ADRIAN InstructFace-Edit - Towards Robust Detection of AI-Manipulated Face Images

Viktoriia Vovchenko Vincenzo Barberi Sergej Schultenkämper and Frederik Simon Bäumer
Bielefeld University of Applied Sciences and Arts
Applied AI Working Group
Interaktion 1, 33619 Bielefeld, Germany
{viktoriia.vovchenko| vincenzo.barberi| sergej.schultenkaemper| frederik.baeumer}@hsbi.de

***Abstract*—Instruction-guided image editing models have made photorealistic facial manipulation accessible through natural language prompts, substantially lowering the barrier to misuse in the cybersecurity landscape. These systems facilitate attacks such as identity theft, phishing, misinformation, and social engineering, while also challenging existing forensic methods that were not designed for prompt-driven edits. However, current facial forensics datasets do not cover this manipulation paradigm. We present ADRIAN InstructFace-Edit, a dataset of 36,000 edited facial image pairs at 1024×1024 resolution generated by four state-of-the-art instruction-guided image editing models (FLUX.1-Kontext-dev, FLUX.2-dev, Qwen-Image-Edit-2511, and Step1X-Edit-v1.2) across six semantically distinct edit types, of which 18,768 are annotated as high quality. The full dataset, including all pairs with quality labels, is made publicly available. To improve data quality, we design a multi-stage filtering pipeline that integrates structural, semantic, and multimodal large language model–based quality checks, with thresholds derived empirically from a human-annotated subset of 1,500 images. Because open-weight vision-language models perform poorly on VIEScore evaluation out of the box, we fine-tune a LoRA adapter for Qwen3-VL-32B-Instruct that improves within-± 1 Semantic Consistency accuracy from near-random to 73.2%. We will publicly release the dataset, generation pipeline, and LoRA adapter to support future research on training, evaluating, and stress-testing detectors for instruction-guided facial manipulation.**

***Keywords*—Deepfakes, Image forensics, Diffusion models, Computer vision**

I. Introduction

The rapid evolution of Generative Artificial Intelligence (GenAI) has fundamentally transformed the landscape of digital media creation and manipulation. While these advancements offer significant creative potential [1], they simultaneously lower the barrier for creating highly realistic, manipulated facial images and deepfakes [2], [3]. Notably, recent instruction-guided image editing models [4], [5] greatly simplify image editing via natural language prompts, effectively removing the need for complex manual workflows required by traditional techniques. This introduces critical risks to digital security, facilitating identity theft, financial fraud, phishing attacks, and severe violations of personal privacy through non-consensual synthetic media.

The increasing realism of manipulated imagery also poses threats to political stability by enabling the rapid spread of misinformation and decreasing trust in digital content [6]–[8]. Consequently, there is a growing need for robust fake image detectors capable of identifying manipulated facial images.

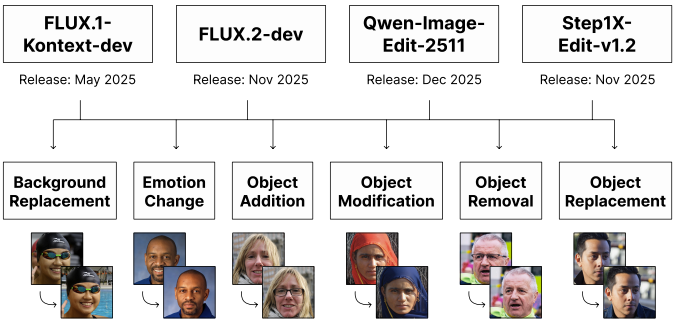

Fig. 1. Overview of the dataset generation framework. Four state-of-the-art models (top) perform six categories of facial manipulations (middle), with representative before-and-after examples shown below (bottom).

A persistent challenge in the field of fake image detection is the "generalization gap" between the release of new generative models and the development of corresponding detection mechanisms [8], [9]. While new models surpass prior standards for photorealism and anatomical consistency, they are often absent from the public datasets used to train detectors. Consequently, existing detectors fail to generalize to novel architectures, resulting in significantly reduced performance on unseen models or manipulation techniques [10]. Crucially, facial images manipulated via instruction-guided image editing models are frequently missing from these current training corpora.

To address these gaps, we propose **ADRIAN InstructFace-Edit (AIFE)**, a dataset comprising 36,000 semantically edited image pairs (original + edited images) at 1024×1024 resolution. This dataset incorporates images generated by state-of-the-art (SOTA) instruction-guided image editing models [5], [11]–[13], which, to the best of our knowledge, have not yet been used in other datasets (see Table I). The creation of our dataset is part of the Authority-Dependent Risk Identification and Analysis in Online Networks (ADRIAN) research project [14], which is dedicated to exploring and developing machine-learning-based methods for detecting potential threats to individuals based on online datasets and Digital Twins (DTs). In this context, detecting these manipulated images is critical,

since facial forgery is actively used for malicious attacks and online exploitation. To improve quality, our dataset underwent a multi-stage filtering process using a combination of automated metrics. We will publicly release the complete dataset alongside our generation and evaluation pipeline. In future work, we plan to further analyze the unique challenges of detecting fake facial images manipulated by instruction-guided image editing models by training and evaluating detectors on our proposed dataset.

The remainder of this paper is organized as follows. Section II reviews related work in instruction-guided image editing, automatic evaluation metrics, and existing facial forensics datasets. In Section III, we detail our comprehensive dataset construction pipeline, covering initial image acquisition, prompt generation, model-based editing, threshold calibration, and VIEScore adaptation via LoRA fine-tuning. Section III-D presents a statistical overview of the resulting dataset. Finally, Section IV discusses practical lessons learned during the large-scale dataset construction process, before Section V concludes the paper.

