Multi-axis Analysis of Image Manipulation Localization

Anonymous authors
Paper under double-blind review

Abstract

Advanced image editing software enables easy creation of highly convincing image manipulations, which has been made even more accessible in recent years due to advances in generative AI. Manipulated images, while often harmless, could spread misinformation, create false narratives, and influence people's opinions on important issues. Despite this growing threat, there is limited research on detecting advanced manipulations across different visual domains. Thus, we introduce Analysis Under Domain-shifts, qualIty, Type, and Size (AUDITS), a comprehensive benchmark designed for studying axes of analysis in image manipulation detection. AUDITS comprises over 530K images from *two distinct sources* (user and news photos). We curate our dataset to support analysis across multiple axes using recent diffusion-based inpaintings, spanning a diverse range of manipulation types and sizes. We conduct experiments under different types of *domain shift* to evaluate robustness of existing image manipulation detection methods. Our goal is to drive further research in this area by offering new insights that would help develop more reliable and generalizable image manipulation detection methods

1 Introduction

Advanced image editing tools provide an ever widening access to realistic and easy to use mechanisms for creating convincing image edits with ease (Goodfellow et al., 2014; Sohl-Dickstein et al., 2015; Rezende & Mohamed, 2015; Shi et al., 2020; Saharia et al., 2022; Yu et al., 2023). While this can be used for benign uses like satire and humor, these image edits can also help create false narratives like misrepresenting key individuals (Guo et al., 2023), resulting in the spread misinformation. To defend against this, many methods have been developed to detect manipulated images and/or localize what has been altered (Liu et al., 2022; Wang et al., 2022; Guo et al., 2023; Triaridis & Mezaris, 2024; Liu et al., 2023; Huang et al., 2025). While they have found that distribution shifts that stem from changing the type of manipulation used to alter the image or the size of the manipulated region, Figure 1 shows this only represents a subset of the challenges these models may face in practice. For example, a model trained on a single image domain may incorrectly learn that any distribution shift is the result of a manipulation, resulting in many false positives on images from a new domain at test time. However, as summarized in Table 1, existing benchmarks do not provide the ability to control for the many factors that may arise in applications of image manipulation localization (*e.g.*, Hsu & Chang (2006); Dong et al. (2013); Wen et al. (2016); MAHFOUDI et al. (2019); Țânțaru et al. (2024); Guillaro et al. (2023); Mareen et al. (2024); Yan et al. (2025)), making it impossible to accurately diagnose issues with current methods.

To address this shortcoming, we introduce Analysis Under Domain-shifts, qualIty, Type, and Size (**AUDITS**), a carefully curated, large-scale image manipulation dataset. The key contribution of this benchmark is the ability to evaluate the ability of methods to generalize across a wider range of settings like domain or image quality shifts that were not possible using prior work. Specifically, AUDITS dataset is designed to enable exploration along four major axes: 1) measuring the effect of domain shifts on image manipulation detectors, 2) explore how manipulation quality (measured via a human perceptual study) impacts detection performance, 3) provide diverse manipulation types to ensure models can generalize across techniques, and 4) evaluate how the size of a manipulation impacts performance. These benefits enable us to study the effect on current methods with some intriguing findings. For example, while many methods in prior work study the

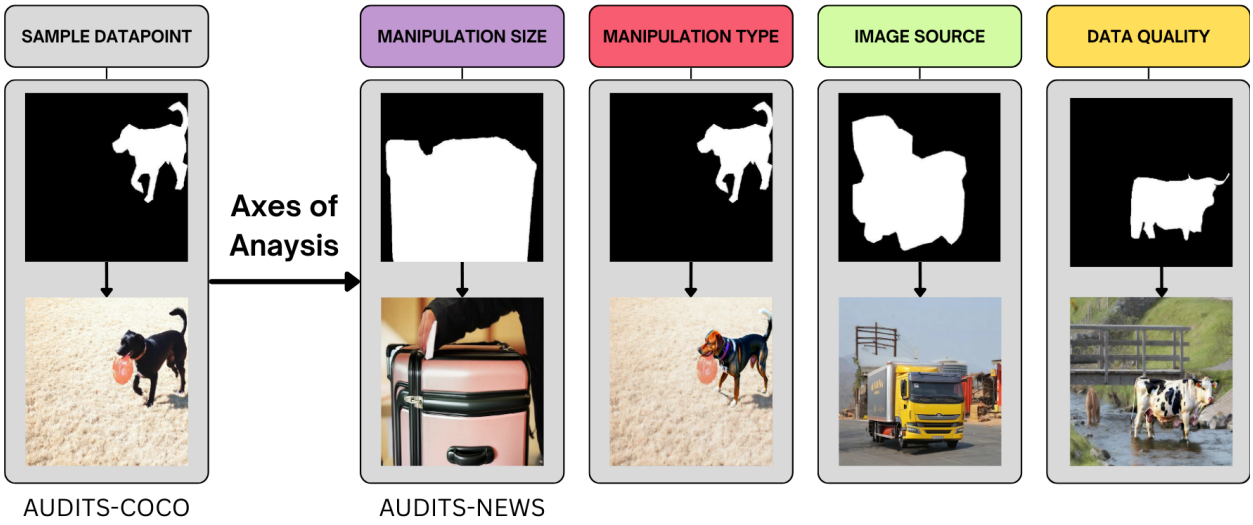

Figure 1: We illustrate the axes of analysis within AUDITS, a comprehensive benchmark for image manipulation detection. Typically, manipulation datasets focus on one axis of analysis, namely unseen manipulation types (*e.g.* Splicing *vs.* CopyMove). In contrast, AUDITS has four such axes: image source (COCO *vs.* VisualNews), data quality (*e.g.* human perception label as low quality *vs.* high quality), manipulation type (*e.g.* Blended Diffusion *vs.* Stable Diffusion) and manipulation size (*e.g.* Small *vs.* Large), and is easily extended to other axes of analysis.

effect of generalizing across manipulation techniques (*e.g.* Țânțaru et al. (2024); Wu et al. (2019); Hu et al. (2020)), we find that a change in image sources also makes a significant impact to performance. This is likely due, in part, to the localization models learning to correlate distribution shifts with manipulations, thereby incorrectly using shifts in image sources as supposed evidence of editing. In addition, we find many methods are biased towards predicting images as not manipulated, resulting in models that struggle with detecting larger manipulations.

One of the most striking observations from our experiments is that many of these issues stemming from our multi-axis shifts are not easily fixed with existing methods. For example, domain generalization methods should make a model more robust shifts in image domain (*e.g.*, Cha et al. (2021); Wortsman et al. (2022); Cha et al. (2022); Krishnamachari et al. (2024)), but we find they do not improve performance on AUDITS. This failure is likely attributed, in part, to the observed brittleness of these methods to multi-axis shifts in prior work (Humblot-Renaux et al., 2024; Wang et al., 2026). Further, while poor quality manipulations may seem like they should be easy for a model to detect, due to the ease as which human users could identify them, we find most models have negligible differences in detection performance due to the caliber of edit. These issues highlight critical failures of current methods that would limit their use. For example, a social media moderator would likely consider a detector unreliable after seeing it miss many obvious image manipulations. These issues highlight the need for our dataset to directly evaluate and explore the ability of image manipulation detectors, while also providing insight into ways that they could be improved.

In summary, our work makes the following contributions:

- We introduce AUDITS, a large-scale image manipulation detection dataset consisting of 531,057 image-mask pairs. Our dataset has diverse distributions across four axes and is conducting using 11 image manipulation techniques.
- Our analysis of existing methods shed light on such unique challenges posed by our dataset, such as a significant challenge in generalizing to OOD images, that could provide valuable insights for future research directions.
- We explore potential solutions to some of our dataset's challenges, *e.g.* integrating domain generalization techniques, and demonstrate that these solutions still struggle to boost performance, further highlighting a need for future work.

Table 1: We provide a comparison of our dataset, AUDITS, to several prior datasets w.r.t.: overall dataset size (Images), contains multiple domains , include diffusion manipulations, controls for sizes, number of manipulation types and includes a human perception label in their dataset. AUDITS stands out with its size, emphasis on diffusion-based inpainting, and inclusion of two image sources (COCO and News), surpassing the scope of previous datasets by incorporating four axes of analysis.

Dataset	Images	Multi Domain	Diffusion MTs	Controlled Sizes	# MT	w/Human Perception Label
COLUMBIA (Hsu & Chang, 2006)	1,845	✗	✗	✗	1	✗
CASIAV1 (Dong et al., 2013)	1,721	✗	✗	✓	2	✓
CASIAV2 (Dong et al., 2013)	12,323	✗	✗	✓	2	✓
COVERAGE (Wen et al., 2016)	200	✗	✗	✗	1	✗
DEFACTO (MAHFOUDI et al., 2019)	229,000	✗	✗	✗	4	✗
IMD2020 (Novozamsky et al., 2020)	2,424	✗	✗	✗	1	✗
DOLOS (Țânțaru et al., 2024)	148,112	✗	✓	✗	5	✗
TGIF (Mareen et al., 2024)	75,000	✗	✓	✗	3	✗
COCOGlide (Guillaro et al., 2023)	512	✗	✓	✗	1	✗
AutoSplice (Jia et al., 2023)	5894	✗	✓	✗	1	✗
MagicBrush (Zhang et al., 2023)	10,388	✗	✓	✗	1	✓
COCO-Inpaint (Yan et al., 2025)	375,000	✗	✓	✗	6	✗
AUDITS (ours)	530,640	✓	✓	✓	11	✓

2 Related Work

Table 1 compares our AUDITS benchmark with datasets proposed in prior work including five main factors that indicate how well a dataset may reflect real-world applications. Notably, prior work only ever covers a subset of the potential axes of analysis, limiting its usefulness. CASIA (Dong et al., 2013) includes a human perception study and MagicBrush (Zhang et al., 2023) provides human annotations, but neither evaluates manipulation localization performance conditioned on human perception labels. In contrast, AUDITS includes a subset of images annotated by human-perceived manipulation quality, enabling analysis of performance with respect to perceptual realism. Another important factor is image source diversity. Having multiple source domains allows us to isolate the effect of domain shifts on image manipulation localization models. However, many existing datasets are derived from a single domain such as faces or generic user images. For example, DOLOS (Țânțaru et al., 2024) uses face datasets FFHQ (Karras et al., 2019) and CelebA (Liu et al., 2015), whereas TGIF (Mareen et al., 2024) and CocoGlide (Guillaro et al., 2023) use COCO Lin et al. (2014). COVERAGE (Wen et al., 2016) consists primarily of consumer photography and therefore represents a single-domain dataset. However, using image editing to perpetuate misinformation via traditional journalism is a significant danger of these methods. Thus, in our benchmark we have samples from both COCO and images in news articles (Liu et al., 2021a), providing greater diversity than prior work that reflects real-world uses of these models.

We evaluate a range of image manipulation localization models with our benchmark that represents diverse design approaches (Liu et al., 2022; Țânțaru et al., 2024; Guo et al., 2023; Triaridis & Mezaris, 2024; Guillaro et al., 2023; Huang et al., 2025; Liu et al., 2021b; 2023). Many of these methods evaluated their ability to generalize across Manipulation Types (MT) or editing sizes (*e.g.*, Țânțaru et al. (2024); Guo et al. (2023); Huang et al. (2025)), but limitations in the benchmarks meant they could not accurately isolate the contribution of other factors like the quality of the manipulation or the effect of image sources. In fact, sometimes models used external datasets that change both image sources and MTs to demonstrate an ability to generalize (*e.g.*, Liu et al. (2022); Triaridis & Mezaris (2024); Liu et al. (2023)), but this coupling meant researchers would not understand if any drop in performance was due to MTs, image sources, or both. Further, while the CASIA (Dong et al., 2013) and MagicBrush (Zhang et al., 2023) datasets have human perception editing labels, they were only used by the original to validate the dataset's quality rather than providing

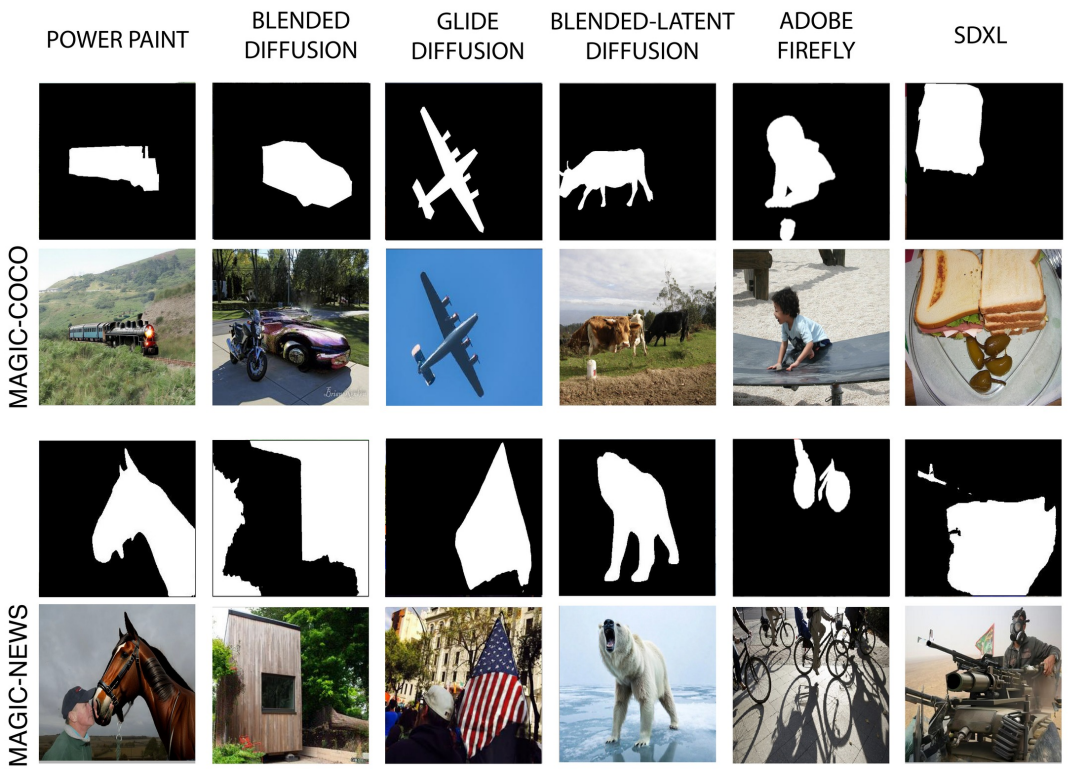

Figure 2: Examples from our benchmark AUDITS, which contains 530,640 images from two visually distinct domains: COCO (Lin et al., 2014) and Visual News (Liu et al., 2021a). The dataset contains 11 state-of-the-art manipulation techniques classified into three categories: removal, replacement, and insertion. Manipulations cover a wide range of sizes, from 1% to 100% of the image area. These examples demonstrate the diversity of manipulation types and sizes in our dataset.

insight into model performance like our work. In contrast, we use these labels along with other controlled factors to provide more extensive and precise insights into model performance to better inform future work.

3 Analysis Under Domain-shifts, quallty, Type, and Size (AUDITS)

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the goal of image manipulation localization is to learn a function $g : I \rightarrow \hat{M}$ that predicts a manipulation mask $\hat{M} \in [0,1]^{H \times W}$, estimating the ground-truth mask:

$$M_{i,j} = \begin{cases} 1, & \text{if pixel } (i,j) \text{ is manipulated} \\ 0, & \text{if pixel } (i,j) \text{ is authentic} \end{cases}$$

To better understand how performance on this task varies under different domains, we introduce our dataset AUDITS that comprises of two distinct subsets. AUDITS-News contains 311,989 images, with 271,527 manipulated images and 40,462 pristine images sampled from VisualNews (Liu et al., 2021a). AUDITS-COCO contains 218,651 images, with 187,982 manipulated images and 30,669 pristine images sampled from COCO (Lin et al., 2014). All the manipulated images are accompanied by ground-truth masks. We design our benchmark such that we can study how different axes of analysis affect image manipulation localization performance across domains.

3.1 Image Manipulation Techniques

While constructing our dataset, we sought to generate a wide variety of image manipulations that would reflect real-world scenarios, offering a significant challenge for models tasked with localizing these alterations.

To achieve this, we employed eleven major diffusion-based manipulation techniques: nine perform replacement, one performs removal, and one performs insertion. These methods are applied to both AUDITS-News and AUDITS-COCO, ensuring high diversity of manipulated images. Additional details follow.

3.1.1 Replacement and Removal Based Manipulations

We employ a diverse set of diffusion-based inpainting models to generate replacement and removal manipulations. These models span multiple generations of diffusion editing pipelines and exhibit different editing behaviors, enabling us to evaluate detector robustness across a wide range of manipulation techniques.

Our dataset includes several widely used text-guided diffusion models, including **Blended Diffusion** (Avrahami et al., 2022), **Blended Latent Diffusion** (Avrahami et al., 2023), **Stable Diffusion** (Rombach et al., 2022), **SDXL** (Podell et al., 2024), and **GLIDE** (Nichol et al., 2022). These models support localized image editing through text-conditioned diffusion and latent-space inpainting. We also incorporate more recent diffusion-based editing frameworks such as **PowerPaint** (Zhuang et al., 2025), **Flux-Inpainting** (Wang et al., 2025), and **HD-Painter** (Manukyan et al., 2025), which introduce improved prompt conditioning, structural consistency, and harmonization mechanisms for producing visually coherent edits. In addition to replacement edits, we include removal-based manipulations using **Latent Diffusion** (Rombach et al., 2022), where the selected object region is removed and inpainted to blend with the surrounding scene. Finally, we include **Adobe Firefly** (Adobe, 2024), a proprietary diffusion-based editing system, to better reflect real-world image editing tools that are widely accessible to users.

All these models are applied to both the AUDITS-News and AUDITS-COCO images. For AUDITS-News, segmentation masks are generated using ODISE (Xu et al., 2023), whereas for AUDITS-COCO, the existing masks provided by the COCO dataset are used. All the models use both text and image data to guide the generation of manipulated images. We identify an object randomly selected from the panoptic mask and replace that region with the new content based on the object class using the object's mask. For example, if the selected mask corresponds to a "cat", the information that the object is a cat is used as a prompt to guide the diffusion models in generating the new content (*i.e.*, another cat).

3.1.2 Insertion-Based Manipulations

We also insert/splice new objects into an existing images. We start splicing by utilizing ODISE to generate panoptic segmentation masks from images sourced from VisualNews and for MS COCO we use the given panoptic masks. Panoptic segmentation provides a detailed and comprehensive breakdown of the scene by categorizing each pixel into specific objects (things) and background elements (stuff). Once the segmentation map is obtained, **GLIGEN** (Li et al., 2023) is employed to generate a new object that fits contextually into the designated area. GLIGEN leverages the spatial and semantic information from the segmentation map to ensure that the newly generated object not only aligns with the surrounding elements in terms of position and scale but also blends seamlessly with the scene's overall aesthetics. The generated object is then carefully inserted into the original image I_{original} using the object mask M (Here the object mask M is randomly chosen from the panoptic segmentation), resulting in a spliced image I_{spliced}, as described by: $I_{\text{spliced}} = \text{GLIGEN}(S) \odot M \oplus I_{\text{original}}$. Where S is the segmentation map, M is the object mask, and I_{original} is the original image.

3.2 Diversity in AUDITS

When constructing the AUDITS-dataset, we curate a wide range of scenarios to better capture the different challenges that an image manipulation detector may encounter in real-world applications. News imagery, in particular, covers an broad variety of content, as shown in Figure 2. To ensure diversity in image content, we select our images from 13 different topics from the VisualNews dataset (further details in Section 8.1 of the Appendix). The VisualNews dataset draws from four major news outlets: USA Today, Washington Post, BBC, and The Guardian. We sample a substantial number of images from each topic and news outlet, details of this can be found in Section 8.1 of our Appendix. MS COCO does not have topic labels; for our AUDITS-COCO subset we sample 82 object categories, including the most common objects (*person*, *car*), and less frequent objects (*hairbrush*, *giraffe*). The combination of news-related imagery and everyday objects

ensures that our dataset represents not just specialized journalistic content but also a broad spectrum of general, real-world scenes.

In the AUDITS-News we manipulate both the stuff (*e.g.* sky, terrain) and thing (*e.g.* person, cat) categories, resulting in a wide variety of manipulation sizes based on the panoptic segmentation masks predicted by ODISE. This is in contrast to prior work that often focuses solely on foreground object manipulation (MAHFOUDI et al., 2019; Novozamsky et al., 2020).

Our AUDITS-dataset exhibits diversity in manipulation sizes. The AUDITS-News subset shows a wide range of manipulation sizes, from small edits to large alterations, requiring detection models to be robust across various levels of visual modifications. In contrast, manipulation sizes in AUDITS-COCO are skewed toward smaller edits, with relatively few medium and large manipulations. In AUDITS-News, there are 142,770 small manipulations covering $< 25\%$ of the image area, 107,872 medium manipulations covering between 25% and 60% of the image area, and 20,885 large manipulations covering $> 60\%$ of the image area. For AUDITS-COCO, the breakdown is 155,106 small, 27,553 medium, and 5,723 large manipulations, respectively. More details are in Section 10.1 of the Appendix.

3.3 Dataset Quality Survey

To assess the quality of the manipulated images and ensure their practical use for manipulation detection, we perform a human evaluation via Amazon Mechanical Turk. A total of 4,950 images were used in our survey (2,750 from AUDITS-News and 2,200 from AUDITS-COCO). We include 5 generators for AUDITS-News and 4 for AUDITS-COCO[1]. We sample 500 images from each generator. Additionally we include 250 authentic images from AUDITS-News and 200 from AUDITS-COCO. Each image is evaluated by 3 people; in total we have 14,850 responses from 1,829 unique workers. The respondents were shown two images: a manipulated image $\boldsymbol{X}$ modified in a region specified by a binary mask $\boldsymbol{M}$, and a copy of $\boldsymbol{X}$ with a mask superimposed on it with a transparency value α. Additional details and examples are in Section 12 of the Appendix.

3.3.1 Survey Questions Details

The respondents are asked the following questions: Q1) *"Do you think this image is manipulated?"* (yes or no), Q2) *"Do you see the **object** in the image (you can use the mask overlay to the right of the Image to better see the object)?"* (yes or no); Q3) *"Does the **object** look realistic?"* (yes, maybe or no), and Q4) *"Does the **object** look natural in the background?"* (yes, maybe, no). For each example, we specify which **object** the respondents should look for (*e.g.* a "giraffe"). We discard the answers to Q3 and Q4 if the respondent answered "No" to Q2, which results in 9,596 responses. Next, we utilize Q3 to create pseudo-labels for "High quality" and "Low quality" images. By combining the "Maybe" and "No" responses and utilizing majority voting, we get 1,507 "Low quality" images. Meanwhile, we have 1,905 "Yes" responses corresponding to "High quality" images, hence we get 3,412 images in total. We evaluate the effect of these pseudo-labels on manipulation detection models in Section 4.5.

4 Experiments

To benchmark the recent manipulation detection models, we employ two different data splits based on manipulation type (11 types) and image subset (News or COCO): one for in-distribution (ID) and another for out-of-distribution (OOD) performance assessment. We compare a number of different image manipulation detection models in our experiments, namely: **PSCC-Net** (Liu et al., 2022), **MMFusion** (Triaridis & Mezaris, 2024), **HiFi** (Guo et al., 2023) ,**EVP** (Liu et al., 2023), **TruFor** (Guillaro et al., 2023), **SIDA** (Huang et al., 2025), **DOLOS** (Țânțaru et al., 2024), **Swin Transformer** (Liu et al., 2021b) used as the base encoder for an Upernet model (Xiao et al., 2018) and additionally utilize domain generalization methods like **SWAD** (Cha et al., 2021), **Model Soups** (Wortsman et al., 2022), **MIRO** (Cha et al., 2022) and **URM** (Krishnamachari et al., 2024). Figure 3 outlines the distribution of manipulation types across the

[1]This evaluation included all manipulation types with the exception of Adobe Firefly (for either subset) and Blended Latent Diffusion / GLIGEN (for AUDITS-COCO).

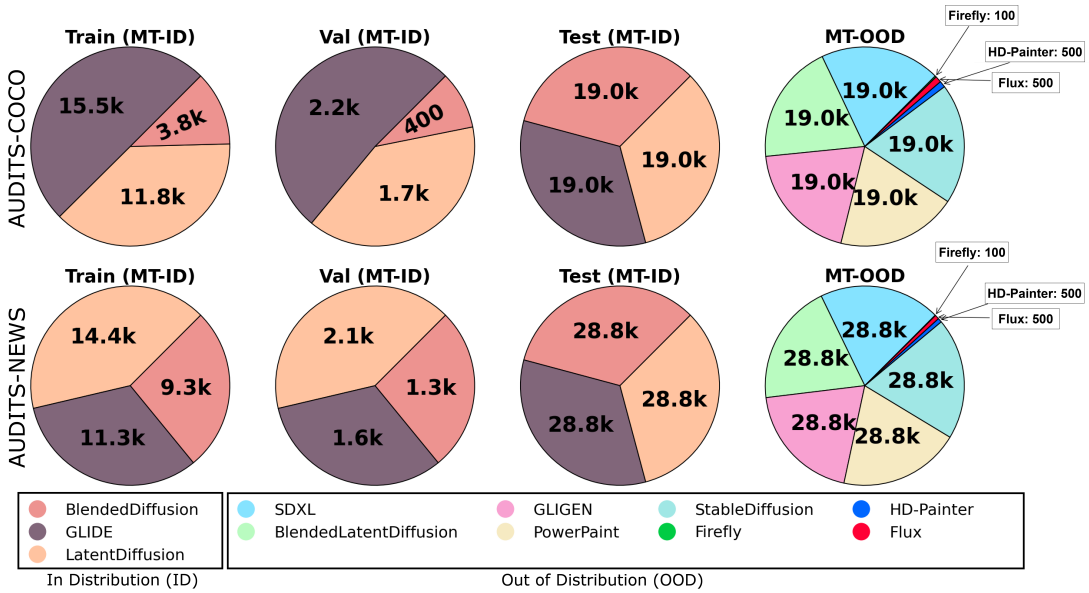

Figure 3: Statistics on manipulation types in AUDITS-News and AUDITS-COCO.

AUDITS-News and AUDITS-COCO subsets, separating manipulation methods into those for training and those reserved for testing as out-of-domain.

4.1 Evaluation Protocol

Manipulation localization is formulated as a pixel-level prediction task. Given an input image I, a detector produces a probability map $\hat{M} \in [0,1]^{H \times W}$ indicating the likelihood that each pixel belongs to a manipulated region. Since predictions are made at the pixel level, we evaluate models using **pixel-level micro-averaged metrics**, pooling predictions across all images in the dataset.

A straightforward evaluation would compare the predicted manipulation map $\hat{M}$ with the binary ground-truth mask M. However, diffusion-based editing models frequently introduce visual changes outside the annotated manipulation mask. These changes may arise from global adjustments such as color harmonization, texture blending, or lighting modifications that occur during the editing process. As a result, detectors may correctly identify regions that differ from the original image even though they fall outside the annotated mask.

Ignoring such regions would penalize detectors that identify real pixel-level changes, while treating them as fully manipulated could reward detections unrelated to the intended edit. To address this ambiguity, we introduce an intermediate category capturing *off-mask drift*.

Ternary ground-truth representation. Let O denote the original image and E the edited image. We compute a per-pixel RGB difference map

$$D_{i,j} = \frac{1}{3} \sum_{c \in \{R,G,B\}} (O_{i,j,c} - E_{i,j,c})^2. \tag{1}$$

Using this difference map and the binary manipulation mask M, we construct a ternary ground-truth representation G:

$$G_{i,j} = \begin{cases} \text{manipulated}, & M_{i,j} = 1, \\ \text{ambiguous}, & M_{i,j} = 0 \text{ and } D_{i,j} > \tau, \\ \text{authentic}, & \text{otherwise.} \end{cases} \tag{2}$$

The threshold τ determines whether visual differences outside the annotated manipulation mask are significant enough to be considered ambiguous. In practice, we compute the mean squared difference between the original and edited images and mark pixels with $D_{i,j} > \tau$ as ambiguous. In all experiments we set $\tau = 0.0025$, which was chosen empirically to detect visible editing artifacts while avoiding small numerical differences. The ambiguous category therefore captures pixels outside the annotated manipulation mask whose appearance differs from the original image. These differences arise from contextual modifications introduced by modern editing pipelines (*e.g.* diffusion-based editing), which may alter surrounding pixels to ensure visual consistency. Identifying these pixels allows us to measure the effect of off-mask drift introduced during the editing process.

One possible way to eliminate ambiguous pixels would be to simply copy the manipulated region into the original image, ensuring that only pixels inside the annotated manipulation mask are modified. This has two major drawbacks. First, this is not the expected usage by general users (*i.e.*, most users would use the diffusion-based edit directly), thereby introducing a difference in evaluation *vs.* its practical use. Second, this procedure would artificially remove contextual changes introduced by modern editing pipelines, particularly diffusion-based editing methods that often modify surrounding pixels to maintain visual consistency (*e.g.* blending or color harmonization) Thus, these edits are often easily identified by humans. In contrast, our goal is to evaluate detectors under realistic editing conditions rather than artificially constrained manipulations, and therefore we retain these off-mask modifications.

Binary evaluation with ambiguous-pixel weighting. Since common evaluation metrics such as AUROC, precision, recall, and F1 are defined for binary classification, we map the ternary labels to a binary prediction task. Pixels labeled as *manipulated* are treated as positive examples, while pixels labeled as *authentic* are treated as negative examples.

Ambiguous pixels fall outside the annotated manipulation mask and are therefore treated as negative examples. However, because these pixels may still reflect visual changes introduced by the editing process, we assign them reduced weight during evaluation. Specifically, we define a per-pixel weight function

$$w(i,j) = \begin{cases} 1, & \text{if pixel } (i,j) \text{ is manipulated,} \\ \alpha, & \text{if pixel } (i,j) \text{ is ambiguous,} \\ 1, & \text{if pixel } (i,j) \text{ is authentic,} \end{cases} \tag{3}$$

where $\alpha \in [0,1]$ controls the contribution of ambiguous pixels. In our experiments we set $\alpha = 0.5$, which reduces the influence of ambiguous regions while still allowing them to contribute to the evaluation. This weighting scheme allows the evaluation to emphasize precise localization of the annotated manipulation region while reducing the impact of ambiguous off-mask changes. If the editing process modifies only pixels inside the annotated manipulation mask (*e.g.* copy–paste manipulations), the ambiguous category is empty and the evaluation reduces to standard binary manipulation localization. These weights are incorporated into the Area Under the ROC Curve (AUROC), precision, recall, and F1 calculations described below.

Metrics. AUROC is computed using a streaming histogram estimator that approximates

$$\text{AUROC} = P(s_p > s_n) + \frac{1}{2} P(s_p = s_n), \tag{4}$$

where s_p and s_n denote prediction scores for positive and negative pixels, respectively. The factor $1/2$ accounts for tied prediction scores, as is standard in ROC analysis. Precision, recall, and F1 score are computed after thresholding prediction scores at $\mu = 0.5$. All metrics are computed using pixel-level micro averaging across the dataset.

The ambiguous-pixel weighted evaluation described above is used only for the pixel-level localization experiments in Section 4.2 and Section 4.3. For the rest of the tables we report standard AUROC and or F1 without ternary mask construction or ambiguous-pixel weighting. This keeps those results directly comparable to prior work while reserving the ambiguous-pixel weighted protocol for experiments where ambiguity outside the annotated manipulation mask is central to the analysis.

Table 2: Comparing AUC and F1 scores of models trained on AUDITS-News or AUDITS-COCO and evaluated on both datasets to measure generalization across image sources and manipulation types. Methods are grouped by base architecture to facilitate direct comparison with their domain generalization (DG) variants. Bold numbers indicate the best performance among baseline methods, while boxed values highlight the best result within each method family (e.g., EVP and its DG variants, MMFusion and its DG variants).

Trained on:	AUDITS-News				AUDITS-COCO				AUDITS-COCO				AUDITS-News			
Tested on:	AUDITS-News								AUDITS-COCO							
	MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
EVP																
EVP	93.2	75.6	**81.6**	52.8	74.4	37.9	**63.5**	20.8	85.9	53.2	76.7	29.8	85.8	**57.4**	66.3	27.3
EVP+SWAD	86.1	62.0	77.3	45.4	77.6	41.9	63.1	22.3	86.8	56.0	72.1	30.1	77.1	45.5	64.0	24.8
EVP+Soup	92.1	70.1	77.7	43.1	83.3	49.8	63.2	15.1	90.5	60.3	71.8	28.9	76.1	39.2	63.4	21.3
EVP+MIRO	95.4	81.5	76.4	44.0	81.7	55.9	61.6	22.6	94.4	72.0	73.2	25.8	92.1	67.5	64.5	22.8
EVP+URM	90.2	62.9	72.6	26.5	84.2	50.9	63.5	12.2	91.2	64.6	69.9	16.5	83.8	52.2	64.4	12.8
MMFusion																
MMFusion	**96.8**	**82.6**	73.8	51.5	**88.6**	**70.9**	56.5	31.3	**96.7**	**75.0**	64.4	30.3	**92.3**	56.7	57.0	22.8
MMFusion+SWAD	96.6	82.5	80.7	60.4	86.8	71.4	62.7	38.9	96.4	77.0	71.0	38.9	91.8	54.0	61.0	28.7
MMFusion+MIRO	96.5	82.7	75.9	55.5	87.0	69.1	58.2	24.2	96.7	78.4	70.6	29.7	89.3	50.4	50.8	21.7
Other Manipulation Detection Methods																
PSCC-Net	89.1	70.5	58.9	33.2	70.7	51.1	45.7	40.6	80.2	45.7	68.3	33.6	85.2	49.7	64.0	25.3
SIDA	82.5	72.0	75.2	**62.4**	68.4	52.2	62.4	41.1	81.0	67.1	**76.8**	**60.0**	80.1	57.1	**67.9**	**41.9**
TruFor	59.4	36.1	55.7	37.9	49.8	44.1	48.0	**44.3**	49.7	25.9	47.7	25.9	64.2	30.5	56.7	25.6
DOLOS	69.0	38.6	61.6	26.6	54.2	5.0	55.7	29.1	66.2	7.5	58.5	25.3	60.2	26.9	44.7	16.1
HiFi	55.1	16.5	50.2	1.9	60.4	45.0	51.4	24.3	75.0	48.7	57.3	26.0	52.9	9.7	49.8	0.8
Upernet	74.2	62.6	64.1	44.3	67.2	51.2	53.4	14.7	77.3	64.5	54.3	16.4	68.6	43.0	53.7	22.0

4.2 Effect of Image Source Domain and Manipulation Type

Table 2 reports results when training models on AUDITS-News or AUDITS-COCO and evaluating on both datasets, allowing us to analyze generalization across manipulation types (MT) and image sources. Overall, performance degrades when models are evaluated on out-of-distribution settings arising from either MT or domain shifts. To isolate the effect of manipulation type, we compare performance when training and testing on the same image source while changing MT. For example, MMFusion trained and evaluated on AUDITS-News has a 23-point decrease. To isolate the effect of image source, we compare performance across datasets while keeping manipulation types fixed. When training on AUDITS-News and testing on AUDITS-COCO (MT-ID), MMFusion has an 8-point decrease. These results suggest that both manipulation type and image source shifts play an important role in building robust detectors. However, most prior work has primarily focused on generalization across MT (Wen et al., 2016; MAHFOUDI et al., 2019; Țânțaru et al., 2024; Guillaro et al., 2023; Mareen et al., 2024), without isolating the effect of image source shifts.

We further compare each base model with its domain generalization (DG) variants to assess whether DG techniques improve robustness to these distribution shifts. For EVP, methods such as SWAD, MIRO, and URM yield comparable performance to the base model, with no single DG approach consistently outperforming EVP across all evaluation settings. A similar trend is observed for MMFusion, where DG variants provide occasional improvements but do not clearly dominate the base model.

Overall, these results suggest that standard DG methods do not consistently improve performance under the multi-axis distribution shifts considered in AUDITS. While MIRO Cha et al. (2022) and URM Krishnamachari

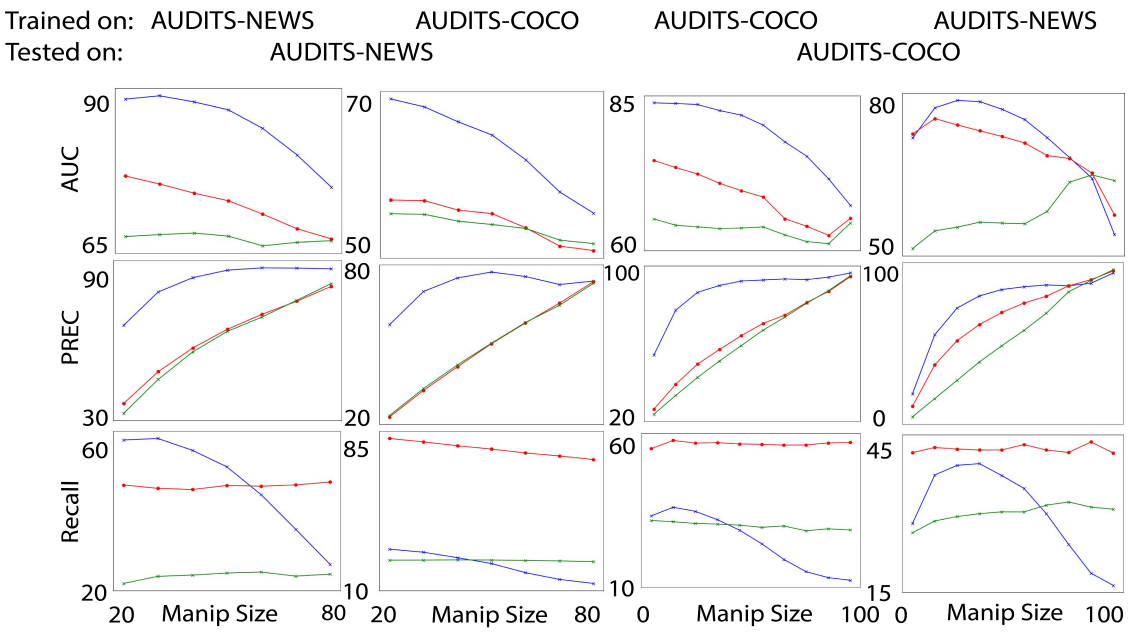

Figure 4: Comparing the AUC/Precision/Recall performance of models trained on the AUDITS-News and AUDITS-COCO to study generalization across image distributions and manipulation sizes. EVP results are in red, PSCC-Net are blue and DOLOS are green. The first two columns are results on AUDITS-News's ID test set and OOD set by getting the average of all the results in the specific size category. The last 2 columns are tested on AUDITS-COCO's ID test set and OOD set by getting the average of all the results in the specific size category.

et al. (2024) provide slight improvements for EVP in some settings, their gains are modest and inconsistent. Notably, URM is a more recent DG method, yet its improvements remain limited, suggesting that advances on standard DG benchmarks do not directly translate to improved performance on AUDITS.

4.3 Effect of Manipulation Size

Another aspect of our dataset is the manipulation size. Here, we determine how SoTA models perform across different sizes of manipulations. We utilize the full range of sizes for AUDITS-News which ranges from 20% - 80% and AUDITS-COCO which ranges from 0% - 100%.

Results. Figure 4 highlights the performance of a range of models, one somewhat counterintuitive observation is that AUC performance on small manipulations is generally higher than large ones. However, this is simply due to the fact that the majority of the image is authentic, so any model that is biased towards predicting that image regions are real would get higher performance. This also manifests itself in the very low recall rates of larger manipulations and corresponding high precision rates on these larger manipulations models are simply less likely to prediction an image region as manipulated unless it is fairly certain it is correct.

Comparing across methods, we see that EVP consistently gets better performance when evaluated on AUDITS-News. However, we obtain mixed results on AUDITS-COCO when trained with AUDITS-News, where best performance bounces back and forth mostly between EVP and PSCC-Net. This helps demonstrate the different properties of both splits, where the AUDITS-COCO manipulations contain a higher portion of small manipulations (see Section 10.1 of the Appendix for size statistics), which can result in significant degradation when training with the generally larger manipulations in AUDITS-News.

Table 3: Performance when evaluated on MagicBrush (Zhang et al., 2023) trained on AUDITS-News

Model	AUC	F1	IoU
PSCC-Net	46.7	25.6	11.8
DOLOS	52.9	26.0	11.1
EVP	**77.0**	**36.8**	**23.3**

Table 4: Performance on CocoGLIDE (Zhang et al., 2023) (trained on AUDITS-News).

Model	AUC	F1	IoU
PSCC-Net (reported in CocoGLIDE)	77.7	51.5	-
TruFor (proposed in CocoGLIDE)	75.2	52.3	-
PSCC-Net	78.0	51.3	38.2
DOLOS	55.6	38.5	25.3
EVP	**83.6**	**57.0**	**42.9**

Table 5: Comparing the AUC results from our human evaluation by utilizing majority voting we use 3412 images (1,507 "Low" quality and 1905 "High" quality images) from our test and OOD set based on question 3 and combined answer choice Maybe with No. An image is considered of "High" quality if more people consider it realistic and "Low" if more people consider it not-realistic. The first 4 columns of results are tested on AUDITS-News's ID test set and OOD set combined and the last 4 columns of results are tested on AUDITS-COCO's ID test set and OOD set combined.

Trained on:	AUDITS-News		AUDITS-COCO		AUDITS-COCO		AUDITS-News	
Tested on:	AUDITS-News				AUDITS-COCO			
Model	Low	High	Low	High	Low	High	Low	High
EVP	80.9	80.7	58.9	59.5	90.3	88.0	69.3	68.0
DOLOS	80.3	79.8	75.4	75.1	66.1	64.5	66.1	64.5
PSCC-Net	72.8	71.8	47.0	48.9	78.2	77.2	49.3	49.0
HiFi	72.6	72.3	48.4	48.7	73.3	71.5	51.5	51.4

4.4 Generalizing Across Inpainting Datasets

We further evaluate whether models trained on AUDITS transfer to external diffusion-based inpainting datasets. Tables 3 and 4 report results on MagicBrush and CocoGLIDE, respectively. These datasets differ from AUDITS in both image source and editing procedure, and therefore provide a complementary test of out-of-dataset generalization.

On MagicBrush, EVP substantially outperforms the other evaluated methods, improving AUC by more than 20 points over the next best method. However, the relative behavior of methods changes on CocoGLIDE: PSCC-Net performs competitively with the previously reported CocoGLIDE result, while DOLOS remains substantially lower. EVP is again the strongest method, outperforming PSCC-Net by roughly 6 AUC points.

These results show that method rankings are not fixed across external inpainting datasets. A method that performs poorly on one external dataset may perform competitively on another, suggesting that evaluation on a single dataset can give an incomplete picture of generalization. This further motivates AUDITS as a multi-axis benchmark for analyzing robustness across image sources, manipulation types, and editing conditions, rather than only reporting aggregate performance on one external test set.

4.5 Effect of Manipulation Quality

An additional important question we explore is how manipulation quality affects these detectors. Low quality manipulations may be easy for humans to detect, suggesting high quality manipulations are more important. However, we argue against this interpretation as many applications require a detector that can detect both low and high quality manipulations. For example, when trying to detect misinformation by identifying whether or not an image has been manipulated, users would not trust a system if the detector could not recognize that an image has been manipulated when it is easy for them to identify. Thus, we explore how human perception of manipulation quality by separating samples into "High" and "Low" quality manipulations using the results of our human study. Table 5 generally reports negligible impact between the two types of

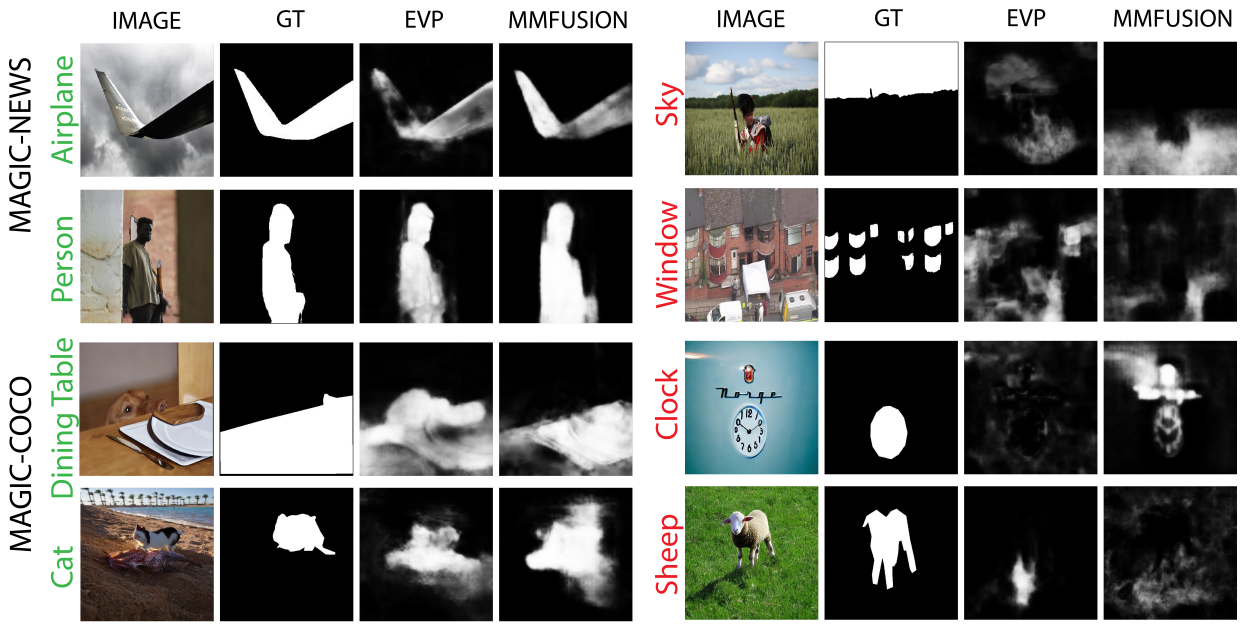

Figure 5: Qualitative comparison of EVP and MMFusion across object categories from AUDITS-NEWS (top) and AUDITS-COCO (bottom). These results suggest that both models struggle with similar object types under distribution shifts.

samples, with detectors often reporting performance with 2% of each other (and quite often less than 0.5%). Notably, the ranking of methods did not change between the two manipulation types, *i.e.*, this suggests that the manipulations types in our dataset are sufficient quality that detectors find that human perception has a negligible impact on their performance. This also suggests that detectors may boost performance by trying to emulate human perception.

4.6 Qualitative Analysis

To further understand the behavior of our selected models, we conducted a qualitative analysis. We examined the manipulated object categories that appeared more than 500 times in each OOD dataset of AUDITS-NEWS and AUDITS-COCO and computed the average AUC for each object category. Our findings indicate that certain objects, such as "person," exhibited strong generalization, whereas other objects like "sky" or "sheep" showed performance degradation. Moreover, we compared results between EVP and MMFusion and observed that for many object categories, their performance remained similar. These insights suggest that certain objects can be challenging to detect across models; additional anlysis is provided in Section 10 of the Appendix.

5 Conclusion

We introduced Analysis Under Domain-shifts, qualIty, Type, and Size (AUDITS), a large image manipulation dataset to study the robustness and generalization capabilities of image manipulation detectors. Our dataset features a range of image sources, manipulation types, and sizes. Notably, we have utilized state-of-the-art image manipulation techniques. Through extensive experiments, we found that while current manipulation localization methods perform well on in-distribution data, they struggle on out-of-distribution samples, underscoring the need for better generalization. In our future research we will focus on integrating this contextual information to add to detection models and enhancing the quality of manipulations to drive further advancements in the field.

References

Adobe. Adobe firefly, 2024. URL https://www.adobe.com/products/firefly.html. Accessed: 2024-09-20.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18208–18218, June 2022.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4):1–11, 2023.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *European conference on computer vision*, pp. 440–457. Springer, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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*. IEEE, July 2013. doi: 10.1109/chinasip.2013.6625374. URL https://doi.org/10.1109/chinasip.2013.6625374.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606–20615, June 2023.

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*, pp. 3155–3165, 2023.

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

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pp. 312–328. Springer, 2020.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B. Moeslund. A noisy elephant in the room: Is your out-of-distribution detector robust to label noise? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22626–22636, June 2024. doi: 10.1109/CVPR52733.2024.02135.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 893–903, 2023.

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

Kiran Krishnamachari, See-Kiong Ng, and Chuan-Sheng Foo. Uniformly distributed feature representations for fair and robust learning. *Transactions on Machine Learning Research*, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6761–6771. Association for Computational Linguistics, November 2021a.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19434–19445, 2023.

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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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)*, pp. 1–5, 2019. doi: 10.23919/EUSIPCO.2019.8903181.

Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. HD-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=6lB5qtdYAg`.

Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallendael, Peter Lambert, and Symeon Papadopoulos. Tgif: Text-guided inpainting forgery dataset. In *Proc. Int. Workshop on Information Forensics and Security (WIFS) 2024*, 2024.

Alexander Quinn 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. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Adam Novozamsky, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 71–80, March 2020.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=di52zR8xgf`.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

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 (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7454–7464, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Konstantinos Triaridis and Vasileios Mezaris. Exploring multi-modal fusion for image manipulation detection and localization. In *Proc. 30th Int. Conf. on MultiMedia Modeling (MMM 2024)*, Jan.-Feb. 2024.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373, 2022.

Siqi Wang, Aoming Liu, and Bryan A. Plummer. Noise-aware generalization: Robustness to in-domain noise and out-of-domain generalization. In *International Conference on Learning Representations (ICLR)*, 2026.

Yikai Wang, Chenjie Cao, Junqiu Yu, Ke Fan, Xiangyang Xue, and Yanwei Fu. Towards enhanced image inpainting: Mitigating unwanted object insertion and preserving color consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23237–23248, June 2025.

Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. Coverage – a novel database for copy-move forgery detection. In *IEEE International Conference on Image processing (ICIP)*, pp. 161–165, 2016.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. PMLR, 2022.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9543–9552, 2019.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2955–2966, June 2023.

Haozhen Yan, Yan Hong, Jiahui Zhan, Yikun Ji, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. Coco-inpaint: A benchmark for image inpainting detection and manipulation localization. *arXiv preprint arXiv:2504.18361*, 2025.

Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023.

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. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 195–211, Cham, 2025. Springer Nature Switzerland.

Dragoș-Constantin Țânțaru, Elisabeta Oneață, and Dan Oneață. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6258–6268, 2024.

Table 6: Comparing AUC and F1 score of models trained only on AUDITS-News or AUDITS-COCO and tested on both to measure generalization across image sources and manipulation types this set contains Authentic Images. We bold the best numbers in the top half, whereas we underline the best method between EVP and combinations with DG methods in the bottom half.

Trained on:	AUDITS-News				AUDITS-COCO				AUDITS-COCO				AUDITS-News			
Tested on:	AUDITS-News								AUDITS-COCO							
	MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Upernet	64.2	48.9	61.4	35.2	60.2	35.0	53.5	8.7	59.2	29.1	53.1	7.0	56.2	25.0	50.7	18.0
DOLOS	64.1	48.0	60.8	43.4	53.2	40.6	56.9	43.0	60.4	21.8	61.8	24.5	56.2	27.7	44.5	25.0
PSCC-Net	74.8	70.4	56.8	50.0	66.0	53.5	47.4	44.9	70.3	36.9	72.0	32.2	72.5	45.8	67.7	30.0
MMFusion	**83.0**	**81.4**	76.9	**82.6**	**76.9**	**70.8**	59.1	41.2	**80.9**	**64.9**	68.1	27.5	74.4	**51.2**	59.0	23.5
HiFi	53.9	14.0	50.2	1.5	57.2	39.6	51.7	39.6	63.7	14.0	56.5	35.1	51.6	6.2	49.8	0.4
EVP	80.9	76.6	**81.5**	61.6	68.3	54.2	**64.6**	**45.5**	76.8	49.1	**81.7**	**39.1**	71.6	45.7	65.3	29.3
TruFor	51.8	40.6	51.2	41.7	50.7	41.0	48.6	40.8	50.3	21.3	48.5	21.2	54.7	24.9	53.4	24.6
SIDA	73.2	68.2	74.8	59.1	62.0	41.6	61.6	33.3	66.5	43.5	63.5	33.0	67.9	46.8	**73.9**	**46.2**
EVP+SWAD	76.7	65.6	78.0	57.2	70.8	56.3	64.5	45.7	77.2	49.6	77.1	36.0	66.7	37.1	62.3	28.0
MMFusion+SWAD	84.0	81.3	80.4	64.0	77.5	69.2	64.3	44.4	82.3	65.9	74.2	36.9	75.5	50.4	59.2	29.0
EVP+Soup	77.7	70.1	77.7	54.6	70.1	56.9	64.3	45.9	77.8	49.3	76.1	34.5	68.9	39.2	63.4	28.3
EVP+MIRO	82.0	78.0	76.8	53.9	73.0	59.2	63.3	44.9	79.4	53.0	77.4	32.9	75.2	49.8	62.9	26.9
MMFusion+MIRO	83.6	81.7	74.1	61.3	77.3	69.6	61.4	43.3	83.1	66.9	74.6	36.1	72.0	48.9	47.5	27.0
EVP+URM	80.0	70.4	73.9	51.1	74.9	64.0	64.9	44.4	79.7	56.6	73.6	30.5	71.9	43.9	65.5	25.6

Appendix

In this Appendix we include additional information about our experiments: 1) We conduct another analysis of our human eval results by looking at the High versus Low quality images. 2) We utilize the captions that are included with our AUDITS-News and AUDITS-COCO dataset to determine if their is any discrepancy between captions that are related or not to the manipulated object. 3) We include an example image of what a human annotator would have seen during the human evaluation. Additionally, we include an example of what the manually create Adobe Firefly manipulations looked like before and after and discuss the ethical considerations of our work.

6 Ethical Considerations

Our work focuses on benchmarking and advancing the methods for detecting manipulated images. We believe this is an important effort aimed at countering misinformation online, especially in the era of advanced generative models. Thus, this work is ethical by its nature. At the same time, since we are producing a dataset with manipulated images, there is some potential for misuse, *i.e.*, it being used for training even more sophisticated falsification methods. However, highlighting these issues should help inform experts when creating defenses.

7 Results with Authentic Images

This can be seen in Table 6 which includes the results with authentic images when calculating the AUC. Because of the fact that in many cases the AUC for authentic images was around 50% for the AUC we reported the model performance with authentic images in this table. We can see that in some cases for example with EVP when looking at columns 1 and 3 it appears that the AUC is higher for MT-ID and MT-OOD, however when we consider the fact the MT-ID contains authentic images, it helps explain why the performance appears lower. This is why we report Table 2 in the main paper which helps us better understand the impact in performance with only manipulations. One thing that is still highlighted in Table 6

Table 7: Comparing the Mean and Standard Deviation (SD) of AUC for models trained using 3 different seeds for models trained on AUDITS-News and AUDITS-COCO, respectively, and tested on both sets to measure generalization across image sources and manipulation types. Model used are EVP (Liu et al., 2023), MMFusion (Triaridis & Mezaris, 2024) EVP+MIRO (Cha et al., 2022).

Trained on:	AUDITS-News				AUDITS-COCO				AUDITS-COCO				AUDITS-News			
Tested on:	AUDITS-News								AUDITS-COCO							
	MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD		MT-ID		MT-OOD	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
EVP	81.1	0.54	79.3	0.97	71.7	0.44	64.5	1.2	79.8	0.14	81.5	0.81	71.8	1.50	61.5	2.77
MMFusion	84.0	0.08	79.5	1.92	76.9	0.69	64.6	2.09	82.3	0.41	76.1	2.05	76.3	0.97	61.9	2.24

Table 8: Comparing the AUC performance of models trained on the AUDITS-News and AUDITS-News respectively subset to generalize across image distributions and caption relevance. *Cap-Ref: manipulated object in caption *Not Ref: manipulated object not in caption. Columns 2-5 are tested on the AUDITS-News test set and columns 6-9 are tested on the AUDITS-COCO test set.

	AUDITS-News		AUDITS-COCO		AUDITS-News		AUDITS-COCO	
	Cap-Ref	Not Ref	Cap-Ref	Not Ref	Cap-Ref	Not Ref	Cap-Ref	Not Ref
EVP	**75.3**	**75.0**	60.7	57.1	**89.6**	**86.1**	**66.2**	**67.5**
DOLOS	72.6	71.8	**67.9**	**66.5**	64.1	63.1	49.8	50.5
PSCC-Net	65.2	64.9	51.2	50.2	75.6	74.5	49.2	49.4

is that the image source performance still drops as can be seen for MMFusion in columns 1 and 5 which highlights the impact of image source and showcases a real world setting which manipulated images can come from many different source.

8 Semantic Saliency

Manipulation Semantic Salience Another aspect of our dataset is looking at the semantic saliency with respect to captions, namely if a manipulated object is mentioned in a caption describing a manipulated image. For this task we utilize the original captions used from Visual News and COCO.

Generalizing across manipulation semantic salience Table 8 refers to the results obtained for related and unrelated captions with respect to the manipulated object being mentioned in the caption. We can see a similar trend to Table 2 in the main paper whereby for both related and unrelated captions EVP tends to be the best performing model for columns 2-3 and 6-9 as to be expected with DOLOS performing the best for columns 4-5. For EVP, columns 4-5 and 6-7 we can see that the *Cap-Ref were the only times it scored slightly higher than *Not Ref, we know for AUDITS-News there are more related captions hence the higher performance for columns 4-5 are expected. However for columns 6-7, being tested on the AUDITS-COCO test set showcases again how challenging the AUDITS-COCO subset can be, as we can see there are fewer related captions. Hence, having a model that has a loosely related caption can possibly highlight challenging manipulations to detect even with the best performing manipulation detection models.

8.1 AUDITS-News Dataset Statistics

We provide additional statistics for the AUDITS-News subset derived from the VisualNews dataset. Table 9 shows the distribution of images across different news topics. The dataset spans a wide range of categories including international news, law and crime, arts, science, and politics, ensuring diversity in scene content and semantic context. This variety helps capture realistic image manipulation scenarios that arise in real-world journalistic settings.

Table 9: Number of images per topic in our AUDITS-News subset.

International	Law and Crime	Arts	World	Science	Sport	Conflict
36,684	35,273	35,867	35,489	33,541	24,022	8,188
Nature	Film	Music	Business	Politics	Disaster	
24,157	19,694	17,374	16,155	15,080	10,382	

Table 10: Number of images per news outlet in our AUDITS-News subset.

Source	Number of Images
The Guardian	120,753
BBC	72,201
USA Today	62,682
Washington Post	56,401

Table 10 presents the distribution of images across news sources. The dataset includes images from four major outlets: The Guardian, BBC, USA Today, and The Washington Post. While the distribution is not perfectly balanced, all sources contribute a substantial number of images, providing diversity in capture styles, editorial practices, and visual content. This variation helps reduce bias toward a single publication and improves the robustness of evaluation across different news domains.

8.2 Traditional Image Manipulation Analysis

We include a different experiment whereby we investigate the performance of the EVP (Liu et al., 2023) model trained on data from either DEFACTO, AUDITS or both. We can see the performance on traditional image manipulation like Copymove and Splicing in Table 11. Based on these results we can see that for traditional manipulations training on either the DEFACTO dataset or the AUDITS+DEFACTO dataset provides much better results for both CASIA and DEFACTO, this is of course expected since our dataset is not composed of any taditional manipulations. However we see the drop in performance when trying to train using traditional manipulation and testing on diffusion based inpaintings as we can see in Table 12. We can see training on DEFACTO has a big drop in performance when testing on diffusion based inpaintings, which highlights the benefit of our dataset. Additionally we can see combining our AUDITS dataset with other traditional manipulation datasets like DEFACTO can provide comparable performance for both traditional and diffusion based manipulations as is seen in both Table 11 and Table 12.

9 Image-Level Detection

To complement our localization experiments, we evaluate image-level manipulation detection on AUDITS. We evaluated the classification task using PSCC-Net (Liu et al., 2022) and HiFi (Guo et al., 2023), as both models include a classification head. The results, summarized in Table 13, show that PSCC-Net outperforms HiFi in most cases despite their similar architecture.

