SAGI: Semantically Aligned and Uncertainty Guided AI Image Inpainting

Paschalis Giakoumoglou[1,2] Dimitrios Karageorgiou[2]
Symeon Papadopoulos[2] Panagiotis C. Petrantonakis[1]
[1]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki
[2]Information Technologies Institute, CERTH
{giakoupg,ppetrant}@ece.auth.gr, {giakoupg,dkarageo,papadop}@iti.gr

Abstract

Recent advancements in generative AI have made text-guided image inpainting—adding, removing, or altering image regions using textual prompts—widely accessible. However, generating semantically correct photorealistic imagery, typically requires carefully-crafted prompts and iterative refinement by evaluating the realism of the generated content - tasks commonly performed by humans. To automate the generative process, we propose Semantically Aligned and Uncertainty Guided AI Image Inpainting (SAGI), a model-agnostic pipeline, to sample prompts from a distribution that closely aligns with human perception and to evaluate the generated content and discard instances that deviate from such a distribution, which we approximate using pretrained large language models and vision-language models. By applying this pipeline on multiple state-of-the-art inpainting models, we create the SAGI ***D****ataset (*SAGI-D*), currently the largest and most diverse dataset of AI-generated inpaintings, comprising over 95k inpainted images and a human-evaluated subset. Our experiments show that semantic alignment significantly improves image quality and aesthetics, while uncertainty guidance effectively identifies realistic manipulations — human ability to distinguish inpainted images from real ones drops from 74% to 35% in terms of accuracy, after applying our pipeline. Moreover, using* SAGI-D *for training several image forensic approaches increases in-domain detection performance on average by 37.4% and out-of-domain generalization by 26.1% in terms of IoU, also demonstrating its utility in countering malicious exploitation of generative AI. Code and dataset are available at* https://mever-team.github.io/SAGI/

1. Introduction

Image inpainting—reconstructing image regions—became highly accessible with powerful generative AI tools, enabling non-experts to create photorealistic edits [84]. Text-guided inpainting, that adds, removes, or alters regions using textual prompts, has advanced significantly through models like Stable Diffusion [57], DALL-E [55], and Imagen [60]. Achieving high-quality results often requires several attempts and iterative refinement [41], as advanced models struggle with complex prompts [11]. However, successful generations can be so realistic that humans struggle to distinguish them from real photographs [11].

Automated inpainting generation faces two key challenges. First, while state-of-the-art models excel at manipulation, they require detailed prompts for realistic results—basic object labels often lack sufficient context [40]. Second, automating realism assessment is difficult, as standard quality metrics may not reflect perceived realism. To address these, we propose SAGI (Semantically Aligned and Uncertainty Guided AI Image Inpainting), a unified framework integrating Semantically Aligned Object Replacement (SAOR) and Uncertainty Guided Deceptiveness Assessment (UGDA). SAOR leverages image semantics and large language models (LLMs) to create context-aware prompts, producing higher aesthetic quality than simple object labels and captions. UGDA is a realism assessment method that uses vision-language models (VLMs) to compare inpainted images with originals, identifying convincing manipulations. We validate this approach through a user study, showing it aligns with human perception of image realism. An overview of the approach is presented in Fig. 1.

Using this framework along with multiple state-of-the-art inpainting models, we create the SAGI-D dataset, the largest and most diverse collection of AI-generated inpaintings to date, establishing a new benchmark for inpainting detection research. The dataset contains 95,839 inpainted images from 78,684 originals across three datasets: MS-COCO [37], RAISE [19], and OpenImages [32]. We provide each inpainted image with its original version, inpainting mask, and text prompt. For evaluation, we create both in-domain and out-of-domain testing splits. The in-domain split uses the same LLM and source images as the train-

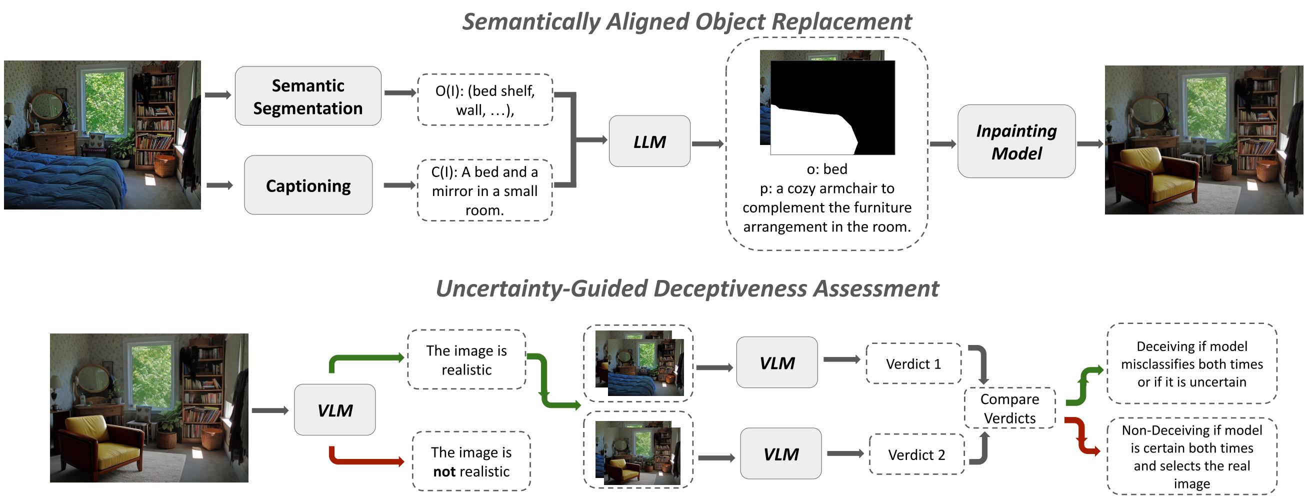


Figure 1. The first row illustrates SAOR: an input image is processed to identify objects (via labels or segmentation) and generate a caption. The language model selects an object and generates a prompt, which, along with the image and mask, guides the inpainting model to produce the result. The second row depicts UGDA: if the inpainted image passes an initial realism check, it undergoes a second evaluation. The VLM compares the inpainted and original images twice, in reverse order. If the VLM's responses differ (indicating uncertainty) or both favor the inpainted image, it is labeled as deceiving; otherwise, non-deceiving.

ing set, for familiar data assessment. The out-of-domain split uses different source images and a different LLM, testing generalization to new data. Beyond advancing inpainting generation, the resulting dataset provides a valuable resource for training and evaluating forensic detection methods that can address challenges regarding potential malicious misuse [73, 78].

