Detective SAM: Adapting SAM to Localize Diffusion-based Forgeries via Embedding Artifacts

Gert Lek [1] **Chaoyi Zhu** [2] **Pin-Yu Chen** [3] **Robert Birke** [4] **Lydia Y. Chen** [2 1]

Abstract

Image forgery localization in the diffusion era poses new challenges as modern editing pipelines produce photorealistic, semantically coherent manipulations that bypass conventional detectors. While some recent methods leverage foundation model cues or handcrafted noise residuals, they still miss the subtle embedding artifacts introduced by modern diffusion pipelines. In response, we develop Detective SAM, which extends the Segment Anything Model by incorporating a blur-based detection signal, learnable coarse-to-fine prompt generation, and lightweight fine-tuning for automatic forgery mask generation. Detective SAM localizes forgeries with high precision. On three challenging benchmarks (MagicBrush, CoCoGlide, and IMD2020), it outperforms prior state-of-the-art methods, demonstrating the power of combining explicit forensic perturbation cues with foundation-model adaptation for robust image forgery localization in the diffusion era.

1. Introduction

The sophistication of modern diffusion models and their local editing techniques has blurred the line between synthetic and real imagery (Ramesh et al., 2021). Deep learning has democratized photorealistic image generation, and synthetic images are now virtually indistinguishable to the naked eye. The wide availability of these techniques has caused our virtual environment to flood with such images. This transformation creates demand for robust and generalizable methods for image forgery localization (IFL) (Kadha et al., 2025).

Modern image-editing systems build directly on broader innovations in deep learning for processing visual information. Among these are Vision Transformers (ViT) (Dosovitskiy et al., 2021) and image foundation models, such as CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and other self-supervised embeddings (Assran et al., 2023). These architectures produce global (coarse, e.g., a 16×16 feature map) and local (fine, pixel-level) representations through large-scale pretraining, yielding powerful features for downstream tasks. The recently released Segment Anything Model (SAM) (Ravi et al., 2024) exemplifies this trend, serving as a strong segmentation foundation model for understanding arbitrary image regions. Moreover, SAM has been fine-tuned via lightweight adapters (Chen et al., 2024) to specialized domains such as shadow detection (Jie & Zhang, 2023) and camouflage detection (Meeran et al., 2024).

Previous forgery methods mainly focused on splicing and copy-move (Chang et al., 2013). The rise of diffusion models such as DALL-E (Ramesh et al., 2022) and their ability to generate realistic local inpaintings has caused previous methods and forensic clues to become outdated (Zhang et al., 2024). The text-guided pipelines of these models allow them to create semantically consistent content. New diffusion-based datasets such as MagicBrush (Zhang et al., 2024) expose the vulnerability of existing IFL approaches, causing a significant drop in performance for IFL tasks (Nguyen et al., 2024). The combination of diffusion-edit datasets (Zhang et al., 2024; Jia et al., 2023) and foundation models has sparked a wide array of IFL models (Lai et al., 2023; Kwon et al., 2024; Zhang et al., 2025; Nguyen et al., 2024) which successfully improve existing diffusion-based IFL benchmarks.

The paradigm shift brought on by diffusion models for image generation initiated the surge in research on stronger forensic clues. Part of this surge has seen robust empirical success with training-free (Ricker et al., 2024; Tsai et al., 2024; He et al., 2024) and zero-shot (Cozzolino et al., 2024) methods that rely on explicit noise artifacts in the embedding space. The text-to-image nature of diffusion edits has also inspired the use of Multimodal Large Language Models (MLLMs) for IFL (Liu et al., 2025). EditScout (Nguyen et al., 2024) benchmarks previous approaches on diffusion-based datasets and proposes such an MLLM-based method.

[1]University of Neuchâtel, Switzerland [2]TU Delft, The Netherlands [3]IBM Research, USA [4]University of Turin, Italy. Correspondence to: Gert Lek <gert.lek@unine.ch>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada.

The concept of leveraging linguistic cues as supplemental semantic information is promising; however, the insufficient use of forensic indicators and image foundation models may explain why performance has not yet reached expectations. Figure 1 shows that our proposed Detective SAM framework masks are more accurate with respect to the ground truth forged region when compared to EditScout.

Figure 1. Left column: source images and instructions. Right column: forged segmentation mask contours generated by EditScout (yellow) and the Detective SAM mask (red) with a white contour for the ground truth forged region. The original EditScout masks are shown in supplementary Figure 7.

Few state-of-the-art techniques have begun incorporating these recent innovations for a robust IFL system. SAM's release prompted papers to adapt the model to various tasks. A successful framework for this is provided by (Chen et al., 2024). The authors propose feature adapters as lightweight modules to fine-tune SAM for detecting different classes of segments. In the problem of IFL, the task requires segmenting the forged region and identifying where it is. This demands automatic prompting of SAM, IMDPrompter (Zhang et al., 2025) proposes a learnable prompting system based on implicit noise filters. Their method relies on dense, handcrafted, and trained noise views instead of artifacts in the embeddings of foundation models and does not beat earlier benchmarks set by (Niu et al., 2024). The IFL field currently lacks a framework that combines these innovations: (1) explicit Gaussian blur perturbation embedding artifacts as a forensic signal (2) feature adapters to fine-tune SAM for the IFL task (3) a learnable prompter based on modern vision architectures, and (4) training on modern diffusion-edited datasets. This paper offers such a unification of ideas and shows that it improves benchmark scores significantly.

