FINE-GRAINED IMAGE TAMPERING LOCALIZATION VIA SEMANTIC PROPOSALS AND ADAPTIVE RESIDUAL REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Proactive image protection enables subsequent tampering to induce observable responses in pretrained vision models, yet these responses are often too coarse and spatially inconsistent to support fine-grained localization. We investigate this observability-to-localization gap in a proactive protection setting, where a protected reference and its manipulated counterpart are available. We propose **SPARR (Semantic Proposal and Adaptive Residual Refinement)**, a response-interpretation framework that converts manipulation-sensitive responses into accurate tampering masks while keeping the upstream protection mechanism unchanged. SPARR first exploits paired response variations under shared spatial prompts to derive robust semantic proposals, and then uses paired RGB residuals to refine response-induced regions toward the actual manipulated structures. To account for sample-dependent reliability, SPARR further introduces an adaptive routing strategy that evaluates whether additional refinement is supported by ground-truth-free evidence and selectively accepts or rejects it. This design enables SPARR to improve localization without relying on a fixed refinement behavior across different manipulation patterns. We evaluate SPARR under controlled same-protection comparisons, five conventional image tampering benchmarks, and four generative-AI editing settings. The results show that reliable downstream response interpretation can substantially improve fine-grained localization without modifying the upstream protection mechanism, while maintaining strong performance across diverse editing settings. The code will be made publicly available upon acceptance.

1 INTRODUCTION

image tampering localization aims to delineate the spatial extent of manipulated content rather than merely determine image-level authenticity. Modern image editing tools and generative models can produce increasingly realistic local modifications, making reliable pixel-level localization an important problem in multimedia forensics. Most existing methods follow a passive paradigm and infer forensic evidence directly from the observed image, exploiting manipulation artifacts, local inconsistencies, boundary cues, or learned forensic representations Zhou et al. (2018); Dong et al. (2023); Guillaro et al. (2023). Despite substantial progress, because such evidence must be recovered after manipulation has already occurred, maintaining reliable localization under previously unseen manipulation distributions remains challenging Dong et al. (2023); Asnani et al. (2023).

A complementary direction is proactive image protection, where manipulation-sensitive information is established before image distribution so that subsequent edits become easier to reveal. Existing approaches introduce protective templates or imperceptible watermarks for later localization Asnani et al. (2023); Zhang et al. (2024; 2025). More recently, Blank Canvas Song et al. (2026) introduces a perception-based protection paradigm that suppresses the normal responses of a pretrained Segment Anything Model (SAM) Kirillov et al. (2023) on the protected image. Subsequent manipulation locally disrupts this suppressed state and exposes manipulation-sensitive responses. This mechanism makes tampering *observable*; however, the exposed responses are not themselves pixel-accurate localization masks. In other words, proactive protection addresses how tampering can be revealed,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

while a separate question remains: *how should the revealed evidence be interpreted into accurate spatial predictions?*

We study this *observability-to-localization gap* under a proactive image protection setting, where a protected reference I^P and its potentially manipulated counterpart I^T are available. We keep the upstream protection mechanism fixed and focus exclusively on downstream response interpretation. Converting observable responses into fine-grained masks involves three characteristic failure modes. First, the exposed responses are prompt-dependent: different spatial prompts and multi-mask outputs may emphasize different suspicious structures, making a fixed probing strategy unreliable across samples. Second, semantic responses need not coincide with the actual edited pixels; a semantically plausible region may still over-cover the manipulation or miss locally modified structures. Third, structural refinement is inherently sample-dependent: a correction that improves one prediction may over-expand, over-contract, or otherwise degrade another. These observations suggest that fine-grained localization requires progressively resolving semantic ambiguity, structural mismatch, and refinement risk rather than simply thresholding a single exposed response or applying a fixed correction to every sample.

To this end, we propose **SPARR**, a Semantic Proposal and Adaptive Residual Refinement framework that progressively interprets manipulation-sensitive evidence while leaving the upstream protection mechanism unchanged. The first stage, Adaptive Multi-Prompt Semantic Proposals (AMSP), aggregates paired multi-mask response variations and adaptively retains informative response-guided prompts to construct a robust semantic proposal. The second stage, Paired RGB-Residual Structural Refinement (PRSR), complements semantic evidence with paired RGB residuals and locally corrects mismatches between the semantic proposal and the physical modification structure. The final stage, Sample-Adaptive Refinement Decision (SARD), avoids applying a fixed correction to every sample by evaluating alternative structural refinements using ground-truth-free sample-level evidence and predicted gains, accepting additional refinement only when sufficiently supported. Together, these stages instantiate a single progressive response-interpretation process, moving from semantic evidence aggregation, through residual-guided structural correction, to sample-adaptive refinement control.