II. Related Work

A. Instruction-guided Image Editing

Recent advancements in generative AI have revolutionized text-to-image synthesis, with diffusion models [15]–[18] producing high-fidelity images from textual descriptions. Despite their strength in image generation, they often struggle with context-preserving edits. To address this, instruction-guided image editing emerged, allowing users to modify existing images through intuitive text prompts.

Early methods, such as Prompt-to-Prompt [19], injected cross-attention maps during denoising, while Imagic [20] optimized text embeddings for real images and partially fine-tuned diffusion models. InstructPix2Pix [4] marked a major paradigm shift by explicitly training a conditional diffusion model on large-scale synthetic instruction-edit datasets generated with Prompt-to-Prompt in combination with Stable Diffusion [16] and OpenAI GPT-3 [21]. Subsequent research addressed early limitations by improving training data quality and better aligning models with human preferences [22]–[25]. Others integrated Multimodal Large Language Models (MLLMs) for better spatial reasoning and complex instruction comprehension [26], [27], or developed unified multi-task architectures using MLLMs, Diffusion Transformers (DiTs) [28], and Mixture-of-Experts (MoE) architectures to improve generation versatility and fidelity [13], [29]–[33].

Current SOTA models unify generation, editing, and multi-image reasoning into general multimodal foundation models. Leading open-source models include Qwen-Image-Edit [12], which utilizes a dual-encoding mechanism for semantic and reconstructive features and a Multimodal Diffusion Transformer (MMDiT) backbone, as well as the FLUX series (FLUX.1 Kontext [5] and FLUX.2 [11]), which leverages generative flow matching models for high-resolution, multi-reference in-context editing. However, proprietary systems like GPT Image [34], [35] and the Gemini Nano Banana series [36]–[38] consistently dominate image editing leaderboards [39], [40]. By natively embedding image generation and editing into MLLMs, they enable conversational, multi-turn workflows with superior instruction adherence, context preservation, and inference speed as well as richer multi-image context and world knowledge use. Despite these advantages, their closed-source nature presents barriers for the research community, as their underlying architectures remain opaque and large-scale inference incurs substantial financial cost.

B. Evaluation metrics

Various evaluation metrics have been proposed to automatically evaluate instruction-guided image editing models. Unlike traditional text-to-image generation, this task introduces a trade-off between perceptual quality, semantic alignment, and content preservation.

To measure perceptual quality, Fréchet Inception Distance (FID) [41] computes the Fréchet distance between Inception-v3 feature distributions of real and generated image sets. While widely used, FID is sensitive to preprocessing and sample size. Kernel Inception Distance (KID) [42] provides an unbiased alternative by utilizing squared Maximum Mean Discrepancy (MMD), making it better suited for smaller datasets. CLIP-MMD (CMMD) [43] replaces Inception features with Contrastive Language-Image Pre-training (CLIP) [44] embeddings to compute MMD. Benefiting from CLIP's extensive training data, CMMD better represents modern generative models and correlates more closely with human judgments. For evaluating content preservation, the Structural Similarity Index (SSIM) [45] quantifies the similarity between an edited image and its reference using local luminance, contrast, and structural statistics. Learned Perceptual Image Patch Similarity (LPIPS) [46] improves upon this by measuring perceptual similarity through deep network feature activations, which aligns better with human perception. To evaluate semantic preservation, CLIP-I computes the cosine similarity between source and edited image embeddings. However, since both successful and unsuccessful edits alter semantics, high scores can falsely indicate a failed edit. Additionally, distillation with no labels (DINO)-based metrics [47]–[50] compute the cosine similarity of embeddings extracted from self-supervised Vision Transformers (ViT) to measure semantic consistency and identity preservation. To evaluate semantic alignment, CLIPScore [51] computes the cosine similarity between the edited image and the text caption using a pretrained CLIP model. Its utility for editing tasks is limited because it ignores the source image and does not account for content preservation. CLIP Directional Similarity [52] addresses this limitation by measuring whether the visual change between the source and edited image aligns with the semantic change between their respective text descriptions.

Recent approaches leverage MLLMs for evaluation. Visual Instruction-guided Explainable Score (VIEScore) [53] uses MLLMs to evaluate Semantic Consistency (SC) and Perceptual Quality (PQ) on a 0 to 10 scale, which are then aggregated into an overall score (O) alongside a generated natural

language rationale to explain its reasoning. SC measures alignment with text instructions, while PQ evaluates visual naturalness by penalizing artifacts or distortions. Although VIEScore provides explainable evaluations that correlate strongly with human judgments, it struggles with fine-grained modifications and performs significantly better with proprietary, closed-source models such as OpenAI GPT-4V [54].

C. Existing datasets

Robust fake image detection relies on the quality, scale, and diversity of the underlying training datasets. Early research predominantly utilized datasets, such as FaceForensics++ [55], DFFD [56], DFDC [57], Celeb-DF [58], and ForgeryNet [59]. However, they are limited by their reliance on older manipulation techniques using Autoencoders (AE) [60] and Generative Adversarial Networks (GANs) [61]. Consequently, these generated images often contain perceptible artifacts that fail to represent the photorealism of modern techniques. This causes detectors trained on them to generalize poorly to current high-fidelity generated imagery.

Due to the introduction of diffusion models, a newer generation of facial forgery datasets has emerged. Datasets, such as DiFF [62], DiffusionFace [63], DF40 [64], and MFFI [65], broadly cover different diffusion-based forgery techniques, including text-to-image synthesis, inpainting, and diffusion-based face swapping. However, these datasets typically manipulate facial images via explicit spatial masks (e.g., masked inpainting) or image-to-image translations rather than utilizing free-form instruction-guided image editing. Furthermore, while general synthetic image datasets like DiffusionForensics [66], DE-FAKE [67], CIFAKE [68], DiffusionDB [69], GenImage [70], and Community Forensics [71] study broader generalization across diverse scenes, they are not tailored for specialized facial forensics.