We propose, *Detective SAM*, a comprehensive framework for image forgery localization. The key novelties of Detective SAM are as follows:

1. **Perturbation-driven forensic signal:** We build upon the success of perturbation embedding sensitivity as a forensic signal to reveal subtle diffusion-induced editing artifacts and integrate this in a learned IFL pipeline.
2. **Coarse-to-fine learnable prompting:** We propose a learnable prompter module that fuses the forensic signal features with the target image features for automatic prompting with coarse-to-fine localization as in ViT architectures.
3. **End-to-End SAM feature adapter integration:** We extend automatic prompting with feature adapters (Chen et al., 2024), enabling lightweight end-to-end fine-tuning of SAM's mask decoding head for forgery segmentation.
4. **State-of-the-art performance:** We demonstrate the effectiveness of this approach on recent diffusion-based IFL datasets such as MagicBrush and CoCoGlide.

2. Related work

Image forgery localization. IFL concerns itself with the task of not only detecting if parts of an image are manipulated, but also pinpointing them pixel-wise. An effective signal or "forensic clue" is required to locate image forgery. These clues/artifacts can include reconstruction error (Vesnin et al., 2024), JPEG compression artifacts (Kwon et al., 2021), explicit noise artifacts (Zhu et al., 2024), or implicit noise artifacts (Zhang et al., 2025).

Recent work has shown explicit noise artifacts in the embedding space of vision-transformer architectures and foundation model embeddings. RIGID (He et al., 2024) and BLUR (Tsai et al., 2024) show that it is possible to detect synthetic images using the DINOv2 (Oquab et al., 2024) image foundation model in a training-free manner by detecting subtle embedding distribution shifts.

SAM in IFL. Adaptations of SAM for IFL have attracted considerable interest (Kwon et al., 2024; Lai et al., 2023; Zhang et al., 2025). These methods seek to distinguish manipulated regions from genuine content by training SAM to segment forged areas in contrast to the conventional object segmentation task. For example, SAM is adapted for

deepfake detection and localization (Lai et al., 2023) with a reconstruction-error signal or used in multi-source forgery partitioning (Kwon et al., 2024) with large-scale contrastive pretraining and a fixed 16x16 point grid. However, diffusion-based tampering often manifests as subtle, irregular artifacts confined to small regions. Therefore, we require learnable prompts that dynamically adjust to the unpredictable patterns of diffusion-based forgeries. IMDPrompter (Zhang et al., 2025) achieves this with a learnable heatmap and box prompts employing various filters as the signal. This technique does not use a perturbation-driven signal or build upon the strong SAM adaptation results from (Chen et al., 2024). Other approaches use SAM's segmentation capabilities without learnable prompts (Su et al., 2024).

3. Detective SAM

3.1. Problem Setting

We consider the task of image forgery localization, where given an RGB image $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ with three channels, height H and width W, we aim to predict a binary mask $\mathcal{B} \in \{0,1\}^{H\times W}$, with $\mathcal{B}_{ij} = 1$ if pixel (i,j) has been edited once, and 0 otherwise. This work mainly focuses on edits produced by diffusion-based image-editing models. A diffusion model processes a text instruction to generate local edits of a source image, as in Figure 2.

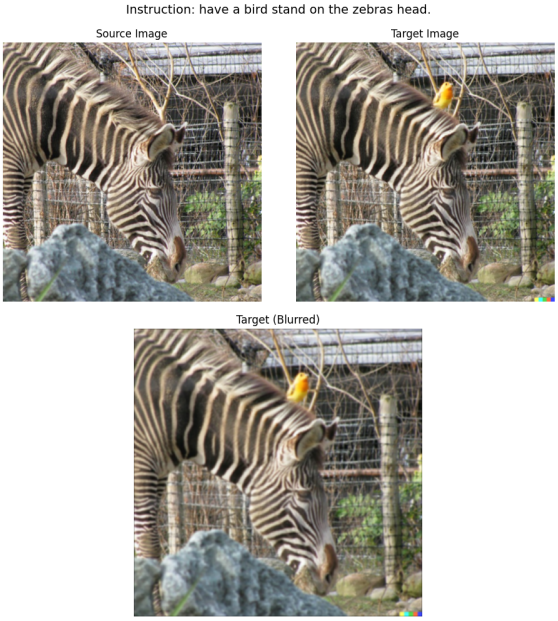

Figure 2. First row: visualization of source and target image with instruction for a sample in the MagicBrush validation set. Second row: Gaussian-blurred version of the same target image ($\sigma = 1$).

3.2. Overview

Detective SAM is a perturbation-driven extension of SAM2 (Ravi et al., 2024) for image forgery localization.
As illustrated in Figure 2, we first apply Gaussian blur to the input image $\mathcal{I}$, producing a perturbed copy $\mathcal{I}'$. Both $\mathcal{I}$ and $\mathcal{I}'$ are then passed through the frozen SAM2 HIERA encoder (Ryali et al., 2023), yielding multi-scale embeddings X_s at scales $s \in \{128, 64, 32\}$ for $(H, W) = (512, 512)$.

For both $\mathcal{I}$ and $\mathcal{I}'$, the smallest scale embedding is used as the image embedding input for our decoder, and SAM's frozen "non-memory" embedding is added to this. This embedding is learned during SAM's training (which also supports video) and flags the input as a single image rather than a video frame, giving us feature $F_{32} = \mathrm{NoMemEmbed}(X_{32})$. The other embedding scales are passed through SAM's frozen convolutional layers, which process these for decoder input; $F_s = \mathrm{ConvSAM}(X_s)$ if $s \in \{64, 128\}$. This yields six feature maps $\{F_s^{\mathcal{I}}, F_s^{\mathcal{I}'}\}$. These features are then passed through a `FeatureAdapter` A_s, which generates a correction term ΔF_s. The adapted feature is computed as: $\widetilde{F}_s = F_s^{\mathcal{I}} + \Delta F_s$.

Additionally, all feature maps $\{F_s^{\mathcal{I}}, F_s^{\mathcal{I}'}, \widetilde{F}_s\}$ are fused by our `MaskAdapter` to produce a low-resolution forgery heatmap $M \in \mathbb{R}^{128\times 128}$. Finally, the adapted features $\{\widetilde{F}_s\}$ and heatmap M are input to the SAM2 mask decoder, which outputs the binary forgery segmentation mask $\mathcal{B} \in \{0,1\}^{H\times W}$. These steps are visualized in Figure 3.

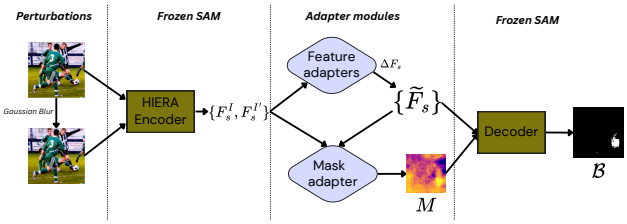

Figure 3. Flow chart of the steps in Detective SAM with our proposed modules in blue and SAM's frozen modules in green.

3.3. Perturbations

Forgery detection/localization methods require a reliable signal to predict forgery maps. Common signals are reconstruction errors (Vesnin et al., 2024), pretrained/fine-tuned embeddings on forgery datasets (Kwon et al., 2024), or perturbation-based signals. Building further upon the success of the BLUR technique (Tsai et al., 2024), we construct a perturbed image as: $\mathcal{I}' = \mathrm{Blur}(\mathcal{I}; \sigma)$ With Blur using the exact 3x3 kernel from (Tsai et al., 2024) and standard deviation $\sigma = 1.0$. Both $\mathcal{I}$ and $\mathcal{I}'$ are passed through the shared SAM2 HIERA encoder. SAM2's HIERA encoder is trained on a large-scale, diverse set of images, serving as our image foundation model. Editing techniques can pro-

duce outputs with higher sensitivity to small perturbations in the embedding space. Empirically, it is noticed that these artifacts show up in the loss landscape of the embedding space (Chen, 2025).

3.4. Adapter Modules

In the spirit of previous work on SAM adaptations (Chen et al., 2024), we create multi-scale feature adapters to fine-tune SAM. Fine-tuning focuses on aligning the feature adapters with SAM for forgery localization adaptation, but is done concurrently with the training of our mask adapter. The feature adapters are designed to incorporate our task-specific information (forensic signal) from the embeddings in the input to SAM's decoder in a lightweight manner. These adapters are single layer convolutional neural networks that process $\{F_s^{\mathcal{I}}, F_s^{\mathcal{I}'}\}$ to get $\{\widetilde{F}_s\}$. The feature adapters fine-tune SAM by modifying the unperturbed feature with the prompt term ΔF_s.

Unlike straightforward adaptations of SAM in prior approaches, we require automatic prompting of the SAM decoder for our target task since it is unknown to the user where the forgery is located. Therefore, a neural network is required to process the forensic signal into a prompt. SAM expects either points, boxes, or heatmap inputs. Thus, we design a mask adapter that takes the feature maps $\{F_s^{\mathcal{I}}, F_s^{\mathcal{I}'}, \widetilde{F}_s\}$ and fuses them with gated feature fusion to produce heatmap M as input to the SAM decoder. The mask adapter is structured to generate a spatially coherent, globally contextualized heatmap by:

1. **Multi-scale feature fusion:** We fuse $\{F_s^{\mathcal{I}}, F_s^{\mathcal{I}'}\ \widetilde{F}_s\}$ with convolutions to an embedding $F_{\text{fuse}} \in \mathbb{R}^{d\times\hat{s}\times\hat{s}}$ with common resolution $\hat{s} = \max\{s\}$, defined as the largest scale. This integrates information from all levels to detect and localize forgery.

2. **Coarse patch scoring:** Significant parts of the image have low importance since they are unforged. To mitigate instability in those regions, we capture high-level spatial importance by tiling F_{fuse} into a grid of size $\left(\lfloor\hat{s}/r\rfloor,\ \lfloor\hat{s}/r\rfloor\right)$ with downscale factor r using a strided convolution. This yields coarse logits $\tilde{L}_{\text{coarse}}$. Cross-attention builds queries from pooled fused features and keys/values from the flattened coarse logits, enabling each patch to gather context from the entire coarse map. This global interaction smooths/removes heatmap "islands" to produce the refined logits L_{coarse}. The number of attention parameters is limited since we work in the downscaled domain.

3. **Full-resolution refinement:** We refine our coarse logits and fused feature map F_{fuse} in a convolutional refine-head to produce high-resolution logits $L_{\text{refine}} \in \mathbb{R}^{\hat{s}\times\hat{s}}$.

4. **Learned gating:** Finally, we blend our coarse and refined logits with $g = \text{sigmoid}(w_1 L_{\text{coarse}} + w_2 L_{\text{refine}} + b)$ with only three parameters: $\{w_1, w_2, b\}$, the final heatmap logits are $M = gL_{\text{refine}} + (1-g)L_{\text{coarse}}$. This suppresses spurious refinements in regions where the coarse mask is confident by allowing the importance of both the coarse and fine components to be learned. Specifically, the gating weight g_{ij} decreases when the coarse logit is high, such that the final mask falls back on the reliable coarse prediction.

Notice that no sigmoid function is applied to mask heatmap M, since SAM's decoder expects logits. M is fed into SAM's decoder as the heatmap prompt to produce the final pixel heatmap $\mathcal{B}$.
We observe that a simple convolution from F_{fuse} to M produces spurious and noisy forgery scores. The idea of dividing the fine space into patches to integrate global context is similar to vision transformers (Dosovitskiy et al., 2021).

3.5. Loss functions

We train the mask and feature adapters using the same objectives as SAM (Chen et al., 2024), combining a focal loss (Lin et al., 2018) with a Dice loss. The Dice loss maximizes the overlap between the predicted and ground-truth masks by penalizing their normalized differences. The focal loss further addresses the class imbalance in IFL by down-weighting well-classified pixels and up-weighting forged pixels relative to the abundant negative background.

Formally, our total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \lambda_{\text{focal}}\,\mathcal{L}_{\text{focal}}^{\alpha,\gamma}.$$

The focusing parameter $\gamma \geq 0$ in $\mathcal{L}_{\text{focal}}^{\alpha,\gamma}$ down-weights well-classified examples. The balance factor $\alpha \in [0, 1]$ re-weights positive vs negative examples to counteract class imbalance. Where we borrow $\lambda_{\text{focal}} = 20$ from the SAM2 paper and sweep over (α, γ).

4. Experiments

4.1. Evaluation

Our performance is evaluated with the pixel-level mean Intersection over Union (IoU) and the mean F1 score. The IoU measures the overlap between the ground truth forged mask and $\mathcal{B}$, and the F1 score serves as the harmonic mean between pixel-level precision and recall.

Datasets: The MagicBrush dataset contains high-quality diffusion-based edits using DALL-E (Ramesh et al., 2022). Magicbrush has multiple edit rounds. Edit rounds are the number of distinct local edits in one image. To align with the training regime of EditScout (Nguyen et al., 2024), we only

use single edits for this paper, giving 4512 samples. Besides MagicBrush, we also use AutoSplice (Jia et al., 2023) for training, which comprises diffusion-based DALL-E edits with 3621 samples.

We train Detective SAM on the training set of MagicBrush (Zhang et al., 2024) and AutoSplice (Jia et al., 2023) to arrive at the same training regime as EditScout (Nguyen et al., 2024) for a fair comparison. Similar to EditScout, we evaluate our model on the 512 samples in the CoCoGLIDE dataset and the 801 samples of the MagicBrush validation and test set. Additionally, we evaluate our model on the 2010 samples in the test set of IMD20 (Novozámský et al., 2020) to compare our model's out-of-sample performance against the benchmarks implemented in (Niu et al., 2024). We choose IMD20 because it is the lowest-scoring dataset in that paper and in (Zhang et al., 2025).

4.2. Results

Quantitative results: The benchmark results in Table 1 are copied from the EditScout paper. The PerfBrush dataset is cut as it is not yet publicly available. We only display the best scores of the table here; the full table is available in the appendix Table 2.

Table 1. Best EditScout benchmark results and Detective SAM.

METHOD	MAGICBRUSH		COCOGLIDE	
	IOU ↑	F1 ↑	IOU ↑	F1 ↑
BEST	30.47	40.35	34.11	45.70
DETECTIVE SAM	**49.22**	**60.24**	**35.54**	**46.86**

Detective SAM outperforms the MagicBrush results by a significant margin, but the CoCoGLIDE results are close in terms of IoU.

Similarly, the full IMD20 benchmarks from (Niu et al., 2024) are in the appendix under Table 3. The best benchmark IoU is 19.2 for MVSS-Net (Dong et al., 2023) and the best F1 is **58.9** for the (Niu et al., 2024) model. Detective SAM achieves an IoU of **41.94** and F1 score of 52.33 on IMD20. The significant divergence between F1 and IoU in this table for some models can be explained by overestimating the small forged region. Conversely, Detective SAM's IoU-to-F1 ratio is consistent across datasets, indicating stronger generalization. The SAM-based IMDPrompter F1 results are added to the table without the IoU metric since it is not reported.

Qualitative Results: Our method produces a refined heatmap M and coarse heatmap $M_{\text{coarse}} = \text{sigmoid}(L_{\text{coarse}})$. In Figure 4, notice that our coarse patches down-weight unconfident regions, yielding a final prediction that successfully highlights the forged area. More samples are visualized in the appendix Section B.

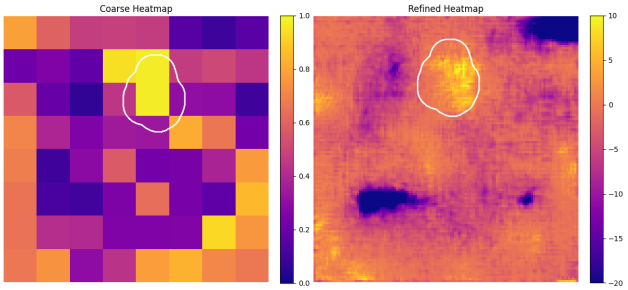

Figure 4. Coarse heatmap M_{coarse} and refined heatmap M, warmer colors indicate higher probabilities/logits of forgery. The contour of the ground truth mask is overlaid.

SAM's decoder processes heatmap M to produce a per-pixel segmentation heatmap (Figure 5), where the adapted SAM pushes background areas into the negative successfully because of its segmentation capabilities and adapted fine-tuning.

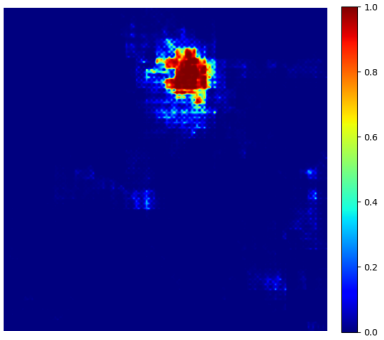

Figure 5. Adapted SAM segmentation output before thresholding at level 0.5, warmer colors indicate higher probabilities of forgery.

Thresholding at 0.5 yields the final binary segmentation mask $\mathcal{B}$ shown in Figure 6. The figure shows that forged regions do not correspond one-to-one with the added object (bird). The divergence is by design: (1) IFL models are trained on the ground truth mask of the edit instruction, (2) the inpainting model often blends the added object with its environment (Figure 2), leading to a forged surrounding region. This observation serves as a case in point on why we require fine-tuning of SAM besides just prompt learning: to change the focus from pure object segmentation to forged region segmentation.

Figure 6. The final binary forgery segmentation mask $\mathcal{B}$ overlaid on the original target image.

5. Conclusion

Detective SAM sets a new bar for localizing (diffusion-based) image forgery with an average increase in IoU of 14.31 over our three test sets. This paper demonstrates that IFL methods significantly improve with strong forensic signals, adapted segmentation models as a backbone, and multi-scale mask refinement. Our evaluation confirms that these ingredients discover subtle artifacts, yielding more accurate masks and better generalization.

5.1. Limitations

Despite these advances, several avenues to improvement remain open. In future work, we plan to conduct a component-wise ablation study and extend evaluation to classical IFL datasets such as Coverage (Wen et al., 2016), Columbia (Hsu & Chang, 2006), CASIA (Dong et al., 2013), and multi-edit datasets. We will explore domain generalization to deepfake and video forgery localization. Last but not least, we will explore varied perturbations and adaptive perturbations. By articulating these steps, we aim to advance the IFL field further to keep pace with the evolving generative editing tools.

5.2. Impact Statement

This paper presents an approach to further the field of image forgery localization. Undetected manipulations of visual content pose serious risks, including the facilitation of misinformation campaigns and the harming of societal trust in digital media. We, therefore, provide a framework to compete with innovations in modern local forgery tools.

Acknowledgments

This research is part of the Priv-GSyn project, 200021E_229204 of Swiss National Science Foundation and the DEPMAT project, P20-22 / N21022, of the research programme Perspectief which is partly financed by the Dutch Research Council (NWO). This work was partly supported by the Spoke 1 "FutureHPC & BigData" of ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, funded by the European Union - NextGenerationEU.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Chang, I.-C., Yu, J. C., and Chang, C.-C. A forgery detection algorithm for exemplar-based inpainting images using multi-region relation. *Image Vision Comput.*, 31(1):57–71, January 2013. ISSN 0262-8856. doi: 10.1016/j.imavis.2012.09.002. URL https://doi.org/10.1016/j.imavis.2012.09.002.

Chen, P.-Y. Computational safety for generative ai: A signal processing perspective, 2025. URL https://arxiv.org/abs/2502.12445.

Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., Li, Z., Sun, L., Mao, P., and Zang, Y. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more, 2024. URL https://arxiv.org/abs/2408.04579.

Cozzolino, D., Poggi, G., Nießner, M., and Verdoliva, L. Zero-shot detection of ai-generated images, 2024. URL https://arxiv.org/abs/2409.15875.

Dong, C., Chen, X., Hu, R., Cao, J., and Li, X. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2023. doi: 10.1109/TPAMI.2022.3180556.

Dong, J., Wang, W., and Tan, T. 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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., and Verdoliva, L. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization, 2023. URL https://arxiv.org/abs/2212.10957.

Guo, K., Zhu, H., and Cao, G. Effective image tampering localization via enhanced transformer and co-attention fusion. In *ICASSP*, 2024.

Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., and Liu, X. Hierarchical fine-grained image forgery detection and localization, 2023. URL https://arxiv.org/abs/2303.17111.

He, Z., Chen, P.-Y., and Ho, T.-Y. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection, 2024. URL https://arxiv.org/abs/2405.20112.

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

Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., and Nevatia, R. 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*, pp. 312–328, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58588-4. doi: 10.1007/978-3-030-58589-1_19. URL https://doi.org/10.1007/978-3-030-58589-1_19.

Jia, S., Huang, M., Zhou, Z., Ju, Y., Cai, J., and Lyu, S. Autosplice: A text-prompt manipulated image dataset for media forensics, 2023. URL https://arxiv.org/abs/2304.06870.

Jie, L. and Zhang, H. Adaptershadow: Adapting segment anything model for shadow detection, 2023. URL https://arxiv.org/abs/2311.08891.

Kadha, V. K., Bakshi, S., and Das, S. K. Unravelling digital forgeries: A systematic survey on image manipulation detection and localization. *ACM Comput. Surv.*, April 2025. ISSN 0360-0300. doi: 10.1145/3731243. URL https://doi.org/10.1145/3731243.

Kwon, M.-J., Yu, I.-J., Nam, S.-H., and Lee, H.-K. Catnet: Compression artifact tracing network for detection and localization of image splicing. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 375–384, 2021. doi: 10.1109/WACV48630.2021.00042.

Kwon, M.-J., Lee, W., Nam, S.-H., Son, M., and Kim, C. Safire: Segment any forged image region, 2024. URL https://arxiv.org/abs/2412.08197.

Lai, Y., Luo, Z., and Yu, Z. Detect any deepfakes: Segment anything meets face forgery detection and localization, 2023. URL https://arxiv.org/abs/2306.17075.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

Liu, J., Zhang, F., Zhu, J., Sun, E., Zhang, Q., and Zha, Z.-J. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization, 2025. URL https://arxiv.org/abs/2410.10238.

Liu, W., Shen, X., Pun, C.-M., and Cun, X. Explicit visual prompting for low-level structure segmentations, 2023. URL https://arxiv.org/abs/2303.10883.

Liu, X., Liu, Y., Chen, J., and Liu, X. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization, 2022. URL https://arxiv.org/abs/2103.10596.

Meeran, M. N., T, G. A., and Mantha, B. P. Sam-pm: Enhancing video camouflaged object detection using spatio-temporal attention, 2024. URL https://arxiv.org/abs/2406.05802.

Nguyen, Q., Vu, T., Nguyen, T.-T., Wen, Y., Robinette, P. K., Johnson, T. T., Goldstein, T., Tran, A., and Nguyen, K. Editscout: Locating forged regions from diffusion-based edited images with multimodal llm, 2024. URL https://arxiv.org/abs/2412.03809.

Niu, Y., Chen, P., Zhang, L., Tan, L., and Chen, Y. Image forgery localization via guided noise and multi-scale feature aggregation, 2024. URL https://arxiv.org/abs/2412.01622.

Novozámský, A., Mahdian, B., and Saic, S. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 71–80, 2020. doi: 10.1109/WACVW50321.2020.9096940.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Ricker, J., Lukovnikov, D., and Fischer, A. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error, 2024. URL https://arxiv.org/abs/2401.17879.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., and Feichtenhofer, C. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023. URL https://arxiv.org/abs/2306.00989.

Su, Y., Tan, S., and Huang, J. A novel universal image forensics localization model based on image noise and segment anything model. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, pp. 149–158, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706370. doi: 10.1145/3658664.3659639. URL https://doi.org/10.1145/3658664.3659639.