Our main contributions are summarized as follows:

- We propose semantically aligned and uncertainty guided image inpainting, a model-agnostic framework, to automate the generation of realistic inpainted images.
- We introduce semantically aligned object replacement to generate semantically coherent prompts.
- We propose uncertainty guided deceptiveness assessment, to estimate the realism of an image at test time.
- We experimentally show that the proposed framework increases quality and aesthetics across several generative models. Through human studies we show that human ability to distinguish inpainted images from real ones drops from 74% to 35% after applying our framework.
- We present the SAGI-D dataset, the largest and most diverse dataset of AI-generated inpainted images to date
- We demonstrate the efficacy of our approach on retraining image forensic methods, enhancing their average in-domain localization performance by 37.4% and out-of-domain generalization by 26.1% in terms of IoU.

2. Related Work

2.1. Image Inpainting

Early image inpainting methods used diffusion for image restoration [8–10, 12] and exemplar-based approaches for object removal [18, 22, 25, 29]. Recent methods are deep learning-based, using CNNs [7], Fourier Convolutions [64], Transformers [72], or Diffusion Models [14]. GAN-based methods are also widely used, to generate coherent images, with techniques such as Context Encoders [53] and dilated convolutions improving results [81]. Diffusion models have emerged as a powerful approach, reconstructing missing regions via noise removal, with models like GLIDE, DALL-E [56] [49] and Stable Diffusion [58] incorporating additional guidance, such as text, for improved control.

2.2. Inpainting Detection

Inpainting detection became crucial with inpainting advancements. Early approaches used patch comparison and connectivity analysis to identify inconsistencies [13, 36, 79, 85]. Deep learning, particularly CNNs, improved detection via filtering, segmentation networks, and encoder-decoder architectures [31, 34, 38, 39, 62, 90]. Transformer-CNN hybrids enhanced long-range dependency and texture analysis [76, 88, 91]. Recent methods integrate frequency-domain, compression artifacts, noise, and semantic features [21, 28, 33, 71]. While effective for traditional inpainting and forgeries like copy-move and splicing, few have been tested on AI-based inpainting, except for evaluations on GLIDE-generated forgeries [21, 28, 71], leaving a gap in assessing performance on modern generative techniques.

2.3. Datasets for Image Inpainting Detection

The development and evaluation of inpainting detection models rely on datasets that capture diverse inpainting techniques and scenarios. Early inpainting datasets used outdated inpainting techniques that produced very poor results

[2, 16, 20, 42, 70, 76]. Several recent datasets [6, 15, 89] employ state-of-the-art generative AI models, but they focus on fully synthetic images rather than inpaintings. This creates a significant gap in the available resources for inpainting research. Among these, CocoGlide [21] and TGIF [44] stand out as the most relevant, utilizing text-to-image inpainting models such as GLIDE, Stable Diffusion, and Adobe Firefly. However, these datasets have notable limitations: CocoGlide is limited in scale, while TGIF, though larger, does not account for more recent and complex inpainting pipelines that produce higher-quality results.

2.4. Image Quality and Aesthetic Assessment

Image quality assessment (IQA) evaluates how distortions affect human perception, evolving from handcrafted features [47, 75] to deep learning methods like NIMA [66] and HyperIQA [63]. Recent approaches utilize CLIP [54, 74, 87], leverage multi-scale inputs and vision-language correspondence but often underperform purely visual methods due to reliance on visual-text similarity, which prioritizes semantic alignment over visual fidelity. Image aesthetic assessment (IAA) is more complex, focusing on attributes like composition and emotional impact [48]. Deep learning dominates IAA, with methods like NIMA [66] and VILA [30] achieving state-of-the-art results. QALIGN [77] advances both IQA and IAA by leveraging VLMs to improve generalization without extensive fine-tuning. However, most approaches are task-specific and struggle with generalization, and none of them focus on realism assessment.

3. Semantically Aligned and Uncertainty Guided AI Image Inpainting

Recent studies have shown that detailed prompts in AI image generation lead to better inpainting results [46, 59], with users employing complex, carefully crafted prompts [41], evaluating image semantics [52] and using multiple models and attempts [67] until achieving desired results. Inspired by these observations, we argue that automatically generating high-quality inpainted images, deceiving to the human eye, requires addressing two key challenges: ensuring semantic coherence in manipulations and assessing the realism of generated images. To this end, we propose SAGI, a model-agnostic pipeline for a) sampling prompts from a distribution that aligns with human perception of reality and b) assessing and discarding generated samples that deviate from such distribution—techniques notably absent in previous works. We consider pre-trained LLMs and VLMs to be the best available approximations for such distribution, due to their training data scale. We employ them in our automated pipeline with two main components: 1) Semantically Aligned Object Replacement for contextually appropriate manipulations and 2) Uncertainty-Guided Deceptiveness Assessment for evaluating realism.

3.1. Semantically Aligned Object Replacement

Generating realistic inpainted imagery requires selecting appropriate objects for replacement and creating contextually relevant prompts that maintain semantic consistency. To address this challenge, we propose Semantically Aligned Object Replacement (SAOR), a method that automates object selection and prompt generation. Formally, let I be an image and $S(I)$ represent its semantic content. We aim to sample a prompt p and object o under the distribution $\mathcal{P}(p, o \mid S(I))$ that characterizes how humans would select regions and generate prompts given image semantics. We approximate $S(I)$ through two components: the image caption $C(I)$ and the set of displayed objects $O(I)$. This decomposition captures both local semantics (through objects) and global relationships (through captions), providing a descriptive representation of image semantics.

