PDF Download
3746275.3762215.pdf
24 February 2026
Total Citations: 0
Total Downloads: 121

Published: 27 October 2025

Citation in BibTeX format

MM '25:The 33rd ACM International Conference on Multimedia
October 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

RESEARCH-ARTICLE

'Humor, Art, or Misinformation?': A Multimodal Dataset for Intent-Aware Synthetic Image Detection

ANASTASIOS SKOULARIKIS, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

STEFANOS IORDANIS PAPADOPOULOS, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

SYMEON PAPADOPOULOS, Centre for Research and Technology-Hellas, Thessaloniki, Macedonia, Greece

PANAGIOTIS C PETRANTONAKIS, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

Open Access Support provided by:

Aristotle University of Thessaloniki

Centre for Research and Technology-Hellas

DHOW '25: Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web (October 2025)
https://doi.org/10.1145/3746275.3762215
ISBN: 9798400720574

“Humor, Art, or Misinformation?”: A Multimodal Dataset for Intent-Aware Synthetic Image Detection

Anastasios Skoularikis
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki, Greece
skouanas@ece.auth.gr

Stefanos-Iordanis Papadopoulos*
Information Technology Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
stefpapad@iti.gr

Symeon Papadopoulos
Information Technology Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
papadop@iti.gr

Panagiotis C. Petrantonakis
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki.
Thessaloniki, Greece
ppetrant@ece.auth.gr

ABSTRACT

Recent advances in multimodal AI have enabled progress in detecting synthetic and out-of-context content. However, existing efforts largely overlook the intent behind AI-generated images. To fill this gap, we introduce *S-HArM*, a multimodal dataset for intent-aware classification, comprising 9,576 “in the wild” image–text pairs from Twitter/X and Reddit, labeled as *Humor/Satire*, *Art*, or *Misinformation*. Additionally, we explore three prompting strategies (image-guided, description-guided, and multimodally-guided) to construct a large-scale synthetic training dataset with Stable Diffusion. We conduct an extensive comparative study including modality fusion, contrastive learning, reconstruction networks, attention mechanisms, and large vision-language models. Our results show that models trained on image- and multimodally-guided data generalize better to “in the wild” content, due to preserved visual context. However, overall performance remains limited, highlighting the complexity of inferring intent and the need for specialized architectures.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Neural networks**; **Supervised learning by classification**; • **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

Multimodal Deep Learning; Synthetic Image Detection; AI-Generated Content; Intent-Aware Classification

DHOW ’25, October 27–28, 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2057-4/2025/10
https://doi.org/10.1145/3746275.3762215

ACM Reference Format:
Anastasios Skoularikis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2025. “Humor, Art, or Misinformation?”: A Multimodal Dataset for Intent-Aware Synthetic Image Detection. In *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web (DHOW ’25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746275.3762215

1 INTRODUCTION

The rise of generative models has significantly affected the landscape of digital media, enabling the large-scale and rapid production of realistic synthetic images that are becoming increasingly indistinguishable from real ones. From the inception of Generative Adversarial Networks (GANs) and their variants [1] to recent advances in diffusion models [2], synthetic image generation has become more powerful and accessible. The wide availability of tools such as Adobe Firefly, DALL·E, and Midjourney allow users to create realistic content with minimal effort and little to no technical expertise. While these tools can foster creative expression, particularly in art and satire, they also raise concerns over the misuse of synthetic content, with potentially serious social, political, and economic consequences [3].

Researchers have responded by developing numerous methods for synthetic image detection [4–11] and identifying multimodal forms of misinformation [12–26]. These efforts largely focus on detecting AI-generated, manipulated images, or decontextualized images; and have made significant progress in recent years. Nevertheless, these approaches often overlook a crucial aspect: the intent behind the creation of synthetic images; whether the image was meant as artistic expression, humor, or deliberate misinformation.

Intent classification can serve as a valuable first step in moderation pipelines by prioritizing harmful content (e.g., misinformation) for human review over benign cases like satire or art. Crucially, intent cannot always be inferred from visual content alone. An image generated for satire or creative purposes may be visually indistinguishable from one designed to mislead. To capture these subtle distinctions, a multimodal approach, integrating both visual and textual signals, is essential for uncovering the creator’s underlying intention. For example, consider Fig. 1, where an AI-generated image is accompanied by the user post: “Ancient stone in Antarctica.

[*] Corresponding Author. Also affiliated with the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki.