To address this gap, we propose a dataset focused on the distinct detection challenges introduced by instruction-guided image editing. Sourced from the Flickr-Faces-HQ (FFHQ) dataset [72], our collection contains 36,000 facial images manipulated exclusively with SOTA instruction-guided models, such as FLUX.1 Kontext [5] and Qwen-Image-Edit [12]. The FFHQ dataset served as the data source because it provides high-resolution facial images and offers broad diversity in terms of age, ethnicity, and backgrounds. Unlike existing datasets, which depend on additional constraints, such as masks or total-image synthesis, these instruction-guided models alter semantic attributes based solely on intuitive text instructions. This makes our dataset a practical resource for researching detection strategies and vulnerabilities associated with modern manipulation techniques.

III. Approach

This section describes our approach for constructing the dataset. The process consists of image acquisition, instruction-guided image editing using multiple generative models, and a sequential multi-stage quality filtering process. Fig. 2 provides an overview of the full pipeline.

TABLE I
COMPARISON WITH EXISTING FACIAL FORENSICS AND FAKE IMAGE DATASETS. (IGIE: INSTRUCTION-GUIDED IMAGE EDITING USED). *"UNKNOWN (-)" INDICATES THAT THE MOST RECENT METHOD USED IN THE DATASET, OR ITS EXACT PUBLICATION YEAR, COULD NOT BE DETERMINED.

Dataset	Newest Methods*	Type	IGIE
FaceForensics++ [55]	NeuralTextures (2019)	Face	×
DFFD [56]	StyleGAN (2018)	Face	×
DFDC [57]	StyleGAN (2018)	Face	×
Celeb-DF [58]	Unknown (-)	Face	×
ForgeryNet [59]	StarGANv2 (2020)	Face	×
DiffusionDB [69]	Stable Diffusion (2022)	General	×
DE-FAKE [67]	Stable Diffusion (2022)	General	×
GenImage [70]	Stable Diffusion (2022)	General	×
DiffusionForensics [66]	Stable Diffusion (2022)	General	×
CIFAKE [68]	Stable Diffusion (2022)	General	×
Community Forensics [71]	Unknown (-)	General	×
DiFF [62]	SDXL (2023)	Face	×
DiffusionFace [63]	DiffSwap (2023)	Face	×
DF40 [64]	PixArt-α (2024)	Face	×
MFFI [65]	Jimeng3.0 (2025)	Face	×
Ours	**Qwen-Image-Edit-2511 (2025)**	**Face**	✓

A. Dataset Construction Pipeline

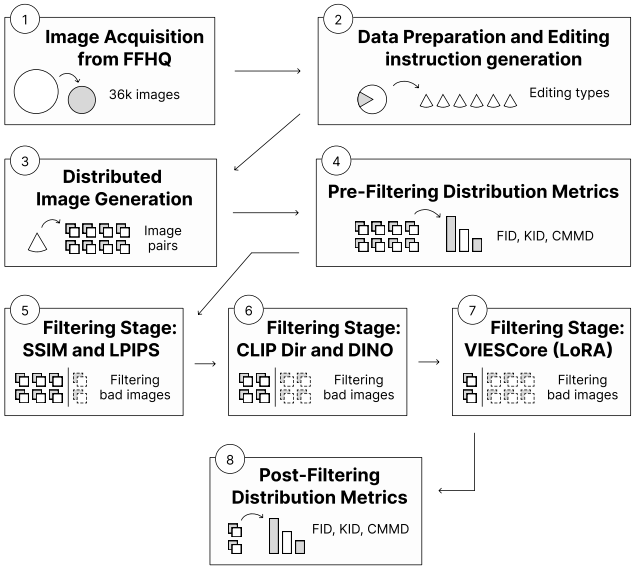

Fig. 2. Overview of our dataset construction pipeline.

Image Acquisition from FFHQ: A subset of 36,000 images is downloaded from the FFHQ dataset [72] and stored as the reference ("real") image in each edited pair.

Data Preparation and Editing Instruction Generation: Six types of edits covering semantically distinct manipulation categories were defined: *Background Replacement, Emotion Change*, *Object Addition*, *Object Modification*, *Object Removal*, and *Object Replacement*. We randomly sampled 9,000 source images from the initial subset for each of the four image editing models used and distributed them evenly across all edit types. This resulted in 1,500 images per edit type per model.

Then, using the GPT-4.1 vision-language model (VLM) via OpenAI's Batch Application Programming Interface (API), each image was annotated with a caption for the original image (`real_caption`), the editing prompt

(`editing_prompt`), a caption for the edited image (`edit_caption`), and a list of strings representing the image elements affected by the edit (`edited_objects`). An example of the metadata for a single image, including these and subsequently added annotations, is shown in Fig. 3. By employing a step-by-step approach - first analyzing the key elements of the original image to generate the initial caption, and subsequently generating the editing prompt and edited caption - we significantly reduced hallucinations and yielded more accurate annotations. Additionally, to optimize editing prompt quality, we experimented with various prompt lengths, levels of detail and specificity. Alongside the findings from our own empirical testing, we also incorporated the best practices outlined in the FLUX.1 prompting guide [73] into our system prompt. The resulting prompt structure yielded more effective editing instructions that significantly enhanced instruction adherence and character preservation across multiple editing models.


```
{
    "id": "img_00001",
    "real_image_path": "...",
    "edit_image_path": "...",
    "model": "FLUX.2-dev-bnb-4bit",
    "edit_type": "Object Modification",
    "edited_objects": ["earrings"],
    "real_caption": "A young woman with straight blonde
    hair and light skin is smiling widely, showing her
    teeth. She has no visible makeup, and is wearing a
    turquoise top and large turquoise hoop earrings. The
    background is softly lit and out of focus.",
    "editing_prompt": "Change the color of the woman's
    earrings to bright red while preserving the woman's
    exact same facial features, hairstyle, and
    expression. Keep the woman in the exact same
    position, scale, and pose. Maintain identical subject
    placement, camera angle, framing, and perspective.",
    "edit_caption": "A young woman with straight blonde
    hair and light skin is smiling widely, showing her
    teeth. She has no visible makeup, and is wearing a
    turquoise top and large bright red hoop earrings. The
    background is softly lit and out of focus.",
    "editing_parameters": {
        "guidance_scale": 1,
        "num_inference_steps": 28,
        "seed": 6485951459952579169
    },
    "ffhq_metadata": { ... },
    "ssim": 0.9586328897993089,
    "lpips": 0.030474014580249786,
    "clip_score": 0.32582157850265503,
    "clip_dir": 0.31859612464904785,
    "dino": 0.9854938983917236,
    "viescore_qwen3vl_lora_v2": 9.0,
    "viescore_qwen3vl_lora_v2_detail": {
        "semantic_consistency": 10.0,
        "perceptual_quality": 8.0,
        ...
    }
}
```