$$S(I) \approx \{C(I), O(I)\} \tag{1}$$

Object masks $M(I) = \{m_1, m_2, \ldots, m_n\}$ and their respective labels $O(I) = \{o_1, o_2, \ldots, o_n\}$ are obtained from labeled data if available; otherwise, a panoptic segmentation model $\Phi_{\text{seg}} : I \to (O(I), M(I))$ is used. Similarly, captions $C(I)$ are taken directly from labeled data or generated by a captioning model Ψ_{cap}. To approximate $\mathcal{P}$, we leverage LLMs, which have been extensively trained on human text, making them ideal candidates to model human reasoning about visual semantics. Specifically:

$$\mathcal{P}(p, o \mid S(I)) \approx \Theta_{\text{llm}}(p, o \mid O(I), C(I)) \tag{2}$$

where Θ_{llm} represents the LLM's conditional probability distribution over generated text. The language model, leveraging the caption to understand the image content, is instructed to produce complex, contextually appropriate prompts, maintaining semantic consistency with the surrounding content. The generated prompt p, the mask m corresponding to the object selected by the LLM, and the original image I are then provided to an inpainting model Γ to generate the inpainted image $\hat{I} = \Gamma(I, m, p)$. The SAOR pipeline is presented at the top of Fig. 1

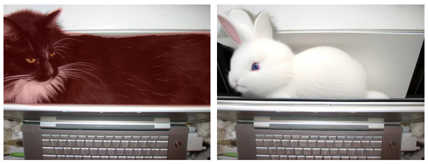
3.2. Uncertainty-Guided Deceptiveness Assessment

Assessing the realism of inpainted images at test-time is crucial for discarding unrealistic samples and refining the generation process. However, quantifying realism is challenging due to its subjective nature [69]. To approximate this subjective judgment, we propose Uncertainty-Guided Deceptiveness Assessment (UGDA), which leverages pretrained VLMs, which have proven successful in synthetic

(a) "a delicious apple to add a pop of color to the scene"

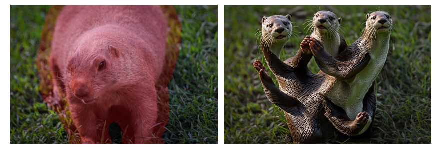
(b) "lush green vines and foliage, intertwining around ancient stone pillars"

(c) "a fluffy golden retriever"



(d) "a fluffy white bunny sitting inside a laptop."

(e) No prompt



(f) "a curious otter exploring the grassy field"

Figure 2. Original (with semi-transparent red inpainting mask) and inpainted images from three datasets, with prompts shown below each pair for text-guided models. The first row shows images classified as deceiving by UGDA 3.2, and the second row shows non-deceiving images. Each column corresponds to a different dataset: MS-COCO [37] (first), RAISE [19] (second), and OpenImages [32] (third).

image detection [80], to compare inpainted images with their originals.

Let I denote the original image and $\hat{I}$ the inpainted version of I. Using a VLM Ω_{vlm}, we perform a two-stage assessment. Based on empirical observations, unrealistic manipulations are often easily identifiable, while realistic ones require a more thorough evaluation. In the first stage, $\hat{I}$ is assessed for realism. Images passing this check undergo a second, more rigorous evaluation. The second stage leverages this key idea: realistic inpainted images align closer with the distribution of semantically coherent images that a pretrained VLM represents. Therefore, aleatoric uncertainty should increase in a proxy task of ranking the original and its corresponding inpainted image according to their semantic alignment with the learned distribution of the VLM: the bigger the aleatoric uncertainty in this proxy task is, the closer the inpainted image is expected to align with the distribution of semantically correct images. To capture this uncertainty increase, we assume that a confident assessment (i.e.low uncertainty), generated by the VLM, should remain stable under semantically invariant perturbations. If the VLM's assessment varies under such perturbations, it indicates uncertainty about the inpainted image's realism. We argue that a suitable perturbation is input reordering. Specifically, we perform two assessments with reversed image order:

$$s_1 = \Omega_{\text{vlm}}(I, \hat{I}) \quad \text{and} \quad s_2 = \Omega_{\text{vlm}}(\hat{I}, I), \tag{3}$$

where $s_1, s_2 \in \{c_I, c_{\hat{I}}, c_{\text{both}}\}$ represents the image assessed as more realistic, with c_I indicating the original, $c_{\hat{I}}$ the inpainted image, and c_{both} equal realism. Variation in responses or consistently choosing the inpainted image indicates uncertainty. The classification rule: $\hat{I}$ is deceiving if $(s_1 = c_{\hat{I}} \vee s_2 = c_{\hat{I}}) \vee (s_1 = s_2 = c_{\text{both}})$. Otherwise, it is non-deceiving. This approach enhances reliability in distinguishing between deceiving and easily identifiable synthetic images through uncertainty-aware, order-based evaluations. Fig. 2 shows representative examples from UGDA across different sources. The first row depicts high-quality inpainting examples classified as deceiving by UGDA, where the manipulations are seamlessly integrated with the original content (see suppl. for detailed classification process). The UGDA pipeline is shown in Fig. 1

3.3. SAGI-D

Using the proposed pipeline, we introduce SAGI-D , the first semantically aligned deceptive dataset for AI-generated inpainting detection, designed to evaluate the effectiveness of our components. The dataset leverages Semantically Aligned Object Replacement (SAOR) for context-aware prompt generation and Uncertainty-Guided Deceptiveness Assessment (UGDA) for realism evaluation, ensuring high-quality and diverse inpainted images.

Sources of authentic images. To ensure diversity and robustness, we leverage datasets spanning multiple domains: general object detection, high-resolution photography, and large-scale segmentation. Specifically, we utilize three publicly available datasets for authentic image sources: (1) MS-COCO [37], which provides images with captions and object masks across 80 categories, (2) RAISE [19], a high-resolution dataset of 8,156 uncompressed RAW images designed for forgery detection evaluation and (3) OpenImages [32], which offers extensive object segmentation data with over 2.7 million segmentations across 350 categories.

SAOR configuration. We use dataset-provided masks and captions for MS-COCO and OpenImages, while for RAISE, we use OneFormer [24] as the segmentation model Φ_{seg} and BLIP-2 [35] as the captioning model Ψ_{cap}. To generate prompts, we employ as the language model Θ_{llm} ChatGPT 3.5 [50] for MS-COCO and RAISE as well as Claude Sonnet 3.5 [3] for OpenImages. The prompt engineering methodology is detailed in the supplementary material.