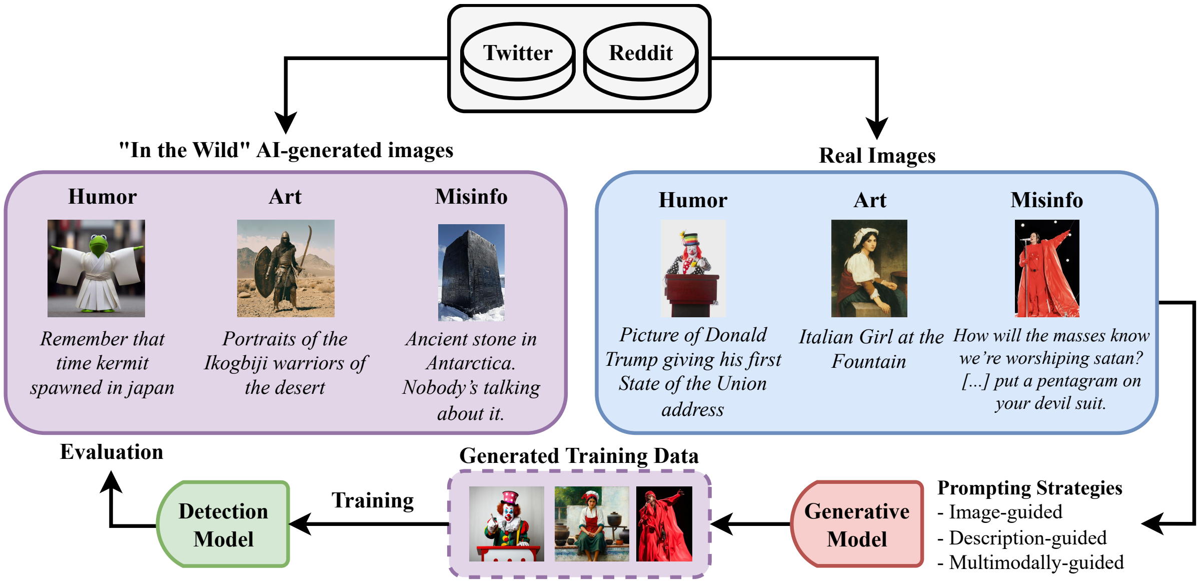

Twitter
Reddit
"In the Wild" AI-generated images
Humor
Art
Misinfo
Remember that time kermit spawned in japan
Portraits of the Ikogbiji warriors of the desert
Ancient stone in Antarctica. Nobody's talking about it.
Real Images
Humor
Art
Misinfo
Picture of Donald Trump giving his first State of the Union address
Italian Girl at the Fountain
How will the masses know we're worshiping satan? [...] put a pentagram on your devil suit.
Evaluation
Detection Model
Training
Generated Training Data
Generative Model
Prompting Strategies
- Image-guided
- Description-guided
- Multimodally-guided

Figure 1: Overview of the proposed pipeline. We collect "in the wild" AI-generated images to build the *S-HArM* evaluation benchmark, and real images to generate synthetic training data via Stable Diffusion using three prompting strategies. Various detection models are trained on synthetic training data and evaluated on the evaluation benchmark.

Nobody's talking about it. I wonder why." The claim is factually incorrect and implies a baseless conspiracy and the AI-generated image is used as "evidence" to support it. However, without the textual modality, the intent behind the generation of this image would be ambiguous.

To address this gap, we introduce *S-HArM* (Synthetic–Humor, Art, Misinformation), a multimodal dataset for intent-aware classification of AI-generated images. Unlike existing datasets that focus on binary classification (i.e., real vs. generated), *S-HArM* supports a three-way classification task: humor/satire, art, or misinformation. *S-HArM* comprises two components: (1) an annotated "in the wild" evaluation benchmark of AI-generated image created and shared by users in social media platforms, and (2) a larger training set of synthetic data generated from real images using Stable Diffusion. To build the evaluation benchmark, we curated content from Twitter/X's Community Notes—identifying examples of misinformation and satire—and from selected subreddits focused on art and humor. For training data, we collected additional Twitter/X and Reddit posts with real (non-generated) images, which were used to generate synthetic versions using three generative prompting strategies: *Image-guided*: the real image serves as the primary input to guide generation, *Description-guided*: a captioning model first produces a textual description of the real image, which is then used as input for text-to-image generation; *Multimodally-guided*: both the real image and its generated description are combined to jointly guide generation. In total, *S-HArM* includes 9,576 annotated evaluation image-text pairs, and over 87,000 synthetic training samples per generative prompting strategy.

Furthermore, we propose an end-to-end pipeline and conduct a comprehensive comparative study for intent-aware classification. As illustrated in Fig. 1, after collecting data and constructing both the "in the wild" evaluation benchmark and three versions of the training set, we train a broad range of models, including: unimodal baselines using either image-only or text-only inputs, multimodal baselines, contrastive learning, reconstruction networks, attention-based models, and large vision–language models (LVLMs), used as zero-shot classifiers. Finally, we evaluate the generalization of models trained on synthetic data to "in the wild" AI-generated content.

Our results show that models trained on data generated through image-guided and multimodally-guided strategies tend to generalize better due to the preservation of crucial visual characteristics. While performance on the synthetic validation set is very high (96.6%), generalization to "in the wild" content remains limited, with the best model reaching 71.6% accuracy. This attests the difficulty of intent-aware classification, which requires a deeper understanding of abstract concepts such as humor, misinformation, and artistic expression—often conveyed through subtle cross-modal cues. Finally, our empirical evaluations confirm that the task needs to be addressed multimodally, since unimodal baselines (image-only or text-only) consistently underperformed, suggesting that intent emerges from the interaction between visual and textual modalities. We share the code and dataset at: https://github.com/Qedrigord/SHARM.

2 RELATED WORK

The rise of generative AI has enabled both creative expression and harmful misuse. While these tools empower users to create digital art, satire, and memes, they are increasingly exploited for misinformation and manipulation [27, 28]. Within algorithmically driven echo chambers, personalized content reinforces existing beliefs, amplifying polarization and radicalization—especially due to cognitive biases like confirmation bias and the continued influence effect [27, 29]. AI-generated images amplify this risk by enabling personalized, deceptive narratives [30], fueling the rise of multimodal misinformation that blends synthetic visuals and text to mislead more effectively [12]. Deepfakes add further threats, including fraud, privacy violations, and political misuse. As manual moderation becomes unsustainable, scalable, automated, and intent-aware detection is urgently needed to protect digital trust [28].

2.1 Multimodal Misinformation Detection

Given the significant social impact of misinformation and the growing prevalence of multimodal forms of misinformation, extensive research has focused on developing effective detection methods, and particularly on Out-of-Context (OOC) misinformation. Recent approaches have leveraged fine-tuned CLIP embeddings [17], self-supervised distillation frameworks [19], and Transformer-based encoders for modeling joint cross-modal attention [20]. Other lines of work incorporate external evidence retrieved from the Web, using models that assess consistency [21], stance [22] or relevance of external evidence towards the image-text pair under examination [23]. Furthermore, there is also growing interest in using LVLMs due to their enhanced reasoning capabilities and potential for explainability [24, 25]. Nevertheless, a recent study has raised critical concerns regarding current benchmarks for OOC detection, showing that models often exploit dataset-specific artifacts rather than genuinely reasoning about factuality [26].

2.2 Multimodal Misinformation Datasets

To support research on multimodal misinformation detection, a range of datasets have been introduced. Early efforts include the MediaEval 2015 and 2016 Twitter datasets [12, 13], as well as the Weibo dataset [14], which contain social media posts annotated as real or fake based on whether the accompanying multimedia accurately reflects the described event. Due to their limited size and manual annotation effort, later works adopted large-scale weak supervision. For example, Fakeddit [15] includes over a million Reddit posts weakly labeled according to the originating subreddit. Other approaches generate synthetic misinformation through algorithmic means. COSMOS [16] randomly samples mismatched image–text pairs to simulate OOC misinformation. NewsCLIPpings [17] builds on legitimate image–text pairs from VisualNews [31], perturbing them via intra- and cross-modal similarity techniques to produce more realistic OOC samples. In contrast, the Multimodal Entity Image Repurposing (MEIR) introduce misinformation by manipulating named entities in textual content [18]. To evaluate the generalization of models trained on weakly annotated or synthetically generated datasets—and to mitigate unimodal biases—VERITE [32] was introduced as a robust benchmark for multimodal misinformation detection.

2.3 Synthetic Image Detection

Beyond multimodal misinformation detection, there has been growing interest in the detection of synthetic images. Early efforts relied on CNN-based detectors, evaluating their ability to generalize across various generative architectures [4, 33]. Other works targets artifacts introduced by CNN-based generators such as GANs and Diffusion models, either in the spatial domain [5] or frequency domain [6, 7]. Similarly, TextureCrop [8] enhances detection performance by extracting high-frequency image patches where generative artifacts are concentrated. Contrastive and self-supervised learning approaches are also widely used—for example, using InfoNCE loss to align image and text representations [9], or leveraging intermediate CLIP features to capture fine-grained inconsistencies [10], or self-supervised spectral reconstruction [11].

2.4 Synthetic Image Datasets

To support the development and evaluation of synthetic image detection methods, several datasets have been introduced. Both the ARTIFACT dataset [34] and Corvi et al. [35] use the COCO dataset [36] as a base to generate synthetic images across diverse categories, employing both Generative Adversarial Networks (GANs) and Diffusion Models. In contrast, Forensynths [4] relies exclusively on GANs and draws from a more varied set of source datasets. Datasets based solely on Diffusion Models include Synthbuster [6], which used the Raise [37] and Dresden [38] datasets to construct prompts, and CIFAKE [39], which generated images using CIFAR-10–inspired prompts to mimic the original dataset. A different approach is taken by TWIGMA [40], which curated AI-generated images directly from Twitter using specific hashtags. While these datasets provide valuable resources for synthetic image detection, they lack annotations related to humor/satire, art, or misinformation, making them unsuitable for our task.

3 CONSTRUCTION OF S-HARM

The Synthetic Humor, Art, or Misinformation (*S-HArM*) dataset consists of two main components: an annotated “in the wild” evaluation benchmark and a synthetic large-scale training set.

3.1 “In the Wild” Evaluation Benchmark

To construct the evaluation benchmark, we collected real-world, or “in the wild”, examples of AI-generated images created and shared by users on two major social media platforms: Twitter/X and Reddit.

3.1.1 *Reddit data.*

Reddit data was sourced from publicly available dumps, obtained using Arctic Shift[1]. To ensure the synthetic origin of the images, we focused exclusively on subreddits dedicated to content generated by AI-based models. We applied several quality control measures to ensure thematic consistency and data relevance: (1) Only subreddits with active moderation and clear thematic focus were considered, as off-topic posts are typically removed by community moderators. (2) *High approval threshold*: Posts were filtered for an upvote-ratio above 0.9, indicating strong community endorsement and alignment with subreddit themes. (3) When available, “post flairs” (tags and labels by users and/or moderators) were employed to assist in identifying content categories. Additional filtering criteria ensured consistency and linguistic relevance: posts had to be in English; each post had to contain an image; post titles were required to have at least four words. For each candidate subreddit, 20 randomly selected posts passing the above filters were

[1] https://github.com/ArthurHeitmann/arctic_shift

Table 1: Number of collected posts from subreddits associated with AI-generated images of *Art* and *Humor/Satire*. Parentheses indicate "flairs": user- or moderator-assigned tags.

Humor/Satire Subreddit	Posts
aigeneratedmemes	2,552
aimemes	599
hellaflyai	10,764
midjourney (Jokes/Meme – Midjourney AI)	1,281
StableDiffusion (Meme)	1,934
Art Subreddit	
midjourney (AI Showcase – Midjourney)	8,202
StableDiffusion (Workflow Not Included)	8,918

manually reviewed to verify content quality and relevance. Subreddits with frequent low-quality or off-topic content were excluded. Data collection was carried out in January 2025. Table 1 shows the number of collected image-post pairs collected from Reddit regarding *Humor/Satire* and *Art*. We were not able to identify any subreddits to consistently represent the *Misinformation* category.

3.1.2 Twitter data. To collect data from Twitter/X, we leveraged the publicly available Community Notes[2], a crowdsourced framework for assessing whether tweet content is misleading or satirical. Each note includes binary field indicators[3] for misinformation and satire. Notes also include a free-text summary where contributors justify their assessments.

For the *Misinformation* class, we selected tweets with at least one misinformation indicator (field) marked as true and both satire indicators were marked as false. For the *Humor/Satire* class, we included tweets with at least one satire-related indicator marked as true. In both cases, we further filtered for tweets whose summary text contained keywords related to generative artificial intelligence (AI, Artificial Intelligence, A.I.); ensuring that the image was likely AI-generated. For the *Art* category, we used the Art Community Generative AI [4], a Twitter/X community dedicated to AI-generated artistic content. To ensure relevance and engagement, we retained only posts with a minimum of five 'favourites_count', as a proxy for user approval. All posts were required to be in English and include at least one image. Data collection took place in January 2025. The amount of data collected from Twitter/X is summarized in the second column of Table 2

3.1.3 Evaluation Benchmark Summary. Our data collection process yielded numerous samples for the *Art* and *Humor/Satire* categories from Reddit, and *Art* from Twitter, whereas the *Misinformation* class was significantly underrepresented and sourced exclusively from Twitter. To address this class imbalance, we applied random under-sampling to Reddit-derived *Art* and *Humor/Satire*, and Twitter-derived *Art*, resulting in 3,192 samples per class, as summarized in Table 2. Due to its curated nature and limited size, this set is meant for use exclusively in evaluation.

Table 2: Number of "in the wild" posts (image–text pairs) from Reddit and Twitter after balancing.

Category	Twitter	Reddit	Total
Humor/Satire	436	2,756	3,192
Art	1,596	1,596	3,192
Misinformation	3,192	–	3,192

Table 3: Number of posts with real (non-synthetic) images from Reddit and Twitter after balancing.

Category	Twitter	Reddit	Total
Humor/Satire	3,789	25,385	29,174
Art	–	29,174	29,174
Misinformation	29,174	–	29,174

3.2 Training Set

Due to the limited availability of "in the wild" AI-generated images, particularly in the *Misinformation* category, we adopted a different strategy for constructing the training set. Unlike the evaluation benchmark, which relies on AI-generated images created and shared by actual users, the training set is based on real (non-AI-generated) images related to the three target categories, collected from Reddit and Twitter. These were then used to synthesize corresponding AI-generated samples using the Stable Diffusion XL[5] model.

3.2.1 Data collection. For Reddit, we selected subreddits related to *Humor/Satire* or *Art* whose users did not primarily post AI-generated content. Similarly, for Twitter, we filtered out posts whose Community Notes summaries mentioned generative Artificial Intelligence. Nevertheless, any accidentally collected AI-generated images would pose no issue here, as they are still reprocessed into synthetic training samples. We only included posts whose summaries referenced visual content (e.g., photo, screenshot) to ensure that most retained posts included an image. All other filtering criteria were consistent with the evaluation set. No suitable art-related community was identified for Twitter in this round of data collection and no misinformation-related subreddit was found in Reddit. Data collection was conducted in February of 2025. As in the case of the test set, the *Misinformation* category was under-represented. We applied random under-sampling to the Reddit-derived *Art* and *Humor/Satire* categories, resulting in 29,174 samples per class, as summarized in Table 3.

3.2.2 Generation Prompt Strategies. To convert real images into synthetic, we used Stable Diffusion XL to generate AI-synthesized images based on three approaches, as illustrated in Fig. 2:

- **Image-guided generation**: The original image was provided as input to the diffusion model (weight: 0.9), along

[2] https://communitynotes.x.com

[3] Misinformation fields: 'misleadingManipulatedMedia', 'misleadingFactualError', 'misleadingOutdatedInformation', 'misleadingMissingImportantContext', 'misleadingUnverifiedClaimAsFact', 'misleadingOther', Humor/Satire fields: 'misleadingSatire', 'notMisleadingClearlySatire'

[4] https://x.com/i/communities/1601841656147345410

[5] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

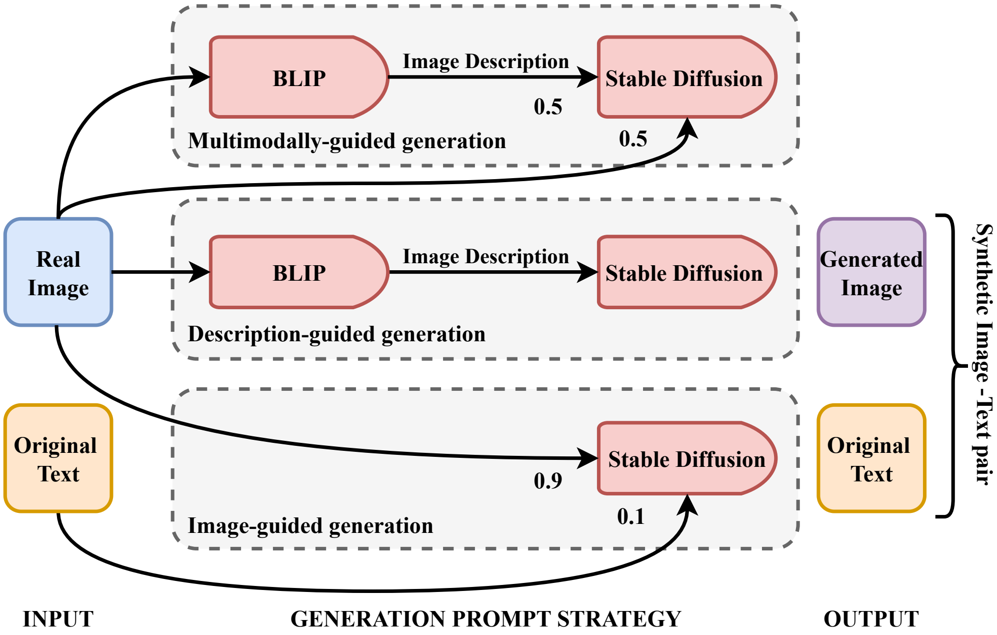


Figure 2: Three prompt strategies used with Stable Diffusion to generate synthetic images based on real images.

with its original caption (weight: 0.1), simulating image-to-image synthesis with low textual influence.

- **Description-guided generation**: We used BLIP [6] to caption each original image, then used that description as a text prompt for text-to-image generation.
- **Multimodally-guided generation**: Both the original image and the BLIP-generated caption were used together as inputs, each with equal weight (0.5), enabling balanced multimodal conditioning.

In all cases, the generated image is combined with the text of the initial post to form a training sample. This process resulted in three versions of the training dataset, each containing 87,522 generated image–text pairs balanced across the three classes. The dataset was split into training and validation with a ratio of 80%-20% resulting in 70,017 samples for training and 17,505 for validation, balanced across the three classes.