Tsai, C.-T., Ko, C.-Y., Chung, I.-H., Wang, Y.-C. F., and Chen, P.-Y. Understanding and improving training-free ai-generated image detections with vision foundation models, 2024. URL https://arxiv.org/abs/2411.19117.

Vesnin, D., Levshun, D., and Chechulin, A. Detecting autoencoder is enough to catch ldm generated images, 2024. URL https://arxiv.org/abs/2411.06441.

Wen, B., Zhu, Y., Subramanian, R., Ng, T.-T., Shen, X., and Winkler, S. Coverage – a novel database for copy-move forgery detection. In *IEEE International Conference on Image processing (ICIP)*, pp. 161–165, 2016.

Wu, Y., AbdAlmageed, W., and Natarajan, P. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhang, K., Mo, L., Chen, W., Sun, H., and Su, Y. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. URL https://arxiv.org/abs/2306.10012.

Zhang, Q., Qi, Y., Tang, X., Fang, J., Lin, X., Zhang, K., and Yuan, C. Imdprompter: Adapting sam to image manipulation detection by cross-view automated prompt learning, 2025. URL https://arxiv.org/abs/2502.02454.

Zhu, J., Li, D., Fu, X., Yang, G., Huang, J., Liu, A., and Zha, Z.-J. Learning discriminative noise guidance for image forgery detection and localization, 2024. URL https://doi.org/10.1609/aaai.v38i7.28608.

A. Method

Table 2. MagicBrush and CoCoGLIDE segmentation results from EditScout (Nguyen et al., 2024) with the Detective SAM row added. The star * denotes additional fine-tuning by the authors of EditScout.

Method	MagicBrush		CoCoGLIDE	
	IoU ↑	F1 ↑	IoU ↑	F1 ↑
PSCC-Net [2022]	8.35	12.30	14.46	20.24
EITL-Net [2024]	7.88	11.38	28.79	35.42
TruFor [2023]	19.47	26.93	29.26	36.08
HiFi [2023]	5.10	8.22	16.55	23.44
CAT-Net [2021]	2.71	4.33	31.63	39.18
PSCC-Net * [2022]	16.82	26.50	15.02	20.75
EITL-Net * [2024]	20.02	28.09	19.15	26.34
CAT-Net * [2021]	30.47	40.35	31.79	41.12
EditScout [2024]	23.77	33.19	34.11	45.70
Detective SAM	**49.22**	**60.24**	**35.54**	**46.86**

Table 3. IMD2020 segmentation results from (Niu et al., 2024), with the Detective SAM row added.

Method	IMD2020	
	IoU ↑	F1 ↑
ManTra [2019]	12.4	18.3
SPAN [2020]	10.0	14.5
PSCC-Net [2022]	12.0	19.7
MVSS [2023]	19.2	26.0
HiFi [2023]	8.0	53.2
EVP [2023]	18.3	23.3
IMDPrompter [2025]	-	30.6
(Niu et al., 2024)	17.0	**58.9**
Detective SAM	**41.94**	52.33

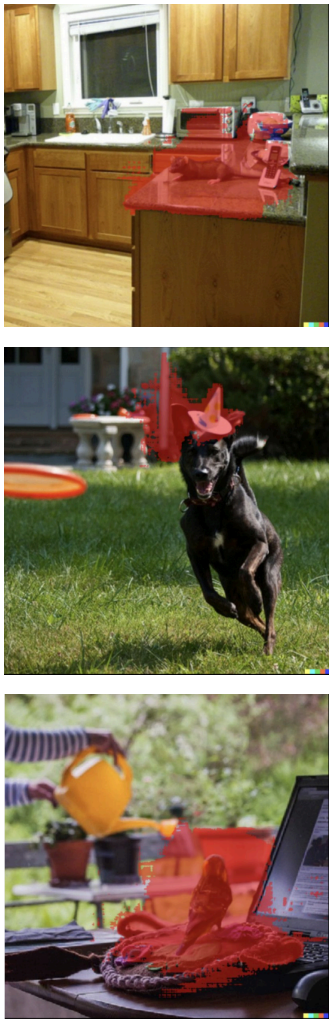

Figure 7. EditScout target images where the contours in Figure 1 are taken from. Images are extracted from the EditScout (Nguyen et al., 2024) paper since the model/code is not publicly available.

B. Experiment visualizations

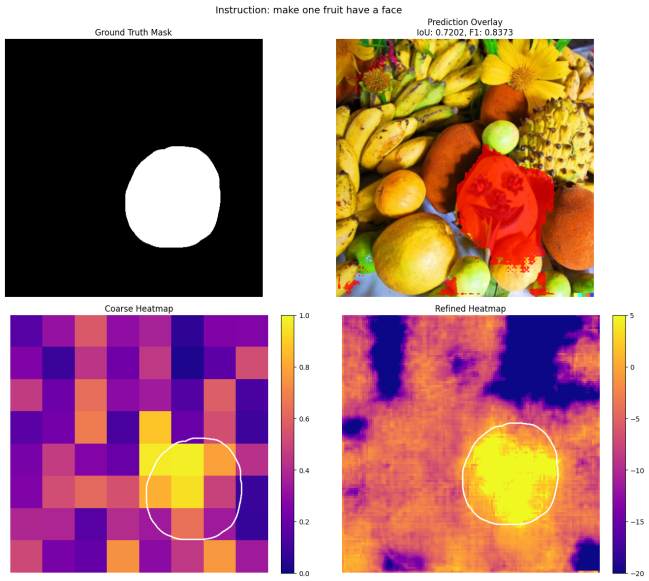

Figure 8. Ground truth mask, forgery mask prediction, coarse heatmap, and refined heatmap visualization for the instruction: ”Make one fruit have a face”.

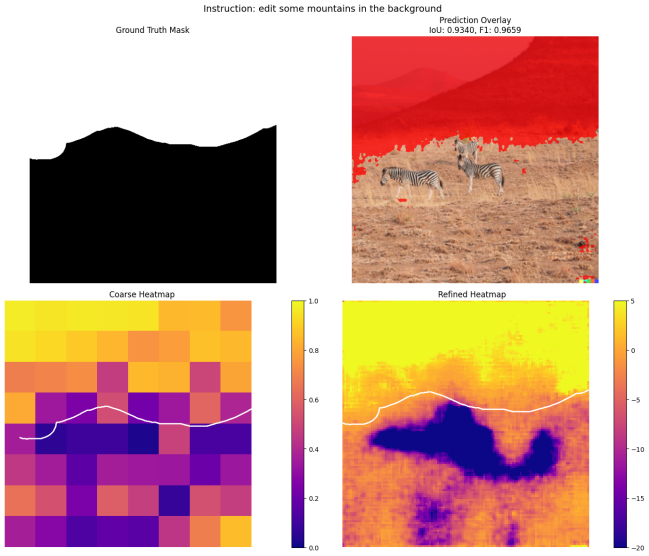

Figure 9. Ground truth mask, forgery mask prediction, coarse heatmap, and refined heatmap visualization for the instruction: ”Edit some mountains in the background”.

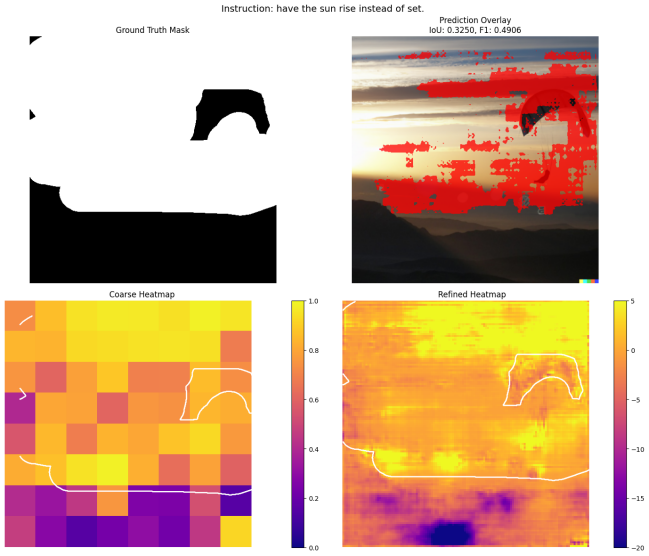

Figure 10. Ground truth mask, forgery mask prediction, coarse heatmap, and refined heatmap visualization for the instruction: ”Have the sun rise instead of set”.

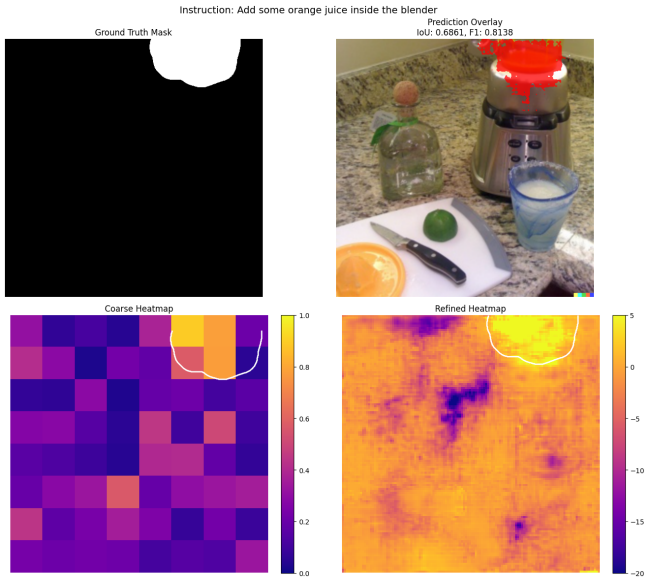

Figure 11. Ground truth mask, forgery mask prediction, coarse heatmap, and refined heatmap visualization for the instruction: ”Add some orange juice inside the blender”.