UGDA configuration. We use GPT-4o [51] as the VLM Ω_{vlm}, chosen for its effectiveness in synthetic image detection [80] and for achieving the best performance in our respective ablation studies presented in Tab. 3. Empirically, we observed that QAlign effectively filters out low-quality images, so we applied UGDA to approximately half of the test inpainted images, selecting those with the highest QAlign scores. The complete prompt engineering methodology is detailed in the supplementary material.

Inpainting Models. Diverse inpainting models are crucial for creating robust detection datasets and high-quality images, capturing varied artifacts, styles, and visual traits. We use five pipelines: HD-Painter [43], BrushNet [26], PowerPaint [92], ControlNet [86], and Inpaint-Anything [83], including its Remove-Anything variant for object removal, with their original padding and normalization strategies maintained. These pipelines support eight models, primarily based on Stable Diffusion [58], except Remove-Anything, which uses LaMa [64]. One-sixth of the images underwent a second inpainting round with different masks and prompts, simulating real-life multi-edit scenarios. Also, due to memory limits, images were resized to a maximum dimension of 2048 pixels, except for Inpaint-Anything, that preserves original dimensions via cropping and resizing.

Preservation of unmasked area. We categorize inpainted images based on how models handle unmasked regions: if the unmasked region is preserved, we refer to them as Spliced (SP) images; if regenerated, as Fully Regenerated (FR) images. Inpaint-Anything preserves unmasked regions through copy-paste (SP), while ControlNet regenerates the full image (FR). BrushNet, PowerPaint, and HD-Painter can produce both SP and FR images depending on post-processing settings (e.g., blending or upscaling). Remove-Anything inherently preserves unmasked regions, thus producing SP images. This diversity in processing approaches contributes to a more comprehensive dataset, as FR images are typically more challenging to detect than SP images [65].

Dataset splits. Our dataset considers both in-domain performance as well as generalization to new data. Regarding in-domain evaluation, we use MS-COCO (60,000 training and nearly all 5,000 validation images) and RAISE (7,735 images processed with Φ_{seg}, yielding 25,674 image-mask-model combinations). To test generalization across different image sources and LLMs, we create an out-of-domain split from OpenImages (6k test images), by also using a different language model Θ_{llm}, i.e. Claude versus ChatGPT for COCO and RAISE. Detailed statistics are provided in the supplementary material.

4. Experimental Evaluation

Our experimental evaluation focuses on two main directions: (1) assessing the quality, aesthetics, and realism of images inpainted with *SAGI*, and (2) evaluating the effectiveness of SAGI-D as a benchmark for inpainting detection.

4.1. Image Quality and Realism Evaluation

We conduct two ablations to validate our design choices. First, we evaluate SAOR's use of language models by comparing three prompting approaches: (1) using just object labels, (2) combining object labels with image captions, and (3) feeding both labels and captions to an LLM to generate enhanced prompts. Second, we assess UGDA's effectiveness through a human study with 42 participants on 1,000 images, comparing human perception against model performance on images classified as *deceiving* or *non-deceiving*. We also ablate the VLM employed in UGDA, by evaluating the ability of several popular VLMs to align with human perception of image realism.

Mdl	**SAOR**			**AS**	**CS Ae**	**QA Ae**	**CS Qlt**	**QA Qlt**
	Obj	**Cap**	**LLM**					
BN	✓			5.55	-0.01	2.63	0.68	3.90
	✓	✓		5.69	0.14	2.68	0.69	3.94
[27]	✓	✓	✓	**5.79**	**0.38**	**2.75**	**0.69**	**3.95**
CN	✓			5.35	-0.07	2.65	0.69	3.88
	✓	✓		5.40	0.01	2.68	0.70	3.92
[86]	✓	✓	✓	**5.46**	**0.14**	**2.71**	**0.70**	**3.92**
HDP	✓			5.74	0.51	2.8	0.71	**4.12**
	✓	✓		5.80	0.49	2.8	0.71	**4.12**
[43]	✓	✓	✓	**5.9**	**0.7**	**2.83**	**0.71**	4.11
IA	✓			5.49	-0.12	2.44	0.66	3.76
	✓	✓		5.60	-0.08	2.50	0.67	**3.83**
[82]	✓	✓	✓	**5.68**	**0.17**	**2.51**	**0.67**	3.81
PPt	✓			5.68	0.46	2.75	**0.71**	**4.06**
	✓	✓		5.76	0.44	2.78	0.70	4.05
[92]	✓	✓	✓	**5.89**	**0.61**	**2.83**	0.70	4.02
	✓			5.56	0.15	2.66	0.69	3.96
Avg	✓	✓		5.65	0.20	2.69	0.69	**3.98**
	✓	✓	✓	**5.74**	**0.40**	**2.73**	**0.70**	3.96

Table 1. Comparison of metrics across inpainting models with different prompt types. The SAOR columns indicate the employed prompting method: object labels (Obj), image captions (Cap), and LLM-generated prompts (LLM). Bold values indicate better performance per model-metric pair. Pipelines: BN (BrushNet), CN (ControlNet), HDP (HD-Painter), IA (Inpaint-Anything), PPt (PowerPaint). Metrics: AS (Aesthetics Score), CS Ae (Clip Aesthetics), QA Ae (QAlign Aesthetics), CS Qlt (Clip Quality), QA Qlt (QAlign Quality). LLM prompts notably enhance aesthetics, while maintaining performance even in quality metrics like CS Qlt, that are typically less sensitive to semantic variations.

SAOR evaluation. To compare LLM-generated prompts with object labels or labels and captions, we evaluated 900 original images across five inpainting models, generating 4,500 images per prompt type. We compared aesthetics and quality metrics, including CLIP Similarity [54] for aesthet-

Metric	Accuracy Top	Accuracy Bottom	Accuracy Diff	Mean IoU Top	Mean IoU Bottom	Mean IoU Diff
AS	54.7	55.1	+0.4	26.6	26.3	+0.2
QA Qlt	56.2	53.5	+2.9	28.1	24.8	+3.3
CS Qlt	61.2	48.5	+12.7	30.6	22.3	+8.3
CS Ae	58.9	50.9	+8.0	28.0	24.39	+3.1
QA Ae	61.3	48.4	+12.9	30.4	22.4	+8.0
	Dec	**Non-Dec**	**Diff**	**Dec**	**Non-Dec**	**Diff**
UGDA	35.2	73.7	**+38.5**	12.7	39.9	**+26.8**

Table 2. Comparison of human accuracy and mean IoU on 500 inpainted images, separated into two groups for each metric. For AS, QA Qlt, CS, and QA Ae, groups are formed by ranking images and splitting them into the top 50% and bottom 50% based on their metric scores; for UGDA, groups are based on its assessment (deceiving/non-deceiving). UGDA can clearly separate deceiving from non-deceiving images.

VLM / Human Eval.	Acc ↓	IoU ↓
w/o UGDA	54.9	26.4
Gemma-3-27b [68]	$45.4^{+9.5}$	$20.0^{+6.4}$
Mistral-S-3.1-24B [1]	$45.2^{+9.7}$	$20.4^{+6.0}$
Qwen2.5-VL-7B [5]	$45.2^{+9.7}$	$21.0^{+5.4}$
Claude-3-7-sonnet [4]	$43.4^{+11.5}$	$17.8^{+8.6}$
Gemini-2.5-flash [17]	$38.6^{+16.3}$	$16.5^{+9.9}$
GPT-4o [51]	$\mathbf{35.2^{+19.7}}$	$\mathbf{12.7^{+13.7}}$

Table 3. VLM ablation for UGDA. Human accuracy and mean IoU for distinguishing inpainted from real images after filtering with different VLMs. Lower values indicate more deceiving inpaintings. All VLMs improve over the unfiltered baseline, with GPT-4o achieving the strongest alignment with human perception.

ics (CS Ae) [23] and quality (CS Qlt) [74] , QAlign [77] for quality (QA Qlt) and aesthetics (QA Ae), and Aesthetic Score (AS) [61]. Tab. 1 shows that LLM prompts consistently outperform object labels and caption prompts across most metrics and models, enhancing both image quality and aesthetics. All models achieve higher aesthetic metrics with LLM prompts. BrushNet and ControlNet excelled across all metrics, while other models showed slight variations in quality metrics. Quality improvements were marginal compared to aesthetic gains, as quality metrics primarily focus on technical aspects of the image which are less influenced by prompt's content. Overall, these results highlight the richer semantic guidance of LLM prompts, significantly enhancing inpainting aesthetics, while also improving some technical aspects of the generated visual content.

UGDA evaluation. We validated UGDA through a user study with 42 participants on 1,000 images: 250 inpainted ones classified by UGDA as deceiving, 250 as non-deceiving and 500 authentic. Images were selected to avoid redundancy, i.e. no authentic-inpainted overlap and no multiple inpainted versions per authentic source image. Participants evaluated batches of 20 images, with each image receiving 3-5 independent assessments. Detailed participant statistics are provided in the supplementary material.

Tab. 2 compares human accuracy and mean IoU on 500 inpainted images, split into two groups per metric. For AS, QA Qlt, CS, and QA Ae, splits are based on top/bottom 50% rankings; for UGDA, on its deceiving/non-deceiving assessment. Existing quality and aesthetic metrics (AS, QA Qlt, CS, QA Ae) show minimal differences among groups (accuracy: 0.4–12.9; IoU: 0.2–8.3), indicating limited correlation with human perception of realism. In contrast, UGDA yields much larger gaps (accuracy: 38.5; IoU: 26.8), effectively identifying realistic inpaintings. This underscores the need for uncertainty guidance to capture realism.

To further validate the efficacy and model-agnostic nature of UGDA, we compare the performance of several VLMs. Tab. 3 compares human ability to distinguish inpainted from real images, both in terms of accuracy and mean IoU, after non-deceiving samples have been filtered out using our methodology. Results demonstrate that our method enables all tested VLMs to better align with human preference and discard non-realistic inpaintings, as evidenced by the positive differences over a baseline without UGDA-based filtering. As GPT-4o shows the strongest alignment with human perception, we employ it as the primary model in our benchmark creation.

Quantitative comparison with state-of-the-art. To the best of our knowledge, SAGI-D is the largest collection of AI-generated inpainted images. Tab. 4 compares SAGI-D with existing datasets. With 95,839 inpainted images from 77,900 originals, it surpasses TGIF (74,976 from 3,124) and CocoGlide (512 from 512) in scale. Unlike single-source datasets, SAGI-D integrates COCO, RAISE, and OpenImages, enhancing diversity. It employs eight inpainting models across five pipelines, considers advanced manipulations such as double inpainting, AI generated content, and object removal, and allows resolutions up to 2048p—exceeding TGIF's 1024p and CocoGlide's 256p. SAGI-D uniquely includes human benchmark and out-of-domain test subsets, making it valuable for evaluating detection under diverse conditions. It outperforms existing datasets in aesthetic and quality metrics, demonstrating superior perceptual alignment and visual fidelity. This positions SAGI-D as a superior resource for advancing detection models and sets a new standard for benchmarking in the field.

Failure Cases While our framework demonstrates strong overall performance, certain limitations remain. A representative failure case for SAOR occurs when, despite the context provided, the LLM generates semantically incoherent prompts. For example, in the first row of Fig. 3, a hot air balloon is suggested as a replacement for a train object. For UGDA, the VLM can exhibit inconsistent behavior by

Model	Orig. Datasets	Orig. Imgs	Inp. Imgs	Inp. Mdls	Inp. Pipes	Double Inp.	Type	Resolution	QA Qlt ↑	QA Ae ↑	CS Qlt ↑	CS Ae ↑	AS ↑
CocoGlide[21]	1	512	512	1	✗	✗	AI	256×256	2.88	1.79	**0.68**	-1.27	5.40
TGIF[45]	1	3,124	74,976	3	✗	✗	AI	up to $1024p$	3.94	2.53	0.63	-0.69	5.65
SAGI-D (ours)	3	77,900	95,839	8	5	✓	AI/OR	up to $2048p$*	**4.06**	**2.84**	**0.68**	**0.27**	**5.69**

Table 4. Comparison of inpainting datasets characteristics. Our dataset surpasses existing ones in scale (number of images), diversity (source datasets, models, pipelines). Resolution varies based on source dataset. AI: AI-Generated Content, OR: Object Removal.

fixating on specific details while ignoring obvious artifacts. As shown in the second row of Fig. 3, in the first permutation, the VLM insists on "couch duplication" and fails to recognize clear blurring artifacts, whereas in the second permutation, it correctly identifies the same blur. These cases highlight the dependency of our approach on the underlying capabilities and biases of foundation models. More examples are discussed in the supplementary material.

