2.2, 2023.

[72] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models, 2022.

[73] Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models, 2023.

[74] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[75] Xiaofei Sun, Yuxian Meng, Xiang Ao, Fei Wu, Tianwei Zhang, Jiwei Li, and Chun Fan. Sentence similarity based on contexts, 2022.

[76] Natnicha Surasit. Criminal exploitation of deepfakes in South East Asia — globalinitiative.net. `https://globalinitiative.net/analysis/deepfakes-ai-cyber-scam-south-east-asia-organized-crime/`.

[77] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019.

[78] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016.

[79] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[80] Iulia Turc and Gaurav Nemade. Midjourney user prompts amp; generated images (250k), 2022.

[81] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[82] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions, 2023.

[83] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

[84] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting, 2023.

[85] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

[86] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[87] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[88] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model, 2022.

[89] Shaoan Xie, Yang Zhao, Zhisheng Xiao, Kelvin C. K. Chan, Yandong Li, Yanwu Xu, Kun Zhang, and Tingbo Hou. Dreaminpainter: Text-guided subject-driven image inpainting with diffusion models, 2023.

[90] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 9291–9298, New York, NY, USA, 2023. Association for Computing Machinery.

[91] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), nov 2023.

[92] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023.

[93] Pu Sun Honggang Qi Yuezun Li, Xin Yang and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[94] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[95] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[96] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[97] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[98] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023.