3.3 Dataset Exploration

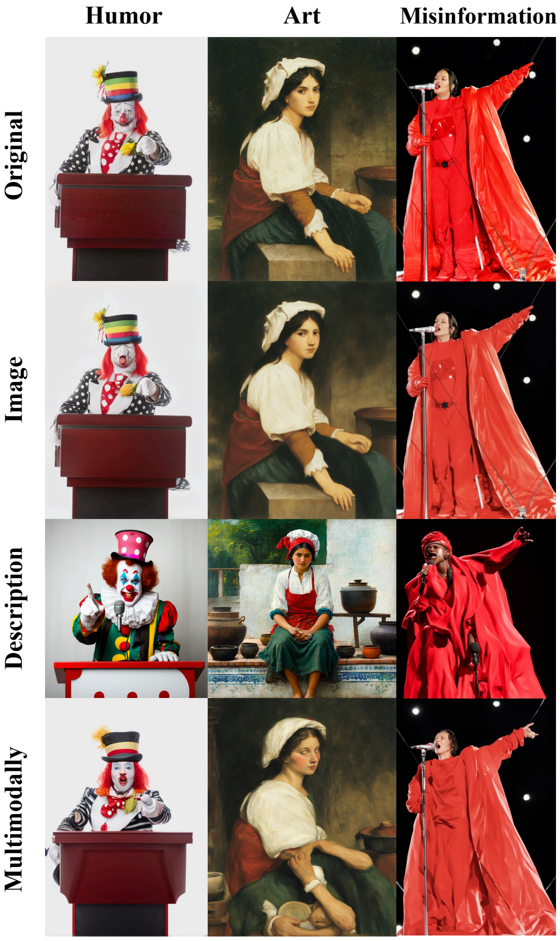
We employed the RINE synthetic image detector [10], updated with a DINOv2 backbone, which has shown strong performance on ‘in the wild’ AI-generated images [41] and flagged 82% of S-HArM “in the wild” samples as synthetic. This suggests that the vest majority of images in the test set are AI-generated and detectable as such. The remaining images—classified as real—may either originate from more advanced generation models that evade RINE’s detection, or be genuine real images that passed our filtering criteria. These findings support our hypothesis that, while general synthetic image detection has seen notable progress, there remains a need for intent-aware detection. Simply knowing an image is synthetic provides little insight into whether it was created to mislead, entertain, or express artistic intent.

Figure 3 illustrates examples from the training data generation pipeline. The top row displays original, real images collected from Reddit or Twitter. Each subsequent row presents AI-generated versions of the originals, using: Image-guided generation (row 2), Description-guided generation (row 3), and Multimodally-guided generation (row 4). The textual prompts used in the description- and multimodally-guided pipelines were generated using BLIP. The captions for the above examples were: (a) “Clown with red hair and a top hat standing behind a podium with a microphone in his hand and pointing at the camera with a finger up and a smile on his face, with a white background.” (b) “Painting of a woman sitting on a ledge with a pot in the background and a pot in the foreground, with a red apron on her head and a white cap on her headband.” (c) “Dressed in red singing into a microphone on stage with a red cape on it’s head and arms outstretched out to the side of the microphone, with a black background, and lights on a black background.”

We observe that Image-guided generation tends to preserve a high degree of similarity with the original image, primarily altering subtle details such as facial features, softening textures, reducing fine details (as in the art painting), or slightly shifting colors—introducing subtle artifacts characteristic of AI-generated images. In contrast, Description-guided generation produces the most divergent outputs, while still retaining key conceptual elements (e.g., a clown with red hair and a top hat, a woman sitting on a ledge, a singer wearing red clothes). These images often exhibit more visible artifacts commonly associated with AI synthesis, such as distorted fingers. Finally, Multimodally-guided generation strikes a balance between fidelity and alteration—maintaining some

[6] https://huggingface.co/Salesforce/blip-image-captioning-large

visual resemblance to the original while introducing more noticeable changes. In this setting, we also observe typical generative artifacts, particularly distorted hands, as seen in the middle image.



Picture of Donald Trump giving his first State of the Union address

Italian Girl at the Fountain

How will the masses know we're worshiping satan? [...] put a pentagram on your devil suit.

Figure 3: Examples of synthetic images from the three generation prompt strategies.

4 EXPERIMENTS

4.1 Experimental Settings

We employed CLIP ViT-L/14 (quickgelu)[7] [42] to extract both image I and text embeddings T. Given input vector $\mathbf{F} \in \mathbb{R}^{d_{\text{in}}}$, we define a classifier consisting of a three-layer Multi-Layer Perceptron (MLP) with GELU activations:

$$\mathbf{y} = \mathbf{W}_3 \cdot \text{GELU}\big(\mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \cdot \mathbf{F})\big) \quad (1)$$

Where $\mathbf{W}_1 \in \mathbb{R}^{512 \times d_{\text{in}}}$, $\mathbf{W}_2 \in \mathbb{R}^{128 \times 512}$, $\mathbf{W}_3 \in \mathbb{R}^{C \times 128}$ and $C = 3$ is the number of output classes. Bias terms were included in the network but omitted from the notation for brevity. The network is trained using the AdamW optimizer based on the Cross-Entropy Loss. Furthermore, we consider multiple modality fusion approaches and detection models to train and evaluate for the task of intent-aware classification:

4.1.1 Baselines. We use the MLP classifier and explore unimodal: image-only ($F = I$) and text-only ($F = T$) as well as multimodal experiments, with concatenation ($F = [I; T]$), addition $F = [I + T]$, subtraction $F = [I - T]$, or multiplication $F = [I * T]$ for modality fusion. We also consider the combination of the aforementioned fusion approaches, termed CASM for short, first introduced in [23], which can be expressed as: $\mathbf{F} = [(I; I + T; I - T; I * T; T)]$.

4.1.2 Contrastive Learning. We explore contrastive learning techniques, which have been widely used in multimodal tasks [9, 43], to the image representations using four different loss functions. The input embedding I was passed through separate projection networks, $\mathbf{z} = \mathbf{W}_{c2} \cdot \text{GELU}(\mathbf{W}_{c1} \cdot \mathbf{x})$, where $\mathbf{W}_{c1} \in \mathbb{R}^{512 \times d_{\text{in}}}$ and $\mathbf{W}_{c2} \in \mathbb{R}^{d_{in} \times 512}$, which are trained with either: Triplet Loss, Quadruplet Loss, InfoNCE Loss, or Supervised Contrastive Loss [44]. The resulting embeddings were concatenated with the text features and passed to the MLP classifier.