4.2. Inpainting Detection Benchmark

To establish a comprehensive benchmark for the presented SAGI-D dataset, we evaluate the performance of several state-of-the-art image forgery detection models.

Problem definition. Given an RGB image $x^{rgb} \in \mathbb{R}^{(H \times W \times 3)}$, the inpainting detection model aims to predict a pixel-level inpainting localization mask $\hat{y}^{loc} \in (0,1)^{(H \times W \times 1)}$ and/or an image-level inpainting detection probability $\hat{y}^{det} \in (0,1)$. The former will be referred to as the localization task, and the latter as the detection task.

Forensics models. We evaluated four inpainting detectors: PSCC-Net [38], CAT-Net [33], TruFor [21], and MMFusion [71]. CAT-Net provides only pixel-level masks, so we used the maximum probability from these masks for image-level detection. The other models output both pixel localization and image-level detection probabilities.

Training protocol. We evaluated the pretrained models, the versions retrained on the SAGI-D and on TGIF [45] following the training protocol on the original papers.

Implementation Details. We retrained all models from scratch, following the training protocol outlined in their original papers. CAT-Net was trained on an NVIDIA A100 GPU, while PSCC-Net, TruFor, and MMFusion were trained on an NVIDIA RTX 4090.

Evaluation metrics. We evaluated model performance at image and pixel levels using distinct metrics. With the positive class representing inpainted regions, for image-level detection, we use accuracy, and for pixel level we measure Intersection over Union (IoU), with a threshold of 0.5. To compare models across datasets, we used the threshold-agnostic Area Under the Curve (AUC) metric. For localization AUC, we resized and flattened localization maps and ground truths into vectors for ROC computation (see supplementary for additional results, qualitative comparisons, and human performance analysis).

4.3. Localization and Detection Results

Figure 3. Failure cases for SAOR and UGDA. Top row shows SAOR generating semantically inappropriate content (a colorful hot air balloon floating in the sky). Bottom row demonstrates UGDA's failure: 1st assessment gets fixated on "unnatural duplication of couch" without recognizing the blur, while 2nd assessment correctly identifies "couch appears altered with blurry texture".

We evaluate SAGI-D using four state-of-the-art inpainting detection models: PSCC-Net, CAT-Net, TruFor, and MMFusion. We assess both pre-trained models and versions retrained on our dataset. Table 5 presents both the localization (IoU) and detection (accuracy) results for in-domain and out-of-domain testing sets and SP and FR images. Retraining on SAGI-D leads to significant performance improvements across all models. TruFor shows the largest gains, with IoU increasing by +52.5% (in-domain) and +46.4% (out-of-domain), while CAT-Net achieves the highest in-domain IoU (69.2%, +29.9% improvement). FR regions remain challenging for original models (IoUs between 5.9% and 19.6%), but retrained models show substantial improvements, with TruFor reaching 72.4% IoU (+52.8%). SP detection is easier, with retrained TruFor achieving 86.8% IoU. TruFor and PSCC-Net maintain consistent cross-domain performance, while CAT-Net and MMFusion exhibit variability due to their designs (e.g., JPEG-specific artifacts and added input complexities). Overall, SAGI-D proves highly effective for improving image forgery detection and localization models.

To further assess SAGI-D as a benchmark, we compare models trained on their original data, TGIF and SAGI-D,

Data	Model	Mean IoU								Accuracy							
		ID		OOD		SP		FR		ID		OOD		SP		FR	
Original	CAT-Net[33]	39.3		20.9		41.4		5.9		59.8		42.4		84.5		39.9	
	PSCC-Net[38]	35.1		24.0		20.1		9.2		59.9		53.0		40.4		33.4	
	MMFusion[71]	19.6		19.1		46.3		16.9		61.7		63.4		62.9		25.5	
	TruFor[21]	12.2		19.1		40.0		19.6		58.2		60.1		41.0		12.5	
SAGI-D	CAT-Net[33]	69.2	+29.9	23.0	+2.1	46.5	+5.1	36.2	+30.3	98.8	+39.0	53.0	+10.6	99.9	+15.4	99.9	+60.0
	PSCC-Net[38]	58.1	+23.0	57.6	+33.6	43.2	+23.1	19.6	+10.4	63.4	+3.5	71.8	+18.8	50.1	+9.8	35.6	+2.2
	MMFusion[71]	63.7	+44.1	41.4	+22.3	73.1	+26.8	50.3	+33.4	92.8	+31.1	80.9	+17.5	85.9	+23.0	81.5	+56.0
	TruFor[21]	64.7	+52.5	65.5	+46.4	86.8	+46.8	72.4	+52.8	96.2	+38.0	92.8	+32.7	95.5	+54.5	94.4	+81.9

Table 5. Evaluation of SAGI on training image forgery detection and localization models. The first column indicates training data, while Mean IoU is reported for localization and Accuracy for detection. Results include in-domain (ID) and out-of-domain (OOD) performance for spliced (SP) and fully-regenerated (FR) images. ID and OOD performance is computed across inpainted and authentic images, while SP and FR highlight the performance on the corresponding type of inpaintings. Green numbers show improvements over original models. Retraining on SAGI-D significantly enhances performance, underscoring the value of semantic alignment for high-quality synthetic data.