Fig. 3. Example JSON metadata record for a single image editing task, detailing the original and edited captions, model parameters, and automated evaluation metrics.

Distributed Image Generation: For each metadata entry, the pipeline invokes a model-specific processor script that loads the model weights and generates the edited image from the source image and its editing prompt. After generation, the random seed used is written back to the metadata file (see example in Fig. 3). This stage produces the full, unfiltered set of (original, edited) image pairs generated by the following open-source image editing models: FLUX.1-Kontext-dev [73], FLUX.2-dev [11], Qwen-Image-Edit-2511 [12], and Step1X-Edit-v1.2 [13], [74]. Our model selection was driven by a balance between cost efficiency – as all selected models are open-source – and output quality, representing the best available models at the time of generation.

Pre-Filtering Distribution Metrics: Before any per-image filtering is applied, FID [41], KID [42], and CMMD [43] are computed over the complete raw generated set as a baseline. Because these are distributional metrics over the full set, they cannot drive per-image filtering decisions; they serve to quantify the quality improvement achieved by the subsequent stages.

Filtering Stage 1: SSIM and LPIPS: The first per-image filtering pass uses SSIM [45] and LPIPS [46], selected as the opening stage due to their low computational cost at scale. SSIM thresholds (> 0.97 and < 0.37) discard pairs where no edit was applied and where severe structural distortion occurred. LPIPS thresholds (< 0.011 and > 0.4) remove perceptually unchanged pairs and those with excessive alteration; the upper bound is waived for *Background Replacement* edits, which legitimately affect large image regions. The resulting distributions are shown in Fig. 4.

Filtering Stage 2: CLIP Directional Similarity and DINO The second stage targets semantic correctness and identity preservation using CLIP Directional Similarity [52] and DINO [48]. Pairs with a negative directional similarity score are rejected, indicating that the visual change contradicts the text instruction. DINO scores below 0.7 are flagged for excessive identity change, with an exemption for *Background Replacement* edits. Distributions are shown in Fig. 4.

Filtering Stage 3: VIEScore The final stage applies VIEScore [53], the most computationally intensive filter. Pairs where either score falls at or below 7 are rejected. Because open-weight VLMs perform significantly worse than proprietary ones on this task [53], we fine-tune a LoRA adapter to substantially improve evaluation accuracy; this is described in Section III-C.

Post-Filtering Distribution Metrics After all filtering stages are complete, FID, KID, and CMMD are recomputed on the final filtered dataset to quantify the overall improvement in distributional quality against the pre-filtering baseline.

B. Empirical Metric Analysis via Quality-Stratified Test Datasets

A central challenge in automated quality filtering is that the absolute values of metrics such as SSIM or LPIPS are not universally meaningful - an SSIM value of 0.8 may indicate an excellent edit in one context and a failed one in another, depending on the expected degree of change [75]. Standard benchmarks derived from other domains or generation tasks do not transfer directly to instruction-based face editing. We therefore develop a data-driven approach to determine what

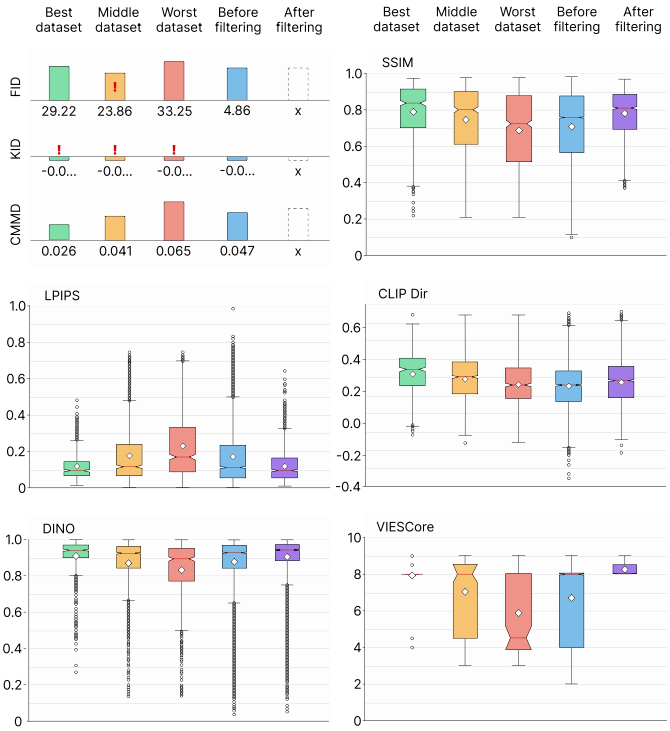

Fig. 4. Metric distributions across the best (0% bad), middle (50% bad), and worst (100% bad) quality datasets, as well as the full dataset before and after filtering.

metric values constitute acceptable quality specifically within our dataset.

Human Annotation: A representative subset of 1,500 image pairs is drawn from the full generated dataset, stratified to preserve the same distribution of models and edit types as the complete dataset. Each pair is manually evaluated by a human annotator and classified as either acceptable or unacceptable. For every image classified as unacceptable, the primary failure reason is recorded from a fixed taxonomy of observed failure modes, including: prompt not executed, deformed image, unrealistic appearance, and others. This annotation serves as the ground truth for all subsequent threshold analysis.

Quality-Stratified Test Datasets: From the annotated subset, three test datasets are constructed (see Fig. 5): a *best* dataset containing 714 good images (100% acceptable quality), a *worst* dataset containing 786 bad images (0% acceptable quality), and a *middle* dataset composed of 1,428 images with equal shares of good and bad pairs (50% acceptable quality). The middle dataset simulates a realistic unfiltered dataset and is used to assess how well each metric separates the two quality classes in a mixed setting.

Threshold Derivation: All metrics were computed across the three test datasets, as well as the full dataset (Fig. 4). Thresholds were established at the point where the worst-dataset distribution clearly diverges from the best-dataset distribution. The exact derived values are detailed in the previous section outlining the filtering stages.

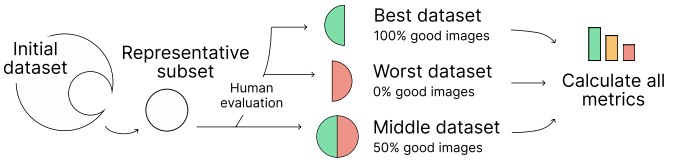

Fig. 5. Creation of the test datasets. Following human evaluation of a representative subset, images are stratified into *best*, *worst*, and *middle* datasets with varying proportions of acceptable images to evaluate metric performance.

C. VIEScore Adaptation via LoRA Fine-Tuning

As discussed above, open-weight VLMs perform significantly worse than proprietary models on VIEScore evaluation [53]. To make VIEScore viable without relying on a closed-source API, we fine-tune a Low-Rank Adaptation (LoRA) [76] for Qwen3-VL-32B-Instruct using the 1,500 human-annotated image pairs collected during the threshold analysis described above.

Training Label Construction Each annotated image pair requires SC and PQ ground-truth scores. Rather than relying on the base model's own predictions - which are known to be unreliable - we derive scores directly from the human annotations. Good images are assigned SC $= 9$, PQ $= 9$. For bad images, each failure category from the human annotation taxonomy is mapped to a characteristic SC/PQ pair reflecting the nature of the defect: for example, a *Prompt not executed* failure receives SC $= 0$, PQ $= 7$, since the image quality is acceptable but the instruction was not followed; a *Deformed image* failure receives SC $= 3$, PQ $= 1$, reflecting both semantic and perceptual failure.

Score Jitter for Regularization A key concern when fine-tuning on a small dataset (1,250 training samples after a stratified 80/20 split preserving model and edit-type distributions) is that the model may memorize fixed score-per-category mappings rather than learning to assess quality on a continuous scale [77], [78]. To mitigate this, a random perturbation of ± 1 - drawn uniformly from $\{-1, 0, +1\}$ - is applied independently to the SC and PQ target scores at every training access. Since the jitter is re-sampled at each epoch, the model observes slightly different numeric targets for the same image across training, effectively augmenting the supervision signal and discouraging memorization.

Training Setup and Results The LoRA adapter targets the query, key, value, and output projection matrices of the attention layers (rank $r = 16$, $\alpha = 32$, dropout 0.05). The model is trained for 5 epochs with a cosine learning-rate schedule (peak 2×10^{-4}, warmup ratio 0.05) in bfloat16 precision using gradient checkpointing. On the held-out test set of 250 samples, the fine-tuned model achieves an SC score within ± 1 of the human annotation in 73.2% of cases and a PQ score within ± 1 in 67.6% of cases, with an overall mean absolute error of 1.65 on the 0–10 scale and zero parse failures.

In contrast, the base Qwen3-VL-32B-Instruct model without fine-tuning produces mean VIEScores of 8.22 on the best-quality dataset and 7.84 on the worst-quality dataset - a gap of only 0.38 points - demonstrating that without adaptation

the model is essentially unable to distinguish between good and bad edits, consistently awarding high scores regardless of actual image quality.

D. Statistical Information

Fig. 6 provides an overview of the dataset composition, illustrating the frequency of targeted objects, categorizations of edit types, distributions of successful versus failed image generations, and the subject demographics.

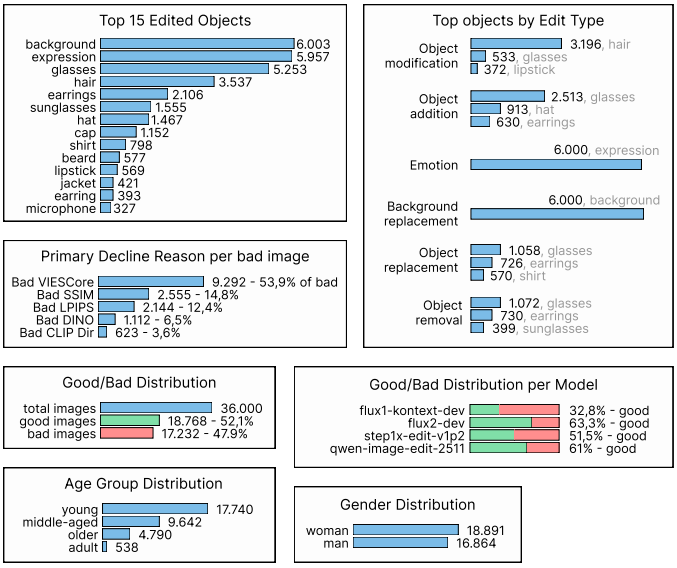

Fig. 6. Dataset statistics: model and edit type distributions, top edited objects, demographic breakdown, model×edit-type heatmap, and a flow diagram linking models to edit types and edited objects.