We evaluate SPARR through controlled same-protection comparisons, five conventional image tampering benchmarks, and four generative-AI editing settings. The controlled setting keeps the upstream protection condition and manipulated image pairs fixed, thereby isolating the contribution of downstream response interpretation. Because passive and proactive approaches operate under different information conditions, passive methods are included as task-level references rather than strictly identical-input competitors. The broader evaluations further examine whether the resulting localization behavior remains effective across the evaluated datasets and editing scenarios. The results show that stronger downstream interpretation can substantially improve fine-grained localization without modifying the upstream protection mechanism.

Contributions. Our contributions are threefold:

- We formulate the *observability-to-localization gap* in proactive image protection, explicitly distinguishing the exposure of manipulation-sensitive responses from their conversion into pixel-accurate localization.
- We introduce **SPARR**, a progressive response-interpretation framework that sequentially addresses prompt-dependent semantic ambiguity, semantic–structural mismatch, and sample-dependent refinement risk while keeping the upstream protection mechanism fixed.
- We provide controlled same-protection evidence that isolates the contribution of downstream response interpretation, and further validate SPARR across conventional tampering benchmarks and diverse generative-AI editing settings.

2 Related Work

Passive Image Tampering Localization. Passive image forensics localizes manipulated content by recovering forensic inconsistencies directly from the observed image. Early deep methods combine RGB-domain artifacts with low-level noise traces or learn manipulation-sensitive anomalies from diverse editing operations Zhou et al. (2018); Wu et al. (2019). Subsequent approaches im-

prove spatial reasoning and generalization through multi-scale attention or correlation modeling Hu et al. (2020); Liu et al. (2022), boundary- and noise-aware representations Dong et al. (2023), hierarchical forgery features Guo et al. (2023), and learned noise-sensitive fingerprints Guillaro et al. (2023). More recent work such as FakeShield Xu et al. (2025) further extends passive localization toward explainable multimodal forensic analysis. Despite substantial progress, passive methods must infer localization evidence from the manipulated observation itself. Their effectiveness therefore depends on whether sufficiently discriminative forensic traces remain after editing, and reliable localization can become more difficult when manipulation distributions, visual content, or editing artifacts differ from those seen during training.

Proactive Image Protection for Tamper Localization. Proactive methods establish manipulation-sensitive information before image distribution. MaLP Asnani et al. (2023) learns a protective template to facilitate subsequent localization, whereas EditGuard Zhang et al. (2024) and OmniGuard Zhang et al. (2025) employ watermark-based protection and extraction for tamper localization and related provenance tasks. Blank Canvas Song et al. (2026) follows a different perception-based paradigm: instead of recovering an embedded signal, it suppresses the normal responses of a pretrained SAM Kirillov et al. (2023), so that later manipulation exposes anomalous model responses. Existing proactive approaches substantially improve manipulation observability, but the exposed evidence is not itself guaranteed to coincide with the spatial extent of the actual edit. Embedded or learned protection signals still require reliable recovery after manipulation, while perception-based responses may remain spatially coarse or locally misaligned with modified pixels. Consequently, accurately converting exposed manipulation-sensitive evidence into fine-grained localization masks remains insufficiently addressed.

3 METHOD

We consider fine-grained tamper localization under proactive image protection. Given a protected reference image I^P, its manipulated counterpart I^T, and a fixed pretrained SAM model Φ, our goal is to predict a pixel-level tampering mask $\hat{M} \in \{0,1\}^{H\times W}$. In our implementation, the upstream protection is instantiated with the basic Blank Canvas mechanism Song et al. (2026), and both the protection stage and Φ remain fixed. Ground-truth masks may supervise the learned selection models during training, but are never accessed when producing test-time predictions.

SPARR progressively converts paired manipulation-sensitive evidence into a fine-grained localization mask through

$$(I^P, I^T) \to M^{\mathrm{sem}} \to M^{\mathrm{prsr}} \to \hat{M}. \tag{1}$$

Adaptive Multi-Prompt Semantic Proposals (AMSP) constructs M^{sem} from paired multi-mask response variations and response-guided prompt selection. Paired RGB-Residual Structural Refinement (PRSR) then uses paired RGB residuals as local corrective evidence to obtain M^{prsr}. Finally, Sample-Adaptive Refinement Decision (SARD) evaluates alternative structural refinements and either accepts a sufficiently supported candidate or falls back to M^{prsr}. Figure 1 summarizes the complete pipeline and the information flow among the three stages.

3.1 ADAPTIVE MULTI-PROMPT SEMANTIC PROPOSALS

AMSP addresses prompt-dependent ambiguity by extracting manipulation-sensitive semantic evidence from paired SAM responses rather than a single prompted prediction. Given I^P and I^T, we apply the same 4×4 Grid-16 spatial prompts $\mathcal{P}_b$ to both images. For each prompt p, the fixed SAM returns K multi-mask logits L_p^P and L_p^T. Their paired variation is aggregated across mask channels as

$$\Delta L_p = |L_p^T - L_p^P|, \qquad R_p(x) = \sqrt{\sum_{k=1}^{K} \big(\Delta L_{p,k}(x)\big)^2}. \tag{2}$$

Each response is robustly normalized using its 1st and 99th percentiles, and the base responses are aggregated by a pixel-wise median:

$$\bar{R}_p = \mathcal{N}_{0.01,0.99}(R_p), \qquad R^{\mathrm{coarse}}(x) = \mathrm{median}_{p\in\mathcal{P}_b}\, \bar{R}_p(x). \tag{3}$$

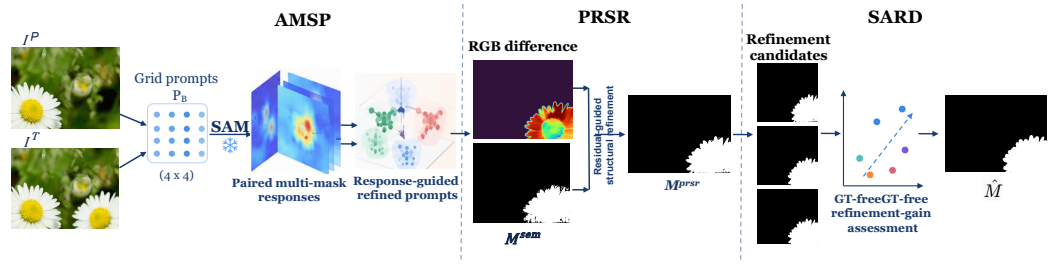


Figure 1: Overall framework of SPARR. The upstream proactive protection stage and pretrained SAM remain fixed. SPARR progressively interprets paired manipulation-sensitive evidence through AMSP semantic proposal construction, PRSR residual-guided structural refinement, and SARD sample-adaptive refinement decision to obtain the final tampering mask.

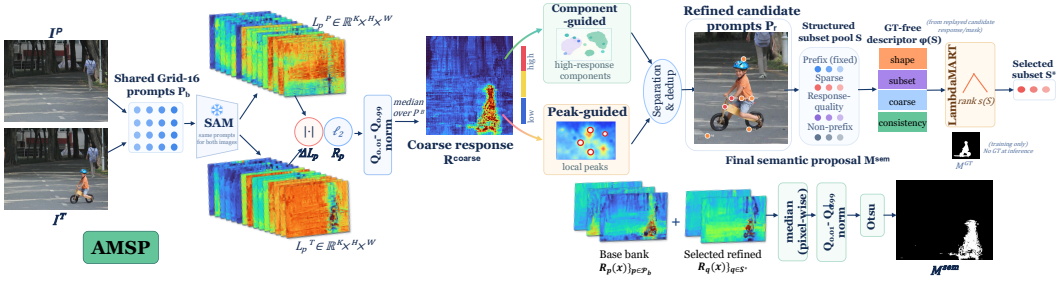


Figure 2: Illustration of Adaptive Multi-Prompt Semantic Proposals (AMSP). Shared paired probing first constructs manipulation-sensitive responses under Grid-16 prompts. The coarse response then guides component-level and peak-level refined prompt generation. A structured candidate pool is ranked using ground-truth-free descriptors, and all base responses are finally fused with the selected refined responses to obtain the semantic proposal M^{sem}. Ground truth is used only for training the ranking model and is not accessed at inference time.

Figure 2 illustrates the complete AMSP pipeline, including paired response construction, response-guided prompt generation, structured subset ranking, and final response fusion.

The coarse response provides a stable global estimate of manipulation-sensitive regions and guides adaptive probing. We generate refined prompt candidates $\mathcal{P}_r$ from connected high-response components and strong local response peaks, followed by spatial separation and deduplication; at most 12 candidates are retained.

Rather than independently scoring individual prompts or selecting a fixed top-k prefix, AMSP performs structured subset selection over $\mathcal{P}_r$. The subset pool $\mathcal{S}$ contains prefix, sparse, response-quality-guided, and non-prefix combinations. Each subset is replayed together with all Grid-16 base prompts to form a candidate fused response and semantic mask. A ground-truth-free candidate descriptor summarizes coarse-response statistics, subset composition, mask geometry, and response–mask consistency. Following the frozen ranking procedure, LambdaMART Burges (2010) first establishes a reference candidate from the base-only/prefix candidates and then ranks the structured candidates using their evidence relative to this reference. The selected subset is

$$S^* = \arg\max_{S \in \mathcal{S}} \big(s(S), -|S|\big)_{\text{lex}}, \tag{4}$$

where subset size is used only as a tie breaker. Ground-truth masks are used to construct ranking supervision during training, but neither the candidate descriptors nor test-time subset selection accesses ground truth.

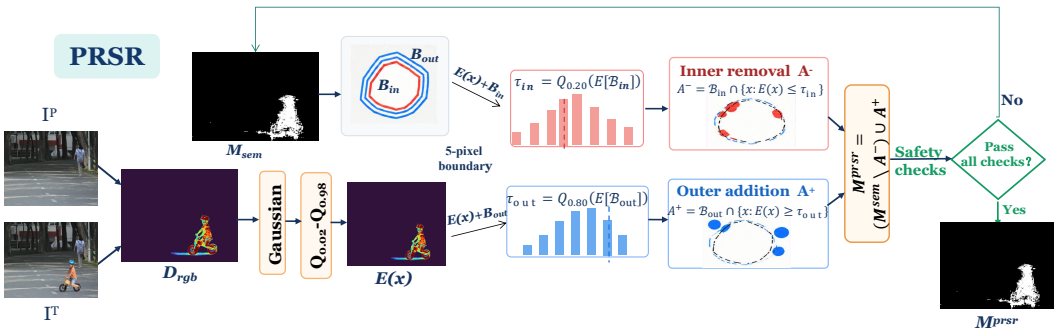


Figure 3: Illustration of Paired RGB-Residual Structural Refinement (PRSR). Starting from the AMSP semantic proposal, PRSR constructs paired RGB residual evidence, restricts correction to inner and outer boundary bands, removes weakly supported inner pixels, recovers strongly supported outer pixels, and applies conservative validity checks before producing M^{prsr}.

Finally, all Grid-16 base responses are retained and augmented only with the refined responses selected by S^*:

$$R^{\mathrm{fuse}}(x) = \mathrm{median}\left(\{\bar{R}_p(x)\}_{p\in\mathcal{P}_b} \cup \{\bar{R}_q(x)\}_{q\in S^*}\right). \tag{5}$$

After robust normalization,

$$\widetilde{R}^{\mathrm{fuse}} = \mathcal{N}_{0.01,0.99}(R^{\mathrm{fuse}}), \tag{6}$$

Otsu thresholding Otsu (1979) produces the semantic proposal

$$M^{\mathrm{sem}}(x) = \mathbf{1}\left[\widetilde{R}^{\mathrm{fuse}}(x) \ge \tau_{\mathrm{Otsu}}\right]. \tag{7}$$

AMSP therefore uses uniform paired probing to determine where manipulation-sensitive evidence emerges, and then adaptively decides where to probe further and which refined prompts to retain.

3.2 Paired RGB-Residual Structural Refinement

AMSP provides a manipulation-sensitive semantic proposal, but its extent may still follow object-level semantics rather than the pixels actually modified by the edit. PRSR addresses this semantic–structural mismatch using paired image-space evidence. Importantly, the residual is used as local corrective evidence rather than an independent localization cue: the semantic proposal determines where manipulation is likely to occur, while paired RGB residuals determine how its local structure should be adjusted. Figure 3 summarizes the PRSR refinement process.

Given M^{sem}, we compute

$$D_{\mathrm{rgb}}(x) = \frac{1}{3}\sum_{c\in\{R,G,B\}} |I_c^T(x) - I_c^P(x)|, \qquad E(x) = \mathcal{N}_{0.02,0.98}\left[G_{\sigma=1.2} * D_{\mathrm{rgb}}\right](x). \tag{8}$$

Rather than thresholding $E(x)$ globally, PRSR restricts correction to 5-pixel inner and outer bands, $\mathcal{B}_{\mathrm{in}}$ and $\mathcal{B}_{\mathrm{out}}$, around the boundary of M^{sem}. Sample-specific thