A Semantically Impactful Image Manipulation Dataset: Characterizing Image Manipulations using Semantic Significance

Yuwei Chen[1], Ming-Ching Chang[1], Mattias Kirchner[2], Zhenfei Zhang[1], Xin Li[1]
Arslan Basharat[2], Anthony Hoogs[2]
[1]University at Albany, 1400 Washington Ave, Albany, NY 12222
[2]Kitware, 1712 US-9, Clifton Park, NY 12065

Abstract

We investigate how to characterize semantic significance (SS) in detecting image manipulations (IMD) for media forensics. We introduce the Characterization of Semantic Impact for IMD (CSI-IMD) dataset, which focuses on localizing and evaluating the semantic impact of image manipulations to counter advanced generative techniques. Our evaluation of 10 state-of-the-art IMD and localization methods on CSI-IMD reveals key insights. Unlike existing datasets, CSI-IMD provides detailed semantic annotations beyond traditional manipulation masks, aiding in the development of new defensive strategies. The dataset features manipulations from advanced generation methods, offering various levels of semantic significance. It is divided into two parts: a gold-standard set of 1,000 manually annotated manipulations with high-quality control, and an extended set of 500,000 automated manipulations for large-scale training and analysis. We also propose a new SS-focused task to assess the impact of semantically targeted manipulations. Our experiments show that current IMD methods struggle with manipulations created using stable diffusion, with TruFor and Cat-Net performing the best among those tested. The CSI-IMD dataset will become available at https://github.com/csiimd/csiimd.

1. Introduction

Image manipulation detection (IMD) and localization has long been a key task in computer vision [6], but the practical application of traditional defensive models in real-world scenarios remains underexplored. Current research often relies on curated academic datasets that fail to reflect real-world contexts. With the rapid advancement of media generation and editing techniques [3, 14, 26, 37], the risks of disinformation and misinformation [29] have surged, as these powerful tools are now easily accessible, even on mobile devices, posing significant threats to online integrity and security. The lack of robust defensive measures alongside this increased accessibility has contributed to growing distrust in the online community. While the cybersecurity and digital forensics sectors are tackling these challenges, the forensics community is hampered by the absence of comprehensive datasets. Most publicly available digital forensics datasets (in Table 1) feature trivial manipulations, leaving many real-world issues unaddressed.

The current research of image manipulation detection and localization is still primitive. Most detection models solely generate pixel-based manipulation masks, which are inadequate for the needs of modern digital forensics. Critical questions remain unanswered when analyzing forged images in real-world scenarios, such as identifying the purposes of the manipulation, understanding how manipulations alter the overall image, and assessing the threat level posed. In security and forensics, merely identifying the type and location of manipulation (the *how* and *where*) is not enough. It is crucial to uncover the intent (the *why* and *what*) behind the manipulation through semantic threat assessment. This involves understanding the purpose behind the perpetrator's actions and identifying the specific harm intended. Figure 1 illustrates this core idea, which motivates the approach of our work.

Malicious actors often create harmful image manipulations, posing significant challenges for analysts who must assess the threat level in forged images. Unfortunately, existing publicly available datasets mostly feature trivial manipulations, making effective evaluation difficult. We advocate for the digital forensics community to analyze manipulated media not only to detect manipulation but also discern whether it is trivial or semantically impactful. To support this shift, we propose a new image manipulation localization dataset with a semantic significance (SS) ranking task. Our goal is to initiate research on assessing digital media manipulation based on semantic threats.

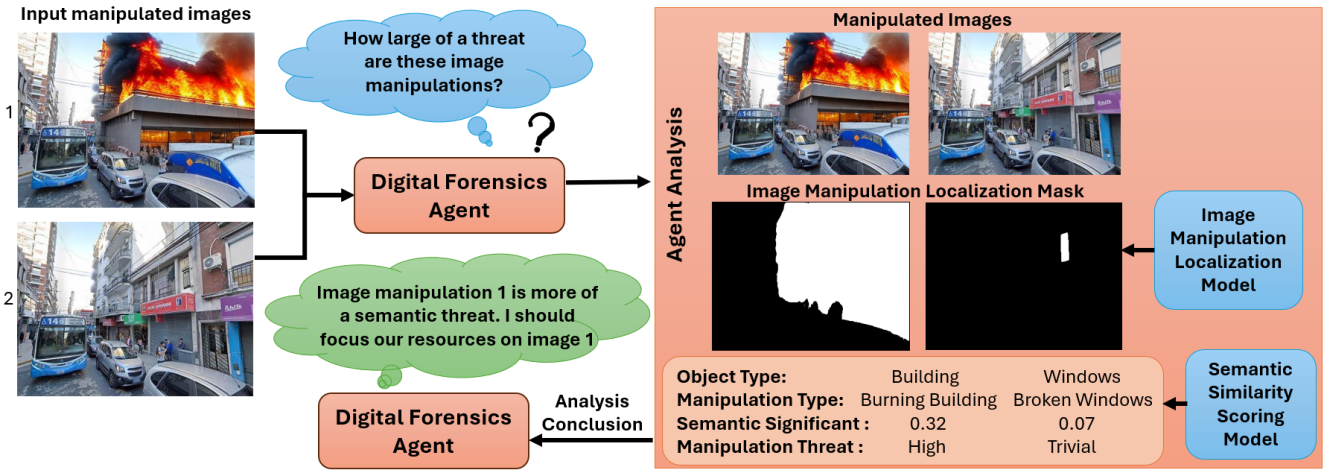

Figure 1. Overview of how the proposed **Characterization of Semantic Impact for Image Manipulation Detection (CSI-IMD)** dataset assists real-world digital IMD forensics.

We provide the following contributions in this paper:

- We investigate a new perspective of image forensics task focusing on evaluating image manipulation under semantic significance, with a benchmark dataset, insights and baseline results provided.
- The CSI-IMD Gold Standard set includes 1,000 manipulated and 1,000 pristine images with six types of annotations for image forensics and semantic significance. Additionally, the extended set offers 500,000 manipulated images with semantic impact for model training.
- We provide evaluation results and analysis on 10 SoTA image manipulation localization on the CSI-IMD dataset.

2. Related Works and IMD Datasets

Image manipulation datasets in computer vision and semantic forensics can be organized into two categories: *fully synthetic* and *partially manipulated* images. Fully synthetic images are typically generated using text-to-image models such as DALLE [26], or deep fake models like DeepFaceLab [25]. Partially manipulated images are altered using techniques such as PhotoShop, involving manipulation types like *splicing*, *copy-pasting*, *morphing*, and *inpainting*.

Traditionally, image manipulation detection and localization have overlooked the semantic implications of manipulations. The focus has mainly been on identifying and outlining manipulated regions, which influenced the design of early image manipulation datasets like Columbia [22] and CASIA [5]. These datasets involved inserting random objects into natural scenes with little regard for coherence or impact. The images typically consist of small thumbnails with simple scenes, minimal entities, and no post-processing enhancements, making them unrealistic compared to real-world instances of image tampering. As a result, many state-of-the-art (SoTA) image manipulation techniques achieve high detection accuracy on these datasets.

Subsequent datasets such as Coverage [32], NIST16 [7], IMD2020 [4], and CIMD [38] introduced more realistic natural image scenes by employing intricate post-processing techniques, making them more representative of real-world image manipulation. However, these datasets share a common limitation: they primarily focus on manipulations involving a single entity, typically a solitary object. The pristine images used often feature overly simplistic scenes with a single entity at the center, usually set against plain backgrounds, offering limited semantic context of the depicted scenes. Additionally, the size of these datasets is restricted by the manual tampering processes involved.

Moreover, the splices in these datasets often exhibit subpar quality, and there is a noticeable lack of diversity in manipulation types. These datasets also overlook the semantic context of manipulations and their impact on the overall image scene. There is little to no focused characterization or rationale behind the choice of specific manipulations, leading to alterations that may appear arbitrary or inconsequential. Consequently, these datasets lack the depth and complexity needed to provide insights into sophisticated and contextually relevant image manipulations that mirror real-world scenarios.

In the past, manipulation techniques primarily involved manual alterations using software like Photoshop, often with systematic, random object insertions. However, recent advancements in computer vision technology have enabled malicious actors to leverage foundation models, such as Large Language Models (LLMs) and generative models like DALL-E [26] and stable diffusion models [27], to create much larger and more complex image manipulation localization datasets.

Dataset	Year	# Manipulated Images	# Pristine Images	Image Size	Format	Manipulation Method
Columbia (Gray) [22]	2004	912	933	128×128	BMP	Random
Columbia (Color) [11]	2006	180	183	757×568 - $1,152 \times 768$	TIF	Random
CASIA v1 [5]	2013	921	800	374×256	JPEG	Manual
CASIA v2 [5]	2013	5,123	7,200	320×240 - 800×600	JPEG	Manual
Coverage [32]	2016	100	100	400×486	TIF	Manual
NIST16 [7]	2016	564	875	500×500 - $5,616 \times 3,744$	TIF	Manual
Realistic Tampering [16]	2016	220	220	$1,920 \times 1,080$	TIF	Manual
IMD2020 [4]	2020	2,010	414	$1,062 \times 866$	JPEG,PNG	Internet
AutoSplice [13]	2023	3,621	2,273	256×256 - $4,232 \times 4,232$	JPEG	LLI Model
CIMD [38]	2024	400	400	$2,048 \times 1,365$	JPEG,TIF	Manual
CSI-IMD Golden Standard (Ours)	2024	1,000	1,000	400×296 - 600×800	JPEG,PNG	Stable Diffusion
CSI-IMD Extended (Ours)	2024	500,000	0	400×296 - 600×800	PNG	Stable Diffusion

Table 1. Comparison of CSI-IMD with existing mainstream image manipulation detection and localization datasets with details.

3. The CSI-IMD Dataset

We create partial image manipulations using generative stable diffusion techniques [28, 30] to produce semantically impactful examples that reflect real-world scenarios. § 3.1 and § 3.2 outlines our dataset creation method and dataset statistics. § 3.3 illustrates how we analyze the image manipulations based on semantic significance. § 3.4 shows the different annotation types in the dataset. § 3.5 explains the methods for generating semantic significance (SS) scores.

3.1. Dataset creation

Figure 2 illustrates our CSI-IMD dataset creation process, which starts with a natural image scene where object detectors and semantic segmentation are applied to identify relevant objects. We use a custom version of YOLOv5 for object detection and the Segment Anything Model [15] (SAM) from Meta Research for semantic segmentation. An object of interest is then automatically selected based on detected objects, cross-referenced with the corresponding segmentation region using an 80 percent overlap threshold. If no suitable objects are found, the image is discarded. The cross-referenced segmentation region is then chosen as the input for manipulation.

For generating manipulations, we experimented with *stable diffusion* models, specifically v1runwayml/stable-diffusion-v1-5 [28] and stabilityai/stable-diffusion-2-inpainting [30] from the Hugging Face library. These models were chosen because generative multi-modal models likely represent the future of image manipulation and require minimal manual interaction, making them ideal for large-scale automated processes. Our qualitative analysis showed that stable diffusion v2 generally produces higher-quality manipulations that blend seamlessly with the surrounding image regions, leaving little to no visual traces. However, this approach has a trade-off: the generated manipulations must remain plausible within the scene's context, or the model may fail to generate the desired object.

Textual prompt engineering: The process of selecting prompts for the generative model involves choosing both positive and negative prompts. Negative prompts are sourced from a predefined list known for producing good outcomes. For positive prompts, one approach is rule-based, selecting manipulations from a list tailored to the object of interest (as outlined in Table 3). For example, applying a 'burning' manipulation to a house (shown in Figure 1) results in a manipulated image with high semantic significance. While limited in variety, this method ensures that each manipulation applied varies in levels of semantic significance.

We utilized 16 types of manipulations in the gold standard set, categorized into three levels of semantic significance (SS). This gold standard set is curated by humans, featuring higher quality manipulations and more detailed annotations compared to the extended set. Some manipulations, like flooding a scene, do not require a specific object of interest. An area of the image is preselected. Table 3 outlines the manipulation types by SS level. For each original image, five manipulations with varying significance were applied to create the dataset. These manipulation types were chosen for their clarity in manipulation significance.

3.2. Dataset statistics

The CSI-IMD dataset comprises two distinct sets: a manually filtered **gold standard set** and an fully automatically generated **extended set**. The gold standard set contains 1,000 pristine images and 1,000 manipulated images, curated through human filtering for experimental analysis and deeper insights characterizing the IMD dataset. In contrast, the extended set contains 500,000 manipulated images generated entirely by our automated pipeline. The extended set contains fewer image annotations aiming to provide a large-scale image manipulation localization dataset featuring semantically impactful manipulations.

All pristine images used for manipulation and testing are

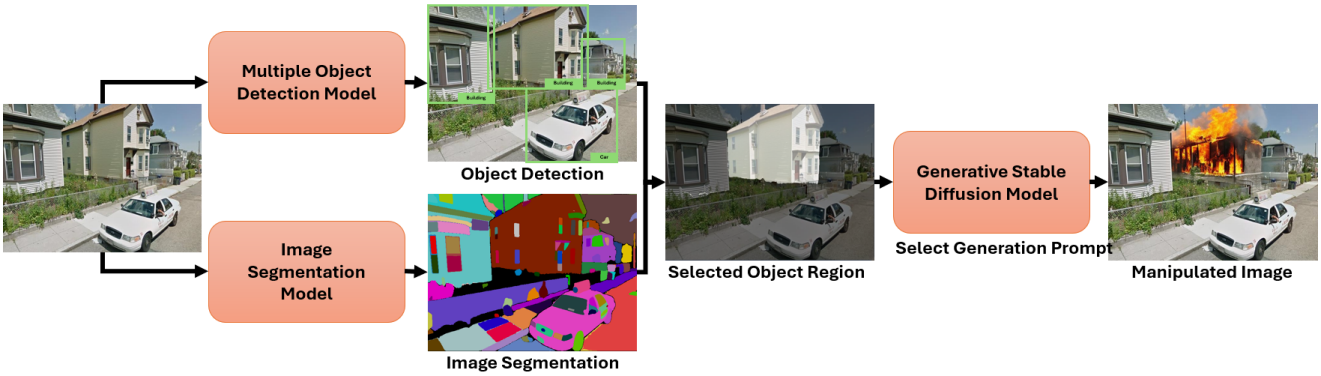

Figure 2. Overview pipeline of our proposed image manipulation generation process.

gathered from publicly available datasets, including GSV-Cities [1] and a Kaggle road vehicle images dataset [36]. These images primarily depict city and urban street scenes. The pristine images are in JPG format, while the manipulated images are in PNG format.

For the gold standard set, we select 200 natural image scenes, resulting in 1,000 manipulated images. These 200 pristine images serve as references for researchers to compare with the manipulated images. These 200 pristine images are separate from the 1,000 pristine images used for evaluation in the manipulation detection task. Annotations for the gold standard set include original pristine image name, image ID, manipulation prompts and numbers, original dataset source, **semantic significance** (SS) scores from each similarity model, and manipulation SS rankings. Table 2 list the provided annotations and ground truths in the CSI-IMD gold standard set.

3.3. Structured semantic analysis of manipulations

We implement a structured approach to analyze the *semantic significance* (SS) of manipulations. Each natural image scene undergoes **five** distinct manipulation prompts, each with varying levels of SS. By using predefined manipulation prompts tailored to the objects of interest, we maintain control over the SS of each manipulation. This ensures that at least one is semantically trivial and another is highly significant.

This uniform application of five manipulations per image scene allows for a direct comparative analysis of the effects within the same context. By isolating the manipulation as the only variable, we can closely examine its semantic significance, providing deeper insights into the significance of each manipulation.

The SS categories were created to prevent penalizing methods for incorrectly ranking manipulations of similar significance. Within each category, manipulations can be ranked in any order without penalty, as long as those with higher or lower significance are ranked correctly relative to one another. This approach ensures that methods are evaluated based on their ability to accurately rank manipulations by their true significance, rather than relying solely on numerical scores.

For a visual example, refer to Figure 4. Observe that the LLMScore method ranked the manipulations sign swap and broken windows differently from other methods. However, since both manipulations are categorized as trivial or low significance, the LLM Score method is not penalized for this variation. All methods correctly identified these manipulations as having the lowest SS compared to the other three manipulations.

3.4. Generating semantically impactful manipulations with SS-relevant annotations

Annotation types: The CSI-IMD dataset includes *six* distinct annotation types detailed in Table 2. The first three are standard in most image manipulation datasets: a binary manipulation flag, the manipulated object class, and a manipulation mask indicating the altered region. The other three notations set our dataset apart, which emphasize the semantic implications of manipulations: a semantically relevant pixel-level manipulation mask, a SS score ranging from 0-1, and a SS ranking ranging from 1-5. These additional annotations aim to support further research in semantic analysis within the image manipulation forensics community.

Given the importance of highlighting the semantic significance of manipulations, we carefully selected additional annotations that would be most relevant. Since we employ generative stable diffusion, the manipulated object may not occupy the entire manipulated region. Therefore, we provide an enhanced pixel-level manipulation mask that specifically highlights the manipulated pixels directly related to the manipulation class, in addition to the standard manipulation mask. This enhanced mask is generated using the Segment Anything Model (SAM) [15] applied to the manipulated region. If a single segmentation mask is obtained, it is used as the enhanced manipulation region. If multiple masks are present, we select the largest one covering more than 70% of the region or combine smaller segments that do

Image Manipulation Ground Truth	Annotation Description
Binary manipulation flag	Indicating if the image is manipulated
Manipulation object class	The object class of the spliced manipulation
Manipulation mask	A pixel-level mask showing all pixels that have been manipulated
Semantically relevant manipulation mask	A pixel-level mask highlighting all pixels related to the object class
Semantic significance (SS) score	A score from 0 to 1 measuring the extent of semantic change in the scene
Semantic significance (SS) ranking	A score from 1 to 5 indicating the manipulation's SS for each image scene

Table 2. Overview of the image manipulation ground truth with annotation descriptions in the CSI-IMD gold standard set.

Low	Medium	High
Car swap	Destroyed car	Burning building
Truck swap	Police car	Destroyed building
Shirt swap	Protesters with signs	Burning car
Sign swap	Police officer	Military tank
Tree swap	N/A	Military officer
Flag swap	N/A	Flood

Table 3. List of manipulation types in the CSI-IMD gold standard set, categorized by the semantic significance (SS).

not overlap with the manipulation region.

For example, consider a manipulation that adds protesters to an empty street in front of a police station. The original image shows an empty street, which is altered to include protesters. The input region for manipulation is the street, with the manipulation prompt being the addition of protesters. This scenario highlights the difference between two types of pixel-level manipulation masks. The standard manipulation mask includes all manipulated pixels (both the road and the protesters). In contrast, the semantically relevant manipulation mask contains only the pixels representing the protesters, as they are directly relevant to the manipulation prompt. The road pixels, although altered, are semantically equivalent to those in the original image.

3.5. Semantic significance score generation

The semantic significance (SS) score assesses the change in context or meaning of an image scene after manipulation. To assess the feasibility of existing scoring methods to capture the semantic significance of a manipulation, We explored the following three methods for calculating semantically focused scores. These methods are illustrated in Figure 3.

(1) Image-to-image score generation: To compute semantic similarity between the original pristine image and its manipulated version, we use a model to compare the two images. The resulting semantic similarity score is then adjusted using a *semantic deviation function*, which focuses on measuring the impact of the manipulation by anchoring the comparison to the pristine image. Details of this adjustment process are shown in Figure 3.