Data	Model	AUC (det)				AUC (loc)			
		TGIF		*SAGI-D*		TGIF		*SAGI-D*	
Original	CN [33]	75.1		53.2		74.9		56.3	
	MM [71]	86.5		75.3		67.4		67.3	
	PS [38]	73.8		62.7		39.0		66.0	
	TF [21]	89.4		79.5		70.5		67.0	
	Average	**81.2**		**67.7**		**63.0**		**64.2**	
TGIF	CN[33]	69.5	-5.6	68.1	+14.9	81.0	+6.1	62.6	+6.3
	MM [71]	95.9	+9.4	82.2	+6.9	93.7	+26.3	82.1	+14.8
	PS [38]	73.5	-0.2	69.6	+7.0	60.9	+21.9	49.6	-16.4
	TF [21]	58.4	-31.0	74.4	-5.1	95.0	+24.5	89.7	+22.7
	Average	**74.3**	-6.9	**73.6**	+5.9	**82.7**	+19.7	**71.0**	+6.8
SAGI-D	CN [33]	84.6	+9.5	92.7	+39.4	73.2	-1.7	84.9	+28.6
	MM [71]	71.9	-14.6	97.1	+21.8	54.6	-12.8	95.9	+28.6
	PS [38]	65.5	-8.3	82.9	+20.2	83.5	+44.6	75.9	+9.9
	TF [21]	91.0	+1.6	99.5	+20.0	91.0	+20.5	98.8	+31.8
	Average	**78.3**	-2.9	**93.1**	+25.4	**75.6**	+12.6	**88.9**	+24.7

Table 6. Performance comparison of image forensics methods CAT-Net (CN), MMFusion (MM), PSCC-Net (PS), and TruFor (TF) across TGIF and SAGI-D datasets. Metrics include detection and localization AUC. *Data* indicates training source. Green and red numbers show performance improvements and decreases over original models. Retraining on SAGI-D yields more consistent improvements, even across datasets, compared to training on TGIF.

and evaluate them on both TGIF and SAGI-D test sets. Tab. 6 shows localization and detection AUC results. Training on SAGI-D consistently improves performance, outperforming TGIF-trained models. Notably, SAGI-D-trained models sometimes surpass TGIF-trained ones even on TGIF (78.3 vs 74.3 detection AUC average). CAT-Net trained on SAGI-D achieves +28.6 and +21.8 improvements on SAGI-D and TGIF respectively, while TGIF-trained CAT-Net drops -5.6 on TGIF. PSCC-Net shows +44.6 and +20.2 localization AUC gains, and SAGI-D-trained TruFor reaches AUCs above 90 on both datasets. Original models perform better on TGIF than SAGI-D, highlighting its greater complexity and benchmark value. These results affirm the superiority of SAGI-D for training detection models.

5. Conclusions

In this work we proposed a model-agnostic framework for generating and evaluating high-quality inpaintings, based on the key ideas of semantic alignment and uncertainty guidance. Looking ahead, our framework can benefit from advances in foundation models to further refine prompt generation and realism assessment and extend uncertainty mechanisms to enhance LLM/VLM performance in tasks involving reference samples, such as image-to-image translation, audio/video editing, and quality assessment of generated content. We aim for this framework to advance both generative quality and forensic detection, supporting safer applications of generative AI.

Limitations: The limited capacity and finite training data of any pre-trained LLM and VLM impose inherent constraints on tasks like semantic alignment and realism assessment. Despite these limitations, we consider them the best approximations available for bridging the gap between the real world and computer-generated visual representations.

Ethical Concerns: While advances in generative AI provide significant societal benefits by automating previously labor-intensive tasks, they also enable potential malicious exploitation, such as spreading misinformation, or manipulating visual evidence. Yet, the fact that our proposed framework can significantly enhance the robustness of image forensic tools, allows mitigating such misuse risks.

Acknowledgments: This work was supported by the Horizon Europe projects vera.ai (grant no. 101070093), ELLIOT (101214398) and AI4TRUST (101070190). Compute resources were granted by GRNET and the HPC infrastructure of the Aristotle University of Thessaloniki.

References

[1] Mistral AI. Mistral small 3.1: Multimodal language model. https://mistral.ai/news/mistral-small-3-1, 2025. Model version 2503, 24B parameters, Apache 2.0 license. 6

[2] Irene Amerini, Lamberto Ballan, Roberto Caldelli, Alberto Del Bimbo, and Giuseppe Serra. A sift-based forensic method for copy–move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, 6(3):1099–1110, 2011. 3

[3] Anthropic. Claude 3.5 sonnet, 2024. Large Language Model. 4

[4] Anthropic. Claude 3.7 sonnet, 2025. Hybrid reasoning large language model. 6

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, et al. Qwen2.5-vl technical report, 2025. 6

[6] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. 3

[7] Y. Bengio and Yann Lecun. Convolutional networks for images, speech, and time-series, 1997. 2

[8] M. Bertalmio. Contrast invariant inpainting with a 3rd order, optimal pde. In *IEEE International Conference on Image Processing 2005*, pages II–778, 2005. 2

[9] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 417–424, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[10] M. Bertalmio, A.L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–I, 2001. 2

[11] Hessen Bougueffa, Mamadou Keita, Wassim Hamidouche, Abdelmalik Taleb-Ahmed, Helena Liz-López, Alejandro Martín, David Camacho, and Abdenour Hadid. Advances in ai-generated images and videos. *International Journal of Interactive Multimedia and Artificial Intelligence*, 9(1):173–208, 2024. 1

[12] Tony F. Chan and Jianhong Shen. Nontexture inpainting by curvature-driven diffusions. *Journal of Visual Communication and Image Representation*, 12(4):436–449, 2001. 2

[13] I-Cheng Chang, J. Cloud Yu, and Chih-Chuan Chang. A forgery detection algorithm for exemplar-based inpainting images using multi-region relation. *Image and Vision Computing*, 31(1):57–71, 2013. 2

[14] Ziyi Chang, George Alex Koulieris, and Hubert P. H. Shum. On the design fundamentals of diffusion models: A survey, 2023. 2

[15] Yiqun Chen and James Zou. Twigma: A dataset of ai-generated images with metadata from twitter, 2023. 3

[16] Vincent Christlein, Christian Riess, Johannes Jordan, Corinna Riess, and Elli Angelopoulou. An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on Information Forensics and Security*, 7(6):1841–1854, 2012. 3

[17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 6

[18] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, 2004. 2

[19] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: a raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, page 219–224, New York, NY, USA, 2015. Association for Computing Machinery. 1, 4

[20] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pages 422–426, 2013. 3

[21] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, 2023. 2, 3, 7, 8

[22] Qiang Guo, Shanshan Gao, Xiaofeng Zhang, Yilong Yin, and Caiming Zhang. Patch-based image inpainting via two-stage low rank approximation. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):2023–2036, 2018. 2

[23] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5, 2022. 6

[24] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation, 2022. 4

[25] Kyong Hwan Jin and Jong Chul Ye. Annihilating filter-based low-rank hankel matrix approach for image inpainting. *IEEE Transactions on Image Processing*, 24(11):3498–3511, 2015. 2

[26] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion, 2024. 5

[27] Xiaodan Ju et al. Brushnet: Plug-and-play image inpainting with user guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5678–5687, 2024. 5

[28] Dimitrios Karageorgiou, Giorgos Kordopatis-Zilos, and Symeon Papadopoulos. Fusion transformer with object mask guidance for image forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4345–4355, 2024. 2

[29] Norihiko Kawai, Tomokazu Sato, and Naokazu Yokoya. Diminished reality based on image inpainting considering background geometry. *IEEE Transactions on Visualization and Computer Graphics*, 22(3):1236–1247, 2016. 2

[30] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining, 2023. 3

[31] Nitish Kumar and Toshanlal Meenpal. Semantic segmentation-based image inpainting detection. In *Innovations in Electrical and Electronic Engineering*, pages 665–677, Singapore, 2021. Springer Singapore. 2

[32] Alina Kuznetsova et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 4

[33] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 2, 7, 8

[34] Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8300–8309, 2019. 2

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 4

[36] Zaoshan Liang, Gaobo Yang, Xiangling Ding, and Leida Li. An efficient forgery detection algorithm for object removal by exemplar-based image inpainting. *Journal of Visual Communication and Image Representation*, 30:75–85, 2015. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 1, 4

[38] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 2, 7, 8

[39] Ming Lu and Shaozhang Niu. A detection approach using lstm-cnn for object removal caused by exemplar-based image inpainting. *Electronics*, 9(5), 2020. 2

[40] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models, 2023. 1

[41] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. Is it ai or is it me? understanding users' prompt journey with text-to-image generative ai tools. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2024. Association for Computing Machinery. 1, 3

[42] Gaël MAHFOUDI, Badr TAJINI, Florent RETRAINT, Frédéric MORAIN-NICOLIER, Jean Luc DUGELAY, and Marc PIC. Defacto: Image and face manipulation dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019. 3

[43] Ara Manukyan. Hd-painter: High-resolution prompt-faithful text-guided image inpainting, 2024. 5

[44] Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. Tgif: Text-guided inpainting forgery dataset, 2024. 3

[45] Hugo Mareen et al. Tgif: A text-guided image forgery dataset for deepfake detection. *arXiv preprint arXiv:2401.56789*, 2024. 7

[46] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization, 2024. 3

[47] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 3

[48] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 3

[49] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 2

[50] OpenAI. Chatgpt-3.5, 2023. 4

[51] OpenAI. Chatgpt-4, 2023. 5, 6

[52] A Pasupathy. The neural basis of image segmentation in the primate brain. *Neuroscience*, 296:101–109, 2015. 3

[53] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 2

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 5

[55] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, pages 8821–8831, 2021. 1

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2, 5

[59] Shachar Rosenman, Vasudev Lal, and Phillip Howard. Neuroprompts: An adaptive framework to optimize prompts for text-to-image generation, 2024. 3

[60] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, and Raphael Gontijo Lopes. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 6

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[63] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[64] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. 2, 5

[65] Matías Tailanian, Marina Gardella, Álvaro Pardo, and Pablo Musé. Diffusion models meet image counter-forensics, 2024. 5

[66] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8): 3998–4011, 2018. 3

[67] Yuying Tang, Ningning Zhang, Mariana Ciancia, and Zhigang Wang. Exploring the impact of ai-generated image tools on professional and non-professional users in the art and design fields, 2024. 3

[68] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, et al. Gemma 3 technical report, 2025. 6

[69] Lucas Theis. What makes an image realistic?, 2024. 3

[70] Dijana Tralic, Ivan Zupancic, Sonja Grgic, and Mislav Grgic. Comofod — new database for copy-move forgery detection. In *Proceedings ELMAR-2013*, pages 49–54, 2013. 3

[71] Konstantinos Triaridis and Vasileios Mezaris. Exploring multi-modal fusion for image manipulation detection and localization, 2023. 2, 7, 8

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[73] Luisa Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 2

[74] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 3, 6

[75] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 3

[76] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2022. 2, 3

[77] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels, 2023. 3, 6

[78] Pin-Yu Wu, C-C Jay Kuo, et al. Defakehop: A light-weight high-performance deepfake detector. *IEEE Transactions on Multimedia*, 24:343–353, 2022. 2

[79] Qiong Wu, Shao-Jie Sun, Wei Zhu, Guo-Hui Li, and Dan Tu. Detection of digital doctoring in exemplar-based inpainted images. In *2008 International Conference on Machine Learning and Cybernetics*, pages 1222–1226, 2008. 2

[80] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. Loki: A comprehensive synthetic data detection benchmark using large multimodal models, 2024. 4, 5

[81] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention, 2018. 2

[82] Ning Yu, Xiang Zhao, and Bo Chen. Inpaint-anything: Segment meets inpaint. *arXiv preprint arXiv:2304.06790*, 2023. 5

[83] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting, 2023. 5

[84] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative ai: A survey, 2024. 1

[85] Dengyong Zhang, Zaoshan Liang, Gaobo Yang, Qingguo Li, Leida Li, and Xingming Sun. A robust forgery detection algorithm for object removal by exemplar-based image inpainting. *Multimedia Tools and Applications*, 77(10):11823–11842, 2018. 2

[86] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 5

[87] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14071–14081, 2023. 3

[88] Yushu Zhang, Zhibin Fu, Shuren Qi, Mingfu Xue, Zhongyun Hua, and Yong Xiang. Localization of inpainting forgery with feature enhancement network. *IEEE Transactions on Big Data*, 9(3):936–948, 2023. 2

[89] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. 3

[90] Xinshan Zhu, Yongjun Qian, Xianfeng Zhao, Biao Sun, and Ya Sun. A deep learning approach to patch-based image inpainting forensics. *Signal Processing: Image Communication*, 67:90–99, 2018. 2

[91] Xinshan Zhu, Junyan Lu, Honghao Ren, Hongquan Wang, and Biao Sun. A transformer–cnn for deep image inpainting forensics. *The Visual Computer*, 39(10):4721–4735, 2023. 2

[92] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting, 2024. 5