Interestingly, PSCC-Net performs particularly well when trained and tested on AUDITS-COCO, likely due to its HRNet (Wang et al., 2020) backbone, which is pre-trained on ImageNet (Deng et al., 2009). However, both models experience a significant drop in performance when tested on OOD images. This highlights the importance of our work in exposing these generalization challenges.

10 Qualitative Analysis of Object Categories

To further illustrate our qualitative findings, we plotted the average AUC for each selected object category across different training and testing setup for our two high performing models EVP and MMFusion. This can be

Figure 6: An example of a Adobe Firefly inpainted image from COCO described in Section 3.1.1 in the main paper. With the left most being the original and the middle being the Firefly Inpainted image.

Table 11: Comparing the performance of the EVP (Liu et al., 2023) model trained on data from either DEFACTO (MAHFOUDI et al., 2019), AUDITS-COCO or both and testing on classic image manipulation datasets that contain Copymove (CM) and Splicing (SP) images, namely CASIAv1, CASIAv2 (Dong et al., 2013) and DEFACTO

	Tested on:											
	CASIAv1				CASIAv2				DEFACTO			
	CM		SP		CM		SP		CM		SP	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Trained on:												
AUDITS+DEFACTO	**67.9**	20.3	86.9	47.6	65.3	15.1	72.8	35.0	85.3	18.0	96.0	93.7
DEFACTO	63.5	18.1	**87.0**	**51.3**	**67.8**	**16.4**	**76.0**	**39.5**	**88.7**	**22.9**	**97.1**	**67.5**
AUDITS	67.1	**20.9**	82.2	40.3	56.8	12.8	54.3	25.3	53.1	4.9	65.6	9.5

seen in Figure 11. The X-axis represents the AUC for models trained and tested on the same subset AUDITS-NEWS or AUDITS-COCO and Y-axis represents the AUC for models trained and tested on different subsets. From this analysis, we noted that objects such as "person" maintained high generalization performance across datasets, while certain object categories like "sky" exhibited a significant drop in performance. Additionally, we compared the performance of EVP and MMFusion across all object categories and found that their success and failure cases are same objects most times, suggesting that both models struggle with similar object types.

10.1 Manipulation Size Distribution

We analyze the distribution of manipulation sizes across the AUDITS-NEWS and AUDITS-COCO subsets. Figure 10 shows the number of manipulated regions categorized as small, medium, and large based on mask area. Both subsets are dominated by small manipulations, with fewer medium and large regions. We also observe differences in the relative proportions between AUDITS-NEWS and AUDITS-COCO. These variations arise from the underlying image content as well as the mask generation procedure. This visualization provides additional insight into the range of manipulation sizes present in AUDITS, which can influence the difficulty of manipulation localization.

Table 12: Comparing the performance of the EVP (Liu et al., 2023) model trained on data from either DEFACTO (MAHFOUDI et al., 2019), AUDITS-COCO or both and testing on our AUDITS dataset to determine the diffusion based inpainting performance

	Tested on:							
	AUDITS-News				AUDITS-COCO			
	MT-ID		MT-OOD		MT-ID		MT-OOD	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Trained on:								
AUDITS+DEFACTO	72.6	50.6	**65.0**	44.4	**88.8**	**51.3**	**82.1**	**41.8**
DEFACTO	55.4	29.3	55.4	29.5	68.7	28.2	72.7	32.7
AUDITS	**74.4**	**54.2**	64.6	**45.5**	85.8	49.1	81.7	39.1

Table 13: Performance of PSCC-Net and HiFi on classification tasks.

Trained on	AUDITS-News	AUDITS-COCO	AUDITS-COCO	AUDITS-News
Tested on	AUDITS-News		AUDITS-COCO	
	AUC/F1	AUC/F1	AUC/F1	AUC/F1
PSCC-Net	69.0/64.1	40.7/37.4	93.0/90.1	66.4/63.7
HiFi	54.7/54.7	50.7/34.9	56.0/55.8	53.9/53.6

11 Different Class Replacements in AUDITS-COCO and AUDITS-News

We conducted a small-scale experiment to assess the impact of image class changes on model performance. Using 100 images from AUDITS-News, we created images where new objects replaced old ones (*e.g.* replacing a car with a bike) while preserving semantic relevance. We performed the same experiment with AUDITS-COCO. Additionally, we included experiments with replacing objects with the same object type (*e.g.* replacing a bike with another bike) as done in prior experiments.

The results, presented in Table 20 and Table 23, show that for both EVP and DOLOS, the performance differences between replacing objects with a new type versus the same type are minimal.

12 Human Evaluation Results

When looking at Table 21 we can see that for Q3, GLIGEN Splicing, Blended and Stable diffusion for News tended to be the worst performing model with a number of people realizing that those images were manipulated. On average for Q3 it appears that AUDITS-COCO tended to perform better with their manipulations for instance with GLIDE and Stable Diffusion being the top performing models. Q1 reveals that most persons did realize the image were manipulated which is to be expected as diffusion based inpaintings are still not perfect on average.

In Section 3.3 of our main paper we explained how we selected 3,412 images [2]. Next, we determine which images belong to the AUDITS-News and AUDITS-COCO ID test sets, and obtain 1,121 images for AUDITS-News and 1,152 for AUDITS-COCO. Table 22 highlights the performance of the models trained in-domain and tested on these images. When comparing these results with those from Table 21 we see that even thought EVP and DOLOS tended to be the better performing models, when tested on AUDITS-COCO. They performed significantly worse than PSCC-Net for Blended Diffusion which could be explained by difficulty of of AUDITS-COCO in general. On average we see that models performed worse on AUDITS-COCO in Table 22 versus AUDITS-News and this follows with what we see in Table 21 for Q3, as persons agreed that

[2]The workers received $0.1 for completing the survey.

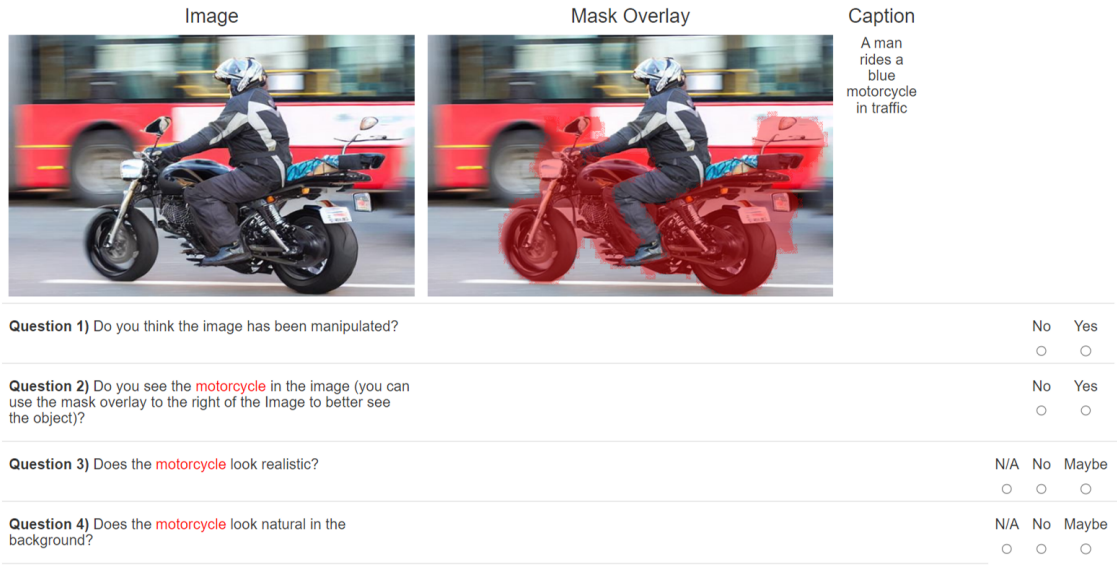

Figure 7: An example of an image from our dataset which human evaluators were given to answer questions on.

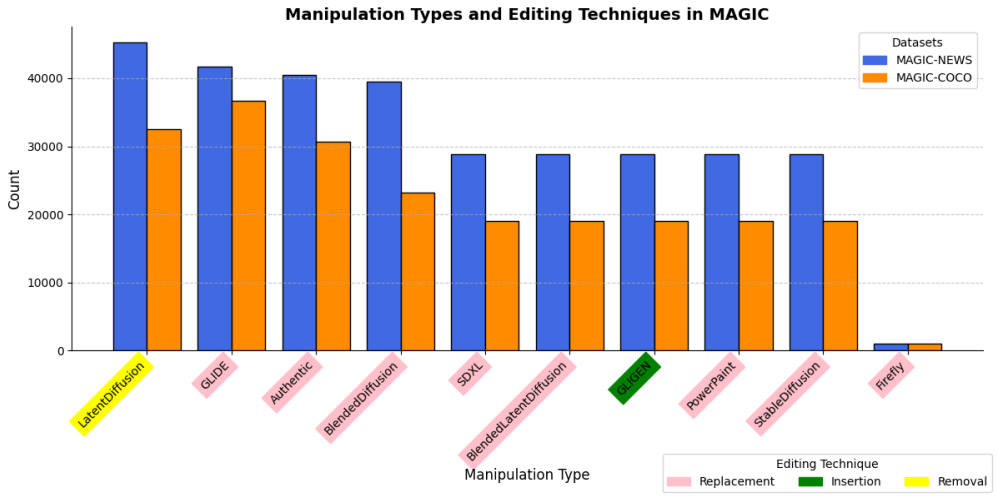

Figure 8: Visualization of manipulation types across different editing techniques from in Figure 3 of the main paper.

the manipulations from AUDITS-COCO were better on average. Hence we have shown that AUDITS-COCO is a more challenging dataset than AUDITS-News for diffusion based inpainting, and, therefore is a promising dataset for manipulation detection models to be tested on.

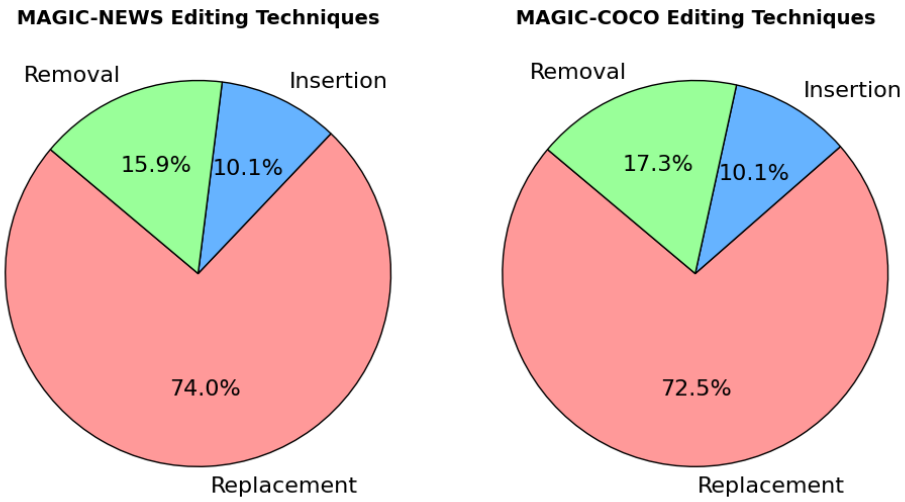

Figure 9: Distribution of different editing techniques.

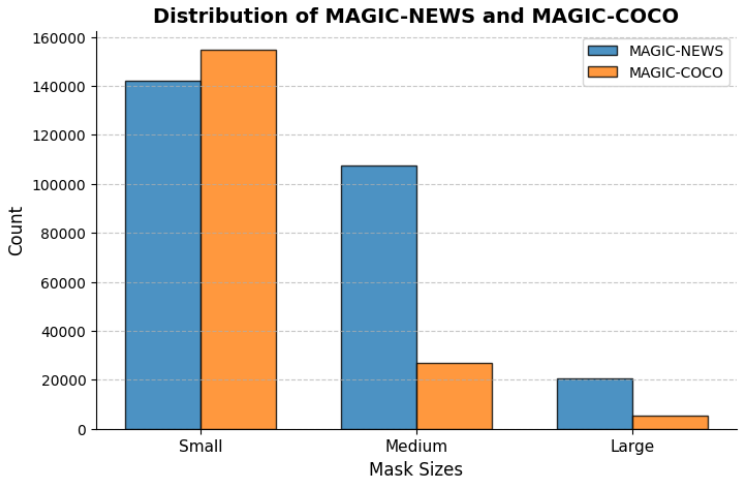

Figure 10: Visualization of manipulation sizes across different image sources from Figure 3 of the main paper.

Table 14: AUC/F1 Performance when evaluated on CasiaV1 (Dong et al., 2013) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	Copy-Move	Splicing
EVP	55.8/16.4	64.3/31.6
MMFusion	54.1/13.2	58.2/20.0
PSCC-Net	67.1/21.3	63.1/33.1
Trained on AUDITS-COCO	Copy-Move	Splicing
EVP	67.1/20.9	82.2/40.3
MMFusion	49.2/13.2	53.3/19.5
PSCC-Net	56.8/16.5	67.5/29.3

Table 15: AUC/F1 Performance when evaluated on CasiaV2 (Dong et al., 2013) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	Copy-Move	Splicing
EVP	56.8/12.8	54.3/25.3
MMFusion	47.1/12.0	48.2/21.8
PSCC-Net	54.2/12.7	50.2/27.1
Trained on AUDITS-COCO	Copy-Move	Splicing
EVP	56.8/12.8	54.3/25.3
MMFusion	52.7/11.4	52.2/21.4
PSCC-Net	52.0/12.5	59.9/27.9

Table 16: AUC/F1 Performance when evaluated on DOLOS (Țânțaru et al., 2024) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	LDM	LAMA	Pluralistic	Repaint-P2
EVP	61.7/57.4	50.0/47.6	58.7/52.6	59.4/53.5
MMFusion	42.5/57.3	42.6/53.7	43.4/55.4	42.8/55.2
PSCC-Net	49.9/0.01	54.5/52.8	48.5/0.09	49.0/13.1
Trained on AUDITS-COCO	LDM	LAMA	Pluralistic	Repaint-P2
EVP	41.8/41.0	44.9/42.1	44.1/42.1	46.0/43.4
MMFusion	43.3/52.6	42.6/48.9	43.4/52.1	43.1/51.4
PSCC-Net	50.3/0.04	54.2/53.5	50.5/0.06	50.6/0.05

Table 17: AUC/F1 Performance when evaluated on TGIF (Mareen et al., 2024) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	PS-SP	SD2-SP	SD2-FR	SDXL-FR
EVP	65.8/14.0	66.1/14.3	66.6/27.0	67.3/16.4
MMFusion	57.9/13.8	57.1/13.5	53.5/25.1	53.3/12.8
PSCC-Net	77.2/25.5	62.3/11.4	50.0/0.01	49.9/0.02
Trained on AUDITS-COCO	PS-SP	SD2-SP	SD2-FR	SDXL-FR
EVP	61.8/13.5	62.7/14.2	65.5/26.5	71.3/17.4
MMFusion	44.8/12.6	46.7/12.9	47.9/24.3	47.5/12.2
PSCC-Net	65.6/14.6	58.1/13.2	51.4/0.02	49.6/0.02

Table 18: AUC/F1 Performance when evaluated on IMD2020 (Novozamsky et al., 2020) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	Manipulated
EVP	53.5/51.1
MMFusion	50.8/55.1
PSCC-Net	49.1/40.5
Trained on AUDITS-COCO	Manipulated
EVP	42.4/36.6
MMFusion	53.2/53.9
PSCC-Net	41.1/46.9

Table 19: AUC/F1 Performance when evaluated on Autosplicing (Jia et al., 2023) trained on AUDITS-News and AUDITS-COCO respectively

Trained on AUDITS-News	JPEG100	JPEG90
EVP	81.3/66.1	77.4/61.3
MMFusion	50.3/50.6	48.8/50.3
PSCC-Net	63.0/25.2	81.6/65.5
Trained on AUDITS-COCO	Copy-Move	Splicing
EVP	71.2/58.3	82.7/63.5
MMFusion	56.9/48.7	58.5/49.0
PSCC-Net	43.2/38.7	75.8/61.5

Table 20: Performance of EVP and DOLOS on AUDITS-News with object replacements.

Trained on	AUDITS-News	AUDITS-News	AUDITS-COCO	AUDITS-COCO
Tested on	AUDITS-News	AUDITS-News	AUDITS-News	AUDITS-News
	New-Object	Replaced-Object	New-Object	Replaced-Object
	AUC/F1/IoU	AUC/F1/IoU	AUC/F1/IoU	AUC/F1/IoU
EVP	59.1/56.2/1.60	58.0/55.8/1.85	60.6/51.1/10.0	61.9/50.6/10.8
DOLOS	54.1/54.6/5.72	55.0/54.9/6.20	51.5/58.3/2.66	47.7/58.4/2.15

Table 21: The data quality survey outcomes, split by the inpainting type. We report majority vote for each question. Q1) *"Do you think this image is manipulated?"*, Q2) *"Do you see the **object** in the image (you can use the mask overlay to the right of the Image to better see the object)?"*; Q3) *"Does the **object** look realistic?"*, and Q4) *"Does the **object** look natural in the background?"*

	AUDITS-News					AUDITS-COCO			
	GLIGEN	Blended	GLIDE	Latent	Stable	Blended	GLIDE	Latent	Stable
Q1↓	84.0	84.0	80.4	81.6	79.2	81.2	80.6	84.2	79.4
Q2↑	72.6	81.4	75.4	72.0	82.0	81.2	80.6	84.2	79.4
Q3↑	48.5	51.0	59.5	54.6	54.1	57.1	60.8	54.8	61.2
Q4↑	56.1	54.2	63.1	59.5	65.2	57.7	61.4	55.8	62.2

Table 22: The analysis of the model performance (AUC) based on our data quality survey outcomes. We train our models on respective subsets and test them in domain on 1,121 images for AUDITS-News and 1,152 for AUDITS-COCO from the total 3,412 images from the Human Evaluation(see main text for discussion).

Tested on:	AUDITS-News			AUDITS-COCO		
Model	Blended	GLIDE	Latent	Blended	GLIDE	Latent
EVP	90.5	88.2	88.6	79.2	90.9	**86.1**
DOLOS	95.9	89.5	**93.4**	69.3	70.3	61.6
PSCC-Net	**97.3**	**94.8**	85.0	**95.7**	**92.7**	68.5

Table 23: Performance of EVP and DOLOS on AUDITS-COCO with object replacements.

Trained on	AUDITS-News	AUDITS-News	AUDITS-COCO	AUDITS-COCO
Tested on	AUDITS-COCO	AUDITS-COCO	AUDITS-COCO	AUDITS-COCO
	New-Object	Replaced-Object	New-Object	Replaced-Object
	AUC/F1/IoU	AUC/F1/IoU	AUC/F1/IoU	AUC/F1/IoU
EVP	87.0/53.1/35.0	86.9/53.2/34.8	60.8/26.0/1.81	61.8/25.9/0.77
DOLOS	67.2/25.0/2.93	66.8/26.1/2.85	44.8/23.7/1.62	46.1/23.9/1.21

Figure 11: Evaluation of mean AUC Object Categories using MMFusion and EVP

Table 24: AUC Performance of EVP, MMFusion and PSCC-Net on recent diffusion based inpainting models when trained on AUDITS-News.

Trained on	AUDITS-News							
Tested on	AUDITS-News				AUDITS-COCO			
	Flux-Inpt	HD-Inpt	Firefly	SDXL	Flux-Inpaint	HD-Inpaint	Firefly	SDXL
EVP	83.1	54.1	75.3	89.2	64.9	53.3	62.4	67.9
MMFusion	47.6	52.4	51.2	80.3	57.3	51.9	51.9	57.4
PSCC-Net	49.9	49.9	77.5	37.8	50.6	50.0	57.4	64.0

Table 25: AUC Performance of EVP, MMFusion and PSCC-Net on recent diffusion based inpainting models when trained on AUDITS-COCO.

Trained on	AUDITS-COCO							
Tested on	AUDITS-News				AUDITS-COCO			
	Flux-Inpt	HD-Inpt	Firefly	SDXL	Flux-Inpaint	HD-Inpaint	Firefly	SDXL
EVP	65.6	47.1	63.7	68.3	54.9	41.4	50.3	77.5
MMFusion	46.3	47.3	47.5	59.0	52.9	55.4	47.2	60.4
PSCC-Net	48.3	49.8	44.2	39.9	49.3	49.4	59.2	71.0