The four models and six edit types each contribute equal shares (25% and ≈16.7%, respectively). The most frequently targeted objects are backgrounds, facial expressions, hair, and eyewear. The source images from FFHQ span a broad demographic range in terms of gender (roughly balanced between women and men) and age (predominantly young and middle-aged subjects).

IV. Discussion

Building a large-scale dataset with SOTA generative models is rarely as straightforward as it appears. We share a few practical observations from this project.

Metrics need validation before use. Standard quality metrics cannot be adopted blindly for instruction-guided face editing. CLIPScore, for example, showed no separation between good and bad pairs in our data and had to be dropped. Every retained metric required empirical calibration against human judgments before it could serve as a reliable filter.

Filtering is inherently a trade-off. Aggressive filtering risks discarding acceptable images alongside unacceptable ones, making precise threshold calibration critical.

Keeping datasets relevant to the SOTA. Throughout this work, new state-of-the-art models were released faster than a single dataset construction cycle. For example, HunyuanImage 3.0 [79] was released shortly before the publication of this dataset, precluding its inclusion. Datasets risk obsolescence between creation and publication. To remain relevant, benchmarks must continuously track the technological frontier.

Human-annotated subsets unlock metric adaptation. Collecting even a small set of human-reviewed image pairs from the evaluated dataset proved disproportionately valuable. Beyond threshold calibration, such annotations open a direct path to *learned* metric adaptation: by fine-tuning an automatic evaluator like VIEScore on this subset via LoRA, it may be possible to shift the model's judgments toward the specific distribution of face-editing quality relevant to this benchmark. We view the investigation of LoRA-based fine-tuning for VIEScore using a small human-reviewed subset curated via our proposed pipeline as a concrete direction for future work.

Heuristic score mappings may introduce bias. The mapping from qualitative failure categories to numeric SC and PQ scores used as training labels for the VIEScore LoRA is inherently heuristic: values such as $\mathrm{SC} = 0, \mathrm{PQ} = 7$ for *Prompt not executed* or $\mathrm{SC} = 3, \mathrm{PQ} = 1$ for *Deformed image* reflect expert judgment rather than empirically grounded ground truth. While score jitter reduces memorization of fixed categories, it does not eliminate the underlying bias: failure categories that are genuinely ambiguous or that overlap in their perceptual and semantic impact may receive scores that do not accurately capture their true severity. As a consequence, the fine-tuned model's predictions inherit the assumptions encoded in the mapping, limiting their interpretability – a predicted SC of 3 means "roughly as bad as a deformed image" rather than conveying an absolute quality level. A more principled alternative would collect continuous numeric ratings directly from annotators for each failure instance, although at substantially higher annotation cost.

Detection impact remains undemonstrated. Despite the dataset's stated motivation of supporting detection of instruction-guided facial manipulations, no detector is trained or evaluated on AIFE in this work. Whether the dataset's scale, diversity, and quality filtering actually translate into improved detection performance is therefore an open question. Future work should benchmark forensic detectors trained on AIFE against those trained on prior datasets to validate the practical utility of the construction choices made here.

Empirical Model Verification. New models do not guarantee high quality. Despite exhaustive testing across numerous prompts and parameter configurations, OmniGen2 failed to generate realistic results and was explicitly excluded to maintain the dataset's high standards.

Reliance on the FFHQ dataset. The exclusive use of FFHQ source images may also limit representativeness. FFHQ provides high-resolution, diverse portrait images and is well suited for controlled dataset construction, but it does not fully represent real-world manipulation settings such as compressed social-media uploads, screenshots, surveillance footage, webcam imagery, group photos, extreme lighting, unusual poses, or low-resolution images. As a result, detectors or analyses

developed with AIFE should be evaluated on additional data before drawing conclusions about deployment in uncontrolled environments.

Single-model prompt generation. The prompt-generation process might introduce another source of bias. All captions, edit prompts, edit captions, and edited-object lists were generated with GPT-4.1. This single-model design improves consistency and scalability, but it may also introduce systematic style, wording, cultural, safety-policy, and object-selection biases. GPT-4.1 prompts may be more explicit, structured, and instruction-complete than real user prompts, adversarial prompts, casual prompts, multilingual prompts, or multi-turn editing requests. Moreover, because the same model produces several metadata fields, errors in captions, edited-object lists, and edit prompts may be correlated. Future versions should compare or mix prompts from multiple prompt generators, human-written prompts, and prompts that better approximate real user and/or attacker behavior.

Intended downstream use cases. Within these limits, the intended downstream uses of our proposed dataset are detector benchmarking on instruction-guided facial edits, cross-generator generalization studies, augmentation of existing training sets and benchmarks, artifact and failure-mode analysis, evaluation of quality-filtering pipelines for edited face datasets, and security-oriented stress testing of forensic detectors under natural-language editing workflows. However, AIFE should not be treated as a complete proxy for real-world facial manipulation attacks.

V. Conclusion

We presented ADRIAN InstructFace-Edit (AIFE), a dataset of 36,000 edited facial image pairs generated by four SOTA instruction-guided image editing models across six semantically distinct edit types. Within this collection, 18,768 pairs are curated as high-quality (acceptable) edits, while the remaining pairs are explicitly annotated as unacceptable but are published alongside the high-quality subset to facilitate broader research into model failures and artifact detection. To ensure effective classification, we developed a multi-stage filtering pipeline combining structural (SSIM, LPIPS), semantic (CLIP Directional Similarity, DINO), and MLLM-based (VIEScore) quality checks, with all thresholds derived empirically from a 1,500-image human-annotated subset. To make VIEScore practical without proprietary APIs, we fine-tuned a LoRA adapter for Qwen3-VL-32B-Instruct, improving SC accuracy within-± 1 from near-random to 73.2%. To the best of our knowledge, AIFE is the first facial forensics dataset built exclusively around instruction-guided editing, filling a critical gap left by existing datasets that rely on older techniques or mask-based inpainting.

For future work, as mentioned in the discussion, we plan to update our dataset with newer models such as HunyuanImage 3.0 [79]. Additionally, we plan to investigate LoRA-based fine-tuning for VIEScore using a small human-reviewed subset. Importantly, we also plan to evaluate existing detectors and train our own detector on our proposed dataset.

We will publicly release the dataset and the generation pipeline to support future work on detecting instruction-guided facial manipulations. The code is available at: https://github.com/vika-v-v/adrian-instructface-edit. The dataset is available at: https://hsbi.sciebo.de/s/pb6ZB4ZeCtmJxGm (password: `adrian`).

Acknowledgment

This research is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

References

[1] K. Roose, "An a.i.-generated picture won an art prize. artists aren't happy." *The New York Times*, 2022. [Online]. Available: https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

[2] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," 2018. [Online]. Available: https://arxiv.org/abs/1812.08685

[3] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, and S. Nahavandi, "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

[4] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[5] Black Forest Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, and C. Diagne, "Flux.1 kontext: Flow matching for in-context image generation and editing in latent space," 2025. [Online]. Available: https://arxiv.org/abs/2506.15742

[6] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media + Society*, vol. 6, no. 1, p. 2056305120903408, 2020. [Online]. Available: https://doi.org/10.1177/2056305120903408

[7] R. Mubarak, T. Alsbou'i, O. Alshaikh, I. Inuwa-Dute, S. Khan, and S. Parkinson, "A survey on the detection and impacts of deepfakes in visual, audio, and textual formats," *IEEE Access*, vol. PP, pp. 1–1, 01 2023.

[8] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis," *Journal of Sensor and Actuator Networks*, vol. 14, no. 1, 2025. [Online]. Available: https://www.mdpi.com/2224-2708/14/1/17

[9] T. Aczel, L. Vettor, A. Plesner, and R. Wattenhofer, "The unwinnable arms race of ai image detection," 2025. [Online]. Available: https://arxiv.org/abs/2509.21135

[10] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24 480–24 489, 2023.

[11] Black Forest Labs, "FLUX.2: Frontier Visual Intelligence," Nov. 2025. [Online]. Available: https://bfl.ai/blog/flux-2

[12] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, and K. Yan, "Qwen-image technical report," 2025. [Online]. Available: https://arxiv.org/abs/2508.02324

[13] S. Liu, Y. Han, P. Xing, F. Yin, R. Wang, and W. Cheng, "Step1x-edit: A practical framework for general image editing," 2025. [Online]. Available: https://arxiv.org/abs/2504.17761

[14] F. Bäumer, S. Denisov, M. Geierhos, and Y. Su Lee, "Towards authority-dependent risk identification and analysis in online networks." in *Proceedings of Artificial Intelligence, Machine Learning and Big Data for Hybrid Military Operations (AI4HMO)*, 2021.

[15] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New York: IEEE, 2022, pp. 10 674–10 685.

[17] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, and J. Müller, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 1862–1874. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf

[18] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, and H. Saini, "Scaling rectified flow transformers for high-resolution image synthesis," in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML'24. JMLR.org, 2024.

[19] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," 2022. [Online]. Available: https://arxiv.org/abs/2208.01626

[20] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, and T. Dekel, "Imagic: Text-based real image editing with diffusion models," in *Conference on Computer Vision and Pattern Recognition 2023*, 2023.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, and P. Dhariwal, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[22] K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: a manually annotated dataset for instruction-guided image editing," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[23] M. Hui, S. Yang, B. Zhao, Y. Shi, H. Wang, and P. Wang, "Hq-edit: A high-quality dataset for instruction-based image editing," 2024. [Online]. Available: https://arxiv.org/abs/2404.09990

[24] H. Zhao, X. Ma, L. Chen, S. Si, R. Wu, and K. An, "Ultraedit: Instruction-based fine-grained image editing at scale," 2024. [Online]. Available: https://arxiv.org/abs/2407.05282

[25] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, and N. Yu, "Hive: Harnessing human feedback for instructional visual editing," *arXiv preprint arXiv:2303.09618*, 2023.

[26] T.-J. Fu, W. Hu, X. Du, W. Wang, Y. Yang, and Z. Gan, "Guiding instruction-based image editing via multimodal large language models," in *ICLR*, 2024. [Online]. Available: https://arxiv.org/abs/2309.17102

[27] Y. Huang, L. Xie, X. Wang, Z. Yuan, X. Cun, and Y. Ge, "Smartedit: Exploring complex instruction-based image editing with multimodal large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8362–8371.

[28] W. Peebles and S. Xie, "Scalable diffusion models with transformers," *arXiv preprint arXiv:2212.09748*, 2022.

[29] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, and O. Ashual, " Emu Edit: Precise Image Editing via Recognition and Generation Tasks ," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2024, pp. 8871–8879. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00847

[30] S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, and R. Yan, "Omnigen: Unified image generation," 2024. [Online]. Available: https://arxiv.org/abs/2409.11340

[31] C. Wu, P. Zheng, R. Yan, S. Xiao, X. Luo, and Y. Wang, "Omnigen2: Exploration to advanced multimodal generation," *arXiv preprint arXiv:2506.18871*, 2025.

[32] Q. Cai, J. Chen, Y. Chen, Y. Li, F. Long, and Y. Pan, "Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer," *arXiv preprint arXiv:2505.22705*, 2025.

[33] P. Wang, Y. Shi, X. Lian, Z. Zhai, X. Xia, and X. Xiao, "Seededit 3.0: Fast and high-quality generative image editing," 2025. [Online]. Available: https://arxiv.org/abs/2506.05083

[34] OpenAI, "Introducing 4o Image Generation," Mar. 2025. [Online]. Available: https://openai.com/index/introducing-4o-image-generation/

[35] OpenAI, "The new ChatGPT Images is here," Dec. 2025. [Online]. Available: https://openai.com/index/new-chatgpt-images-is-here/

[36] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, and I. Dhillon, "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities," Jul. 2025, arXiv:2507.06261 [cs]. [Online]. Available: http://arxiv.org/abs/2507.06261

[37] D. Sharon and N. Brichtova, "Image editing in Gemini just got a major upgrade," Aug. 2025. [Online]. Available: https://blog.google/products/gemini/updated-image-editing-model/

[38] Google, "Nano Banana 2: Combining Pro capabilities with lightning-fast speed," Feb. 2026. [Online]. Available: https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/

[39] Artificial Analysis, "Artificial Analysis Image Editing Leaderboard," https://artificialanalysis.ai/image/leaderboard/editing, 2026.

[40] Arena Intelligence, Inc., "Image Edit Arena," https://arena.ai/leaderboard/image-edit, 2026.

[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6629–6640.

[42] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *International Conference on Learning Representations*, 2018.

[43] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, " Rethinking FID: Towards a Better Evaluation Metric for Image Generation ," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2024, pp. 9307–9315. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00889

[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and S. Agarwal, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[47] M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, and P. Bojanowski, "Emerging properties in self-supervised vision transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233444273

[48] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, and V. Khalidov, "Dinov2: Learning robust visual features without supervision," 2023.

[49] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, and C. Jose, "DINOv3," 2025. [Online]. Available: https://arxiv.org/abs/2508.10104

[50] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, " DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation ," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 22 500–22 510. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02155

[51] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7514–7528. [Online]. Available: https://aclanthology.org/2021.emnlp-main.595/

[52] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM Trans. Graph.*, vol. 41, no. 4, Jul. 2022. [Online]. Available: https://doi.org/10.1145/3528223.3530164

[53] M. Ku, D. Jiang, C. Wei, X. Yue, and W. Chen, "VIEScore: Towards explainable metrics for conditional image synthesis evaluation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12 268–12 290. [Online]. Available: https://aclanthology.org/2024.acl-long.663/

[54] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, and

I. Akkaya, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[55] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," 2019. [Online]. Available: https://arxiv.org/abs/1901.08971

[56] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, " On the Detection of Digital Face Manipulation ," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2020, pp. 5780–5789. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00582

[57] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, and M. Wang, "The deepfake detection challenge (dfdc) dataset," 2020. [Online]. Available: https://arxiv.org/abs/2006.07397

[58] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213.

[59] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, and L. Song, " ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis ," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2021, pp. 4358–4367. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00434

[60] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[61] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, and S. Ozair, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf

[62] H. Cheng, Y. Guo, T. Wang, L. Nie, and M. Kankanhalli, "Diffusion facial forgery detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 5939–5948. [Online]. Available: https://doi.org/10.1145/3664647.3680797

[63] Z. Chen, K. Sun, Z. Zhou, X. Lin, X. Sun, and L. Cao, "Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis," *ArXiv*, vol. abs/2403.18471, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268723887

[64] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, and J. Zhu, "Df40: toward next-generation deepfake detection," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[65] C. Miao, Y. Zhang, M. Luo, W. Feng, K. Zheng, and Q. Chu, "Mffi: Multi-dimensional face forgery image dataset for real-world scenarios," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 13235–13242. [Online]. Available: https://doi.org/10.1145/3746027.3758280

[66] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, and H. Chen, "Dire for diffusion-generated image detection," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 388–22 398.

[67] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," 2023. [Online]. Available: https://arxiv.org/abs/2210.06998

[68] 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.

[69] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 893–911. [Online]. Available: https://aclanthology.org/2023.acl-long.51/

[70] M. Zhu, H. Chen, Q. YAN, X. Huang, G. Lin, and W. Li, "Genimage: A million-scale benchmark for detecting ai-generated image," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 77 771–77 782. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf

[71] J. Park and A. Owens, "Community forensics: Using thousands of generators to train fake image detectors," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 8245–8257.

[72] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.2970919

[73] Black Forest Labs, "FLUX.1 Kontext Prompting Guide - Image-to-Image," https://docs.bfl.ml/guides/prompting_guide_kontext_i2i, 2025.

[74] F. Yin, S. Liu, Y. Han, Z. Wang, P. Xing, R. Wang *et al.*, "Reasonedit: Towards reasoning-enhanced image editing models," 2025. [Online]. Available: https://arxiv.org/abs/2511.22625

[75] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao, and Y. Sui, "A survey on image quality assessment: Insights, analysis, and future outlook," 2025.

[76] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, and S. Wang, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[77] R. Vavekanand, A. K. Ruzimbaevich, and M. Jumaniyozova, "A study on the extraction of training dataset from fine-tuned language models," *Computer Assisted Methods in Engineering and Science*, vol. 32, no. 3, p. 293–315, 2025. [Online]. Available: https://cames.ippt.pan.pl/index.php/cames/article/view/1781

[78] B. Zhu, M. I. Jordan, and J. Jiao, "Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf," *arXiv preprint arXiv:2401.16335*, 2024. [Online]. Available: https://doi.org/10.48550/arxiv.2401.16335

[79] S. Cao, H. Chen, P. Chen, Y. Cheng, Y. Cui, and X. Deng, "Hunyuanimage 3.0 technical report," *arXiv preprint arXiv:2509.23951*, 2025.