4.1.3 Reconstruction Networks. We investigate the use of Reconstruction Networks, which have been previously applied to tasks such as image super-resolution [45] and deepfake detection [46]. In our context, inspired by [47], we aim to leverage reconstruction as a means of restoring latent features of the original image used to generate the synthetic image. Given the image embeddings I of a synthetic image, we train a neural network to reconstruct the embeddings of the original image I^o—that is, the real image from which the synthetic version was derived. The model is trained by minimizing the Mean Squared Error (MSE) between the reconstructed embedding I^r and the target embedding I^o. We explore several architectural variations:

- **Shared Reconstructor (Replace):** A single reconstructor is trained across all categories. The reconstructed embedding I^r replaces I, and the classifier receives input $F = [I^r; T]$.
- **Shared Reconstructor (Combine):** A single reconstructor is trained across all categories. Both I and I^r are combined with T, forming $F = [I; I^r; T]$, which is then passed to the classifier.
- **Class-Specific Reconstructors (Replace):** Separate reconstructors R_h, R_a, and R_m are trained for the humor, art, and misinformation classes, respectively. Only reconstructed embeddings are used, yielding $F = [I_h^r; I_a^r; I_m^r; T]$, as classifier input.
- **Class-Specific Reconstructors (Combine):** Same as above, but I is also included in the classifier: $F = [I; I_h^r; I_a^r; I_m^r; T]$.

4.1.4 Attention Mechanisms. We investigate the use of self-attention mechanisms to capture interactions between image and text modalities. In addition, we adapt RED-DOT [23], a state-of-the-art Transformer model for multimodal misinformation detection. For our purposes, we exclude the Relevant Evidence Detection (RED) module, as it is not applicable in this setting.

[7] https://github.com/mlfoundations/open_clip

4.1.5 Implementation Details. To ensure consistency and reliable comparisons, each trainable model was run 10 times using different random seeds, and the results were averaged. All experiments used a learning rate of 5×10^{-5}, weight decay of 1×10^{-4}, and a batch size of 32. Each model was trained for up to 30 epochs.

4.1.6 Large Vision-Language Models. We explore the use of LVLMs for zero-shot intent-aware classification of image-text pairs. Specifically, we use Llama-3.2-11B[8]. Given the image-text pair as input, we use the “Direct Classification” prompt:

> **“Direct Classification” prompt**
>
> *“You are given a title and an AI generated image. Your task is to classify the pair into one of the following categories: ‘misinformation’, ‘satire’, or ‘art’. Respond with only the category name, no punctuation, no explanation, and no additional text.”*

Due to high confusion between satire and misinformation, we also added the *“Between ‘misinformation’ and ‘satire’ choose ‘misinformation”* sentence; denoted as “Direct Classification (Nudged)”

Furthermore, we explore a “Two-Stage Prompting” approach, in which image analysis is performed before classification. In the first stage, the image is provided as input along with the following prompt:

> **First-stage Prompt (Description)**
>
> *“You are given an AI-generated image. Please describe in detail what the image contains. Describe the emotional tone conveyed. Respond with a detailed analysis, including what stands out and why.”*

In Stage 2, the image, original caption, and the description from Stage 1 are passed together with the following prompt:

> **Second-stage Prompt (Detection)**
>
> *“You are given a title, an AI-generated image, and a detailed analysis of the image. Your task is to classify the image-text pair into one of the following categories: ‘misinformation’, ‘satire’, or ‘art’. Between ‘misinformation’ and ‘satire’ choose ‘misinformation’. Respond with only the category name, no punctuation, no explanation, and no additional text.”*

4.2 Experimental Results

4.2.1 Comparative study across generation prompt strategies. Table 4 presents the performance of all methods trained on the synthetic *S-HArM* training set and evaluated on the “in the wild” benchmark. We observe that the highest overall performance is achieved by the MLP classifier with concatenated image and text features ($F = [I; T]$), reaching 71.6% accuracy with a standard deviation of 0.54 when trained on image-guided data. Among contrastive methods, pre-training with supervised Contrastive Loss reached the second best performance (71.5%) when trained on multimodally-guided data. Meanwhile, among reconstruction-based approaches, the Class-Specific Reconstructors (Combine) performed well when trained on image-guided data, with an average accuracy of 71.16%, ranking fourth across all methods. We also observe that attention-based models underperformed compared to simpler architectures. Self-attention peaked at 70.05% on image-guided data, and RED-DOT reached 70.16% with multimodally-guided data; trailing behind the “MLP (CASM)” baseline without any attention mechanism.

We observe that the three prompting strategies–image-guided, description-guided, and multimodally-guided–favor different model architectures, suggesting that each introduces distinct patterns or artifacts that certain models are better equipped to exploit. As shown in the last row of Table 4, the description-guided approach consistently underperforms, likely due to the absence of direct visual input during generation, resulting in missing visual cues essential for classification. In contrast, the image- and multimodally-guided methods retain these visual features, supporting stronger generalization across models.

4.2.2 Performance of Large Language-Vision Models. As shown in Table 5, LVLMs underperform compared to task-specific models. This is expected, as LVLMs operate in a zero-shot setting without fine-tuning for the classification task. The baseline Direct Classification prompt yields limited accuracy (50.09%). However, the ‘Nudged’ prompt–designed to bias the model toward selecting misinformation when uncertain between satire and misinformation–leads to a substantial improvement, reaching 62.28%. Moreover, the two-stage prompting strategy, which first elicits a description of the image before classification, further increases accuracy to 66.65%. While LVLMs still fall short of fully trained models, these findings indicate promising directions for improvement, such as few-shot prompting, more advanced models, and the integration of chain-of-thought reasoning. This suggests that with appropriate prompting strategies, LVLMs may become viable components in future intent-aware and multimodal misinformation detection systems.

4.2.3 Model Generalization. Table 6 reports per-class accuracy for the best-performing method, MLP (concatenation) trained on image-guided data, across the validation set and “in the wild” data from Twitter and Reddit. First, we observe that accuracy on the validation set is notably high, averaging 96.63% across the three categories. Nevertheless, this this performance does not directly translate to “in the wild” synthetic images. This gap highlights the distributional shift between our synthetic training data and “in-the-wild” AI-generated content, which is more diverse, produced by a range of generative models, and curated by users for quality.

Performance on the real-world benchmark is especially low for satire/humor content from Twitter (13.56%), which corresponds with the under-representation of this class–source pair in the dataset, as shown in Tables 2 and 3. The resulting data imbalance likely hampers the model’s ability to learn meaningful patterns. In contrast, the *Art* and *Misinformation* classes achieve higher accuracy on Twitter (72.71% and 85.59%, respectively), likely due to the longer average text length of Twitter posts compared to Reddit, offering richer textual context and aiding classification.

4.2.4 Unimodal Analysis. As shown in Table 4, text-only models consistently outperform image-only models across all three generation procedures. This suggests that text provides more informative signals for classification, likely because synthetic images–regardless

[8] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct

Table 4: Comparison of detection models trained on datasets from three generation prompt strategies. We report the average accuracy and standard deviation over 10 random seeds. Bold indicates the best overall performance; underlining highlights the best result per generation prompt strategy.

Method / Variation	Accuracy % (Std. Dev. %)		
	Image-guided	Description-guided	Multimodally-guided
MLP (Image only)	62.94 (1.33)	61.62 (1.02)	63.80 (0.68)
MLP (Text only)	65.43 (0.45)	65.43 (0.45)	65.43 (0.45)
MLP (Concatenation)	**71.60** (0.54)	70.05 (0.53)	70.32 (0.39)
MLP (Addition)	69.61 (0.62)	68.37 (0.70)	68.10 (0.97)
MLP (Subtraction)	65.47 (0.46)	62.29 (0.92)	65.02 (0.69)
MLP (Multiplication)	59.69 (1.04)	57.13 (0.72)	59.26 (1.02)
MLP (CASM)	71.37 (0.89)	70.05 (0.71)	70.28 (0.53)
Contrastive Triplet Loss	69.51 (0.96)	69.94 (0.59)	70.61 (0.86)
Contrastive Quadruplet Loss	68.67 (0.68)	69.43 (0.49)	69.87 (0.60)
Contrastive InfoNCE	70.88 (0.53)	68.79 (0.88)	70.39 (0.45)
Supervised Contrastive	70.03 (0.65)	70.13 (0.49)	71.50 (0.57)
Shared Reconstructor (Replace)	70.54 (0.80)	70.77 (0.58)	71.14 (0.64)
Shared Reconstructor (Combine)	70.75 (0.79)	69.94 (0.81)	69.89 (1.07)
Class-specific Reconstructors (Replace)	69.64 (1.96)	69.22 (1.41)	68.92 (1.42)
Class-specific Reconstructors (Combine)	71.16 (0.76)	69.79 (1.17)	70.23 (1.44)
Self-Attention	70.05 (0.31)	69.37 (0.23)	68.90 (0.32)
RED-DOT	70.05 (1.04)	69.25 (0.93)	70.16 (0.90)
Average Accuracy	68.67 (0.81)	67.74 (0.74)	68.46 (0.73)

Table 5: Zero-shot performance of LVLM-based classification.

Prompt	Accuracy
Direct Classification	50.09
Direct Classification (Nudged)	62.28
Two-Stage (Nudged)	66.65

Table 6: Per-class accuracy of MLP (Concatenation) on validation and 'in the wild' Twitter and Reddit samples.

Category	Validation set	Twitter	Reddit
Humor/Satire	94.61	13.56	69.59
Art	99.53	72.71	61.36
Misinformation	95.75	85.59	-
Balanced Accuracy	96.63	75.49	66.57

of their intended purpose–are often visually similar. In contrast, captions offer stronger cues about the creator's intent. Unsurprisingly, the text-only results remain identical across generation types, since all methods retain the original, unmodified captions. Nevertheless, models trained on both modalities consistently outperform unimodal counterparts, highlighting the importance of cross-modal interactions for accurate classification; validating our hypothesis that intent-aware classification of synthetic images should be addressed as a multimodal task.

5 CONCLUSION

In this work, we introduced *S-HArM*, a novel dataset for intent-aware classification of synthetic multimodal content. Unlike previous efforts that focus on detecting synthetic or decontextualized images, *S-HArM* targets a more nuanced challenge: understanding the intent behind AI-generated images by categorizing them as humor/satire, art, or misinformation. *S-HArM* includes a real-world "in-the-wild" evaluation benchmark sourced from Twitter and Reddit, alongside three synthetically generated training sets produced using Stable Diffusion through image-guided, description-guided, and multimodally-guided generation strategies.

We conducted a comparative study on a wide range of models, including various modality fusion techniques, contrastive learning, reconstruction networks, attention mechanisms and LVLMs. These methods achieved limited model generalization, highlighting the difficulty of inferring intent—a high-level, abstract concept—from visual and textual cues alone. Our findings indicate that current architectures are ill-suited for intent-aware classification, highlighting the need for models that can reason about intent, contextual framing, and social subtext–alongside more diverse and representative training data. We hope that *S-HArM* will serve as a foundation for advancing research in this direction.

ACKNOWLEDGMENTS

This work is partially funded by the Horizon Europe projects "DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies" under grant agreement no. 101079164, "vera.ai - VERification Assisted by Artificial Intelligence" under grant agreement no. 101070093, and "AI-CODE - AI services for COntinuous trust in emerging Digital Environments" under grant agreement no. 101135437. The authors also acknowledge the support and computational resources provided by the IT Center of the Aristotle University of Thessaloniki (AUTh) throughout the progress of this research work.

REFERENCES

[1] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–49, 2021.

[2] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[3] S. Karnouskos, "Artificial intelligence in digital media: The era of deepfakes," *IEEE Transactions on Technology and Society*, vol. 1, no. 3, pp. 138–147, 2020.

[4] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," 2020. [Online]. Available: https://arxiv.org/abs/1912.11035

[5] S. Sinitsa and O. Fried, "Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis," 2024. [Online]. Available: https://arxiv.org/abs/2303.10762

[6] Q. Bammey, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 1–9, 2024.

[7] Y. Li, Q. Bammey, M. Gardella, T. Nikoukhah, J.-M. Morel, M. Colom, and R. G. Von Gioi, "Masksim: Detection of synthetic images by masked spectrum similarity analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024, pp. 3855–3865.

[8] D. Konstantinidou, C. Koutlis, and S. Papadopoulos, "Texturecrop: Enhancing synthetic image detection through texture-based cropping," 2025. [Online]. Available: https://arxiv.org/abs/2407.15500

[9] H. Wu, J. Zhou, and S. Zhang, "Generalizable synthetic image detection via language-guided contrastive learning," 2025. [Online]. Available: https://arxiv.org/abs/2305.13800

[10] C. Koutlis and S. Papadopoulos, "Leveraging representations from intermediate encoder-blocks for synthetic image detection," 2024. [Online]. Available: https://arxiv.org/abs/2402.19091

[11] D. Karageorgiou, S. Papadopoulos, I. Kompatsiaris, and E. Gavves, "Any-resolution ai-generated image detection by spectral learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 706–18 717.

[12] C. Boididou, S. Papadopoulos, L. Apostolidis, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2015," in *Working Notes Proceedings of the MediaEval Workshop*, 2015.

[13] C. Boididou, S. E. Middleton, Z. Jin *et al.*, "Verifying information with multimedia content on twitter," *Multimedia Tools and Applications*, vol. 77, pp. 15 545–15 571, 2018. [Online]. Available: https://doi.org/10.1007/s11042-017-5132-9

[14] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[15] K. Nakamura, S. Levy, and W. Y. Wang, "r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," 2020. [Online]. Available: https://arxiv.org/abs/1911.03854

[16] S. Aneja, C. Bregler, and M. Nießner, "Cosmos: catching out-of-context image misuse using self-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 12, 2023, pp. 14 084–14 092.

[17] G. Luo, T. Darrell, and A. Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," 2021. [Online]. Available: https://arxiv.org/abs/2104.05893

[18] E. Sabir, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Deep multimodal image-repurposing detection," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1337–1345.

[19] M. Mu, S. Das Bhattacharjee, and J. Yuan, "Self-supervised distilled learning for multi-modal misinformation identification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2819–2828.

[20] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. Petrantonakis, "Synthetic misinformers: Generating and combating multimodal misinformation," in *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, 2023, pp. 36–44.

[21] S. Abdelnabi, R. Hasan, and M. Fritz, "Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 940–14 949.

[22] X. Yuan, J. Guo, W. Qiu, Z. Huang, and S. Li, "Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4268–4280.

[23] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Red-dot: Multimodal fact-checking via relevant evidence detection," 2024. [Online]. Available: https://arxiv.org/abs/2311.09939

[24] F. Zhang, J. Liu, Q. Zhang, E. Sun, J. Xie, and Z.-J. Zha, "Ecenet: explainable and context-enhanced network for muti-modal fact verification," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1231–1240.

[25] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, "Sniffer: Multimodal large language model for explainable out-of-context misinformation detection," 2024. [Online]. Available: https://arxiv.org/abs/2403.03170

[26] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection?" in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, February 2025, pp. 5570–5579.

[27] D. Xu, S. Fan, and M. Kankanhalli, "Combating misinformation in the era of generative ai models," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 9291–9298. [Online]. Available: https://doi.org/10.1145/3581783.3612704

[28] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314222001114

[29] S. Muhammed T and S. K. Mathew, "The disaster of misinformation: a review of research in social media," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 271–285, May 2022. [Online]. Available: https://doi.org/10.1007/s41060-022-00311-6

[30] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, Feb. 2023. [Online]. Available: https://doi.org/10.1007/s13278-023-01028-5

[31] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visual news: Benchmark and challenges in news image captioning," 2021. [Online]. Available: https://arxiv.org/abs/2010.03743

[32] S. I. Papadopoulos, C. Koutlis, S. Papadopoulos *et al.*, "VERITE: A robust benchmark for multimodal misinformation detection accounting for unimodal bias," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 4, 2024. [Online]. Available: https://doi.org/10.1007/s13735-023-00312-6

[33] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, "Online detection of ai-generated images," 2023. [Online]. Available: https://arxiv.org/abs/2310.15150

[34] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, and S. A. Fattah, "Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2200–2204.

[35] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," 2022. [Online]. Available: https://arxiv.org/abs/2211.00680

[36] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

[37] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "Raise: a raw images dataset for digital image forensics," in *Proceedings of the 6th ACM Multimedia Systems Conference*, ser. MMSys '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 219–224. [Online]. Available: https://doi.org/10.1145/2713168.2713194

[38] T. Gloe and R. Böhme, "The 'dresden image database' for benchmarking digital image forensics," in *Proceedings of the 2010 ACM Symposium on Applied Computing*, ser. SAC '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1584–1590. [Online]. Available: https://doi.org/10.1145/1774088.1774427

[39] J. J. Bird and A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15 642–15 650, 2024.

[40] Y. Chen and J. Zou, "Twigma: A dataset of ai-generated images with metadata from twitter," 2023. [Online]. Available: https://arxiv.org/abs/2306.08310

[41] D. Konstantinidou, D. Karageorgiou, C. Koutlis, O. Papadopoulou, E. Schinas, and S. Papadopoulos, "Navigating the challenges of ai-generated image detection in the wild: What truly matters?" 2025. [Online]. Available: https://arxiv.org/abs/2507.10236

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[43] ——, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[44] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2021. [Online]. Available: https://arxiv.org/abs/2004.11362

[45] H. Yan, Z. Wang, Z. Xu, Z. Wang, Z. Wu, and R. Lyu, "Research on image super-resolution reconstruction mechanism based on convolutional neural network," 2024. [Online]. Available: https://arxiv.org/abs/2407.13211

[46] Z. He, W. Wang, W. Guan, J. Dong, and T. Tan, "Defeating deepfakes via adversarial visual reconstruction," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2464–2472. [Online]. Available: https://doi.org/10.1145/3503161.3547923

[47] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Latent multimodal reconstruction for misinformation detection," *arXiv preprint arXiv:2504.06010*, 2025.