(2) Image-to-text score generation: To assess semantic similarity between an image and a textual description, we first use ChatGPT [24] or LLava [19] to generate a short description of the manipulated image. Then methods such as LLMScore [21] then calculates a semantic similarity score between the original pristine image and this description. This score is adjusted using the semantic deviation function to produce the final SS score. The underlying idea is that a more impactful manipulation will result in a lower similarity score between the original image and the manipulated description, justifying the need for this adjustment.

(3) Text-to-text score generation: To evaluate semantic similarity between textual descriptions, we generate descriptions for both the original and manipulated images using ChatGPT [24] or LLava [19]. We then apply the RoBERTa model, specifically, twitter-roberta-base-sentiment-latest [23] from the Hugging Face library to assign sentiment scores to these descriptions, capturing their emotional tone. The SS is quantified by calculating the absolute difference between the sentiment scores of the original and manipulated images. This approach is particularly valuable for images involving sensitive topics, such as a forgery photo used in a police brutality fake news, where the original image may carry a strongly negative sentiment. Using the sentiment score of the original image as a baseline allows us to measure the semantic shift caused by the manipulation, effectively assessing its overall impact.

It is important to note that these numeric SS scores are not absolute indicators of a manipulation's significance. Rather, they reflect the extent to which the overall meaning of the semantic scene has deviated from the original image. As such, a SS score of 0.4 should not be interpreted as being twice as significant as a score of 0.2. In light of this, we have implemented measures to best mitigate the affects of concrete scores within the newly proposed SS ranking task. In § 3.3, we outlined the steps taken to ensure that the only difference between the pristine and manipulated images is the manipulation itself.

4. Experimental Evaluation

We conducted extensive experiments on the newly introduced CSI-IMD dataset to benchmark mainstream image manipulation localization methods, focusing on the semantic aspects of image forgery detection and localization. Section § 4.1 contains details of the evaluation set up for tradi-

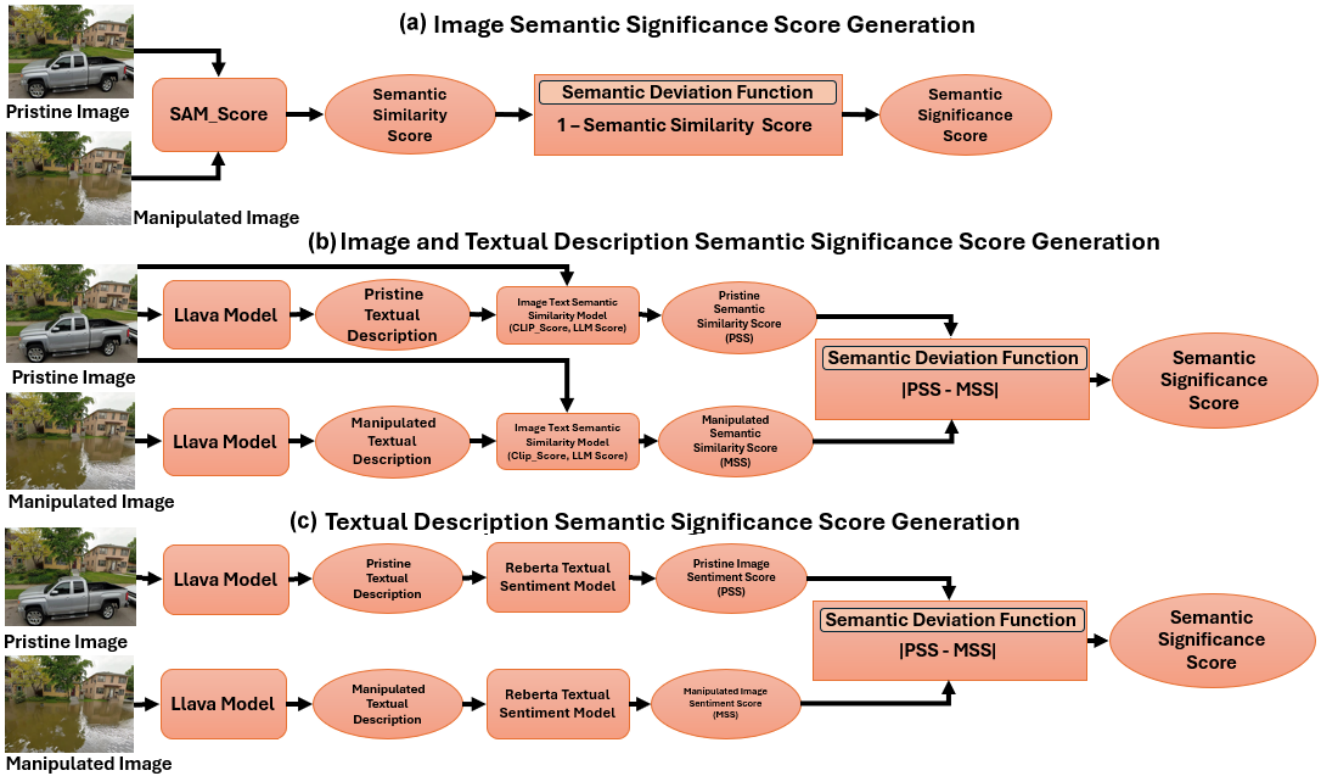

Figure 3. Pipeline showing the three different approaches we explored on generating **semantic significance (SS)** scores. (a) **Image-based:** Use a semantic similarity model to compare the original and manipulated images. (b) **Image-to-text description:** Convert images to textual descriptions and apply image-text similarity models such as LLMScore to assess the SS. (c) **Text-based SS analysis:** Since meaningful manipulations often alter the sentiment of the image topic, we measure SS by analyzing changes in sentiment.

tional pixel based image manipulation localization. Section § 4.2 contains details of the evaluation set up for our proposed SS assessment. Section § 4.3 contains quantitative results and analysis of the two evaluation assessments. Section § 4.4 contains discussion based topics such as generative AI ethics and limitations of the existing dataset.

Our evaluation metrics include average IoU, accuracy, precision, recall, and F1 score to measure the complexity of manipulations generated by stable diffusion models. These metrics measure the difficulty of image manipulations carried out by generative stable diffusion models. To calculate average precision and recall, we conducted a pixel-based evaluation for each manipulated image, producing a confusion matrix to calculate the precision and recall. These values were then averaged across all images to obtain the final scores. Pixel-based IoU was calculated by dividing the overlap between the predicted region and the ground truth by the union of those regions, while accuracy was measured as the overlap over the ground truth region.

4.1. Evaluation of manipulation localization

To evaluate image manipulation localization, we utilized ten state-of-the-art (SoTA) models outlined in Table 4 who are known for their high performance on publicly available image manipulation localization datasets outlined in Table 1. The evaluation was conducted at the pixel level, highlighting the SS of image manipulation detection, with quantitative results presented in Table 4. To ensure consistency and fairness among all methods, a fixed threshold of 0.5 was applied across all methods when converting outputs

IMD Methods	IoU	Acc.	Prec.	Recall	F1
ManTraNet [34]	0.009	0.010	0.176	0.010	0.020
CAT-Net v2 [17]	0.614	0.772	0.671	0.772	0.718
TruFor [8]	0.652	0.786	0.713	0.762	0.737
CRCNN [35]	0.002	0.004	0.060	0.003	0.007
HiFi-Net [9]	0.018	0.024	0.117	0.024	0.040
IF-OSN [33]	0.041	0.075	0.204	0.076	0.110
ObjectFormer [31]	0.006	0.008	0.035	0.008	0.012
PSCC-Net [20]	0.139	0.226	0.239	0.226	0.232
RRU-Net [2]	0.102	0.180	0.278	0.179	0.218
SPAN [12]	0.000	0.000	0.007	0.000	0.000

Table 4. Pixel-level image manipulation localization results using only manipulated images from the gold standard set. A fixed threshold of 0.5 is used when converting to a binary mask. The results for SPAN is 0 percent as the prediction localization masks were either empty or contains incorrect small regions.

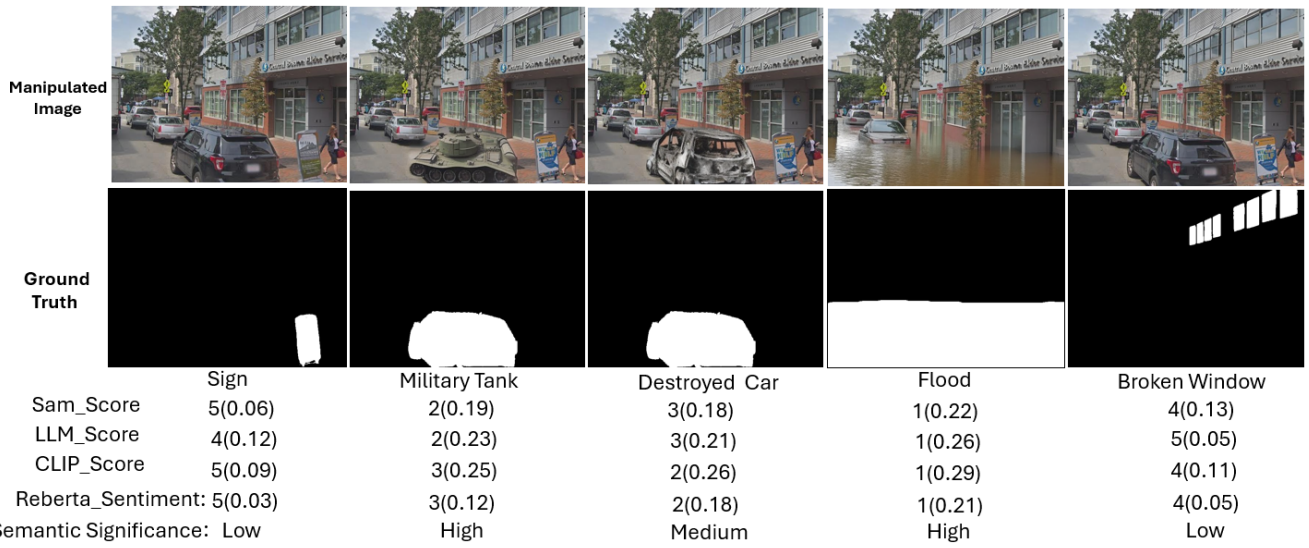

Figure 4. A visual example showing dataset annotations and our proposed semantic significance (SS) detection task. Five manipulations with varying SS are applied to a single image scene, with SS scoring methods displayed alongside their scores and rankings. The visualization illustrates an analyst's process of evaluating these manipulations using each scoring method.

to binary masks.

A high-level analysis of the results reveals that our proposed dataset poses significantly greater challenges compared to previous datasets. As a result, most mainstream methods struggle to accurately localize the manipulations generated by stable diffusion models.

4.2. Semantic significance ranking

We introduce a manipulation semantic significance (SS) ranking task to evaluate the semantic importance of manipulations in the dataset. In each contextual scene, manipulations are ranked by their significance. annotated rankings are manually established for all 200 natural image scenes. Semantic similarity scoring methodologies are then used to generate corresponding rankings based on the scores they produce. As outlined in the § 3.3, the manipulation prompts are designed to ensure that each scene includes manipulations of *trivial*, *moderate*, and *high* significance.

Semantic Significance assessment: To evaluate the SS of image manipulations in the CSI-IMD benchmark, we explored various methods for scoring the semantic similarity. These methods fall into three categories: (1) **Image-based scores:** Methods that compute semantic scores solely from images. (2) **Image and text scores:** Methods that use both images and brief textual descriptions to derive scores. (3) **Text-only scores:** Methods that rely exclusively on textual features for scoring. Details of these methods are provided in § 3.5. Our goal is to offer the research community a structured baseline and benchmark tools (dataset with annotations) to analyze digital media manipulations based on their SS and threat level.

The ranking of manipulation SS scores: We introduce a novel task for ranking image manipulations based on their semantic significance. To ensure the evaluation focuses solely on the significance of the manipulations, we apply five distinct manipulations of varying significance to the same pristine image scene. Since all five manipulations are performed on the same image, the only difference among them is the manipulation itself. This method enables a more precise analysis of how each manipulation alters the scene's semantic context. A qualitative example illustrating this ranking task is provided in Figure 4.

To quantitatively assess the performance of each mainstream semantic similarity scoring methods in Table 5 and Table 6, we calculate an accuracy score based on how well each method ranks manipulations, by calculating the number of correct rankings achieved by each SS score generation method across all images within the gold standard set. Each method ranks all five manipulations for every pristine image scene. These rankings are then compared to the manually assigned rankings for each scene, resulting in a count of correctly ranked manipulations out of 1,000. To ensure fairness, all accuracy scores are adjusted according to the manually assigned SS categories..

4.3. Evaluation results

Evaluation of analysis of SoTA IMD and localization methods: Quantitative results are shown in Table 4. Most state-of-the-art (SoTA) and mainstream image manipulation detection (IMD) and localization methods struggled to detect and localize manipulations generated by diffusion models. Only newer methods like TruFor [8] and CAT-Net

v2 [17] performed adequately. A likely reason for the low performance scores, as shown in Table 4, is the domain shift between the training and test data. Existing SoTA methods are typically trained on datasets, listed in Table 1, where manipulations are manually created using image editing software. Since generative inpainting models are relatively new, these methods have not yet adapted to this form of manipulation. Additionally, we used a fixed threshold of 0.5 when converting to a binary localization mask to ensure fairness across all methods. While this could lower performance for some methods, our qualitative analysis suggests that fine-tuning the threshold would yield similar results.

Semantic significance ranking results: Tables 5 and 6 show two sets of performance results on the CSI-IMD gold standard set. We used both ChatGPT [24] and LLava [19] to generate textual descriptions of the manipulated images. These descriptions were then input into models like LLM Score [21], CLIP Score [10], and ReBERTa Sentiment [23] to calculate semantic similarity scores.

Overall, the methods performed well in this semantic ranking task, demonstrating their ability to effectively capture the semantic significance of the manipulations. However, it is important to note that these methods had access to the original pristine image, which would not be available in real-world scenarios where defensive models only have the manipulated images to analyze. The goal of this task is to provide the digital forensics community with tools for future research on the semantic analysis of image manipulations. Through the baseline experiments described in § 4.3, we aimed to assess the feasibility of capturing and ranking the SS of image manipulations.

4.4. Discussions

Subjectivity in manipulation characterization and mitigation measures: We acknowledge the inherent subjectivity in evaluating the semantic significance of manipulations. To address this, we carefully select manipulation prompts that align with widely accepted notions of trivial and highly impactful changes. Additionally, we categorize manipulations into semantic significance groups to reduce reliance on arbitrary numeric scores for ranking. This approach allows models participating in the semantic similarity ranking task to avoid penalties for inaccurately ranking manipulations with closely related significance. While not flawless, this framework is designed to address key concerns in forensic characterization evaluation.

AI Ethics: To prevent the misuse of generated images for disinformation, we followed strict guidelines when creating our dataset. Significant manipulations are limited to common objects, avoiding images of cultural landmarks, celebrities, or political figures. Furthermore, any generated humans are intentionally of lower quality, making them easily identifiable as artificial and not real individuals.

Semantic Similarity Models	Accuracy
SAM Score [18]	0.87
LLM Score [21]	0.72
CLIP Score [10]	0.79
ReBERTa Sentiment [23]	0.80

Table 5. Semantic significance ranking accuracy results with textual descriptions generated by ChatGPT

Semantic Similarity Models	Accuracy
SAM Score [18]	0.87
LLM Score [21]	0.72
CLIP Score [10]	0.73
ReBERTa Sentiment [23]	0.75

Table 6. Semantic significance ranking accuracy results with textual descriptions generated by LLava

Limitation: Our study addresses the characterization of image manipulations, acknowledging that assessing their semantic significance can be subjective. To reduce this subjectivity, we focus on manipulations that are either highly impactful or trivial, following specific guidelines. For a more balanced and accurate assessment, we propose crowdsourcing the annotation process. Involving a broader community will help achieve a stronger consensus on the semantic significance of image manipulations, especially as more nuanced, context-dependent manipulations emerge.

5. Conclusion

We introduced CSI-IMD, a large-scale image manipulation localization dataset designed to assess the semantic significance of image manipulations. The dataset includes six distinct types of annotations to support the digital forensics community in analyzing manipulations based on their semantic significance. Additionally, we have proposed a novel image manipulation ranking task centered on semantic significance along with results and insights. Results are provided for 10 SOTA image manipulation localization methods. We hope that this dataset and the new ranking task will inspire further research into semantically impactful image manipulations and advance the field of digital forensics. The dataset and results will be released on github at a later date upon this papers acceptance.

Future works: We believe that extending the CSI-IMD dataset to include manipulations based on deeper understanding of the surrounding contexts within each image will advance the semantic analysis and characterization of image manipulation forensics. Additionally, we aim to explore highly impactful manipulations that are physically small in size. Finally, we plan to investigate using LLMs such as ChatGPT to generate manipulation prompts from keywords or image descriptions, which could broaden the scope of possible manipulations but may introduce challenges in maintaining prompt quality.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguere. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 4

[2] Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 6

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[4] dam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, page 71–80, 2020. 2, 3

[5] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. *IEEE China summit and international conference on signal and information processing*, page 422–426, 2013. 2, 3

[6] Hany Farid. Image forgery detection. *IEEE Signal Processing Magazine*, 26(2):16–25, 2009. 1

[7] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 63–72. IEEE, 2019. 2, 3

[8] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7

[9] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023. 6

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *Conference on Empirical Methods in Natural Language Processing*, page 7514–7528, 2021. 8

[11] Y.-F. Hsu and S.-F. Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. *International Conference on Multimedia and Expo*, 2006. 3

[12] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European Conference on Computer Vision (ECCV),*, pages 312–328. Springer, 2020. 6

[13] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 3

[14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 3, 4

[16] Pawel Korus and Jiwu Huang. Evaluation of random field models in multi-modal unsupervised tampering localization. *IEEE international workshop on information forensics and security (WIFS)*, pages 1–6, 2016. 3

[17] 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 (IJCV)*, 2022. 6, 8

[18] Yunxiang Li, Meixu Chen, Wenxuan Yang, Kai Wang, Jun Ma, Alan C, Bovik, and You Zhang. Samscore: A semantic structural similarity metric for image translation evaluation. 8

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 5, 8

[20] 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. 6

[21] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation, 2023. 5, 8

[22] Tian-Tsong Ng, Shih-Fu Chang, , and Q Sun. A data set of authentic and spliced image blocks. *Columbia University ADVENT Technical Report*, 2004. 2, 3

[23] Cardiff NLP. cardiffnlp/twitter-roberta-base-sentiment-latest, 2024. 5, 8

[24] OpenAI. Chatgpt, 2024. 5, 8

[25] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. Ldeepfacelab: Integrated, flexible and extensible face-swapping framework. 2

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 2

[27] 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. 2

[28] runwayml. runwayml/stable-diffusion-v1-5, 2024. 3

[29] Muhammed T Sadiq and Saji K Mathew. The disaster of misinformation: a review of research in social media. *International journal of data science and analytics*, 13:271–285, 2022. 1

[30] stabilityai. stabilityai/stable-diffusion-2-inpainting, 2024. 3

[31] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022. 6

[32] Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. a novel database for copy-move forgery detection. *IEEE international conference on image processing (ICIP)*, page 161–165, 2016. 2, 3

[33] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2022. 6

[34] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[35] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 6

[36] Ashfak Yeafi. Road vehicle images dataset, 2024. 4

[37] Zhenfei Zhang and Ming-Ching Chang. Two-stage dual augmentation with clip for improved text-to-sketch synthesis. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 1–6. IEEE, 2023. 1

[38] Zhenfei Zhang, Mingyang Li, and Ming-Ching Chang. A new benchmark and model for challenging image manipulation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7405–7413, 2024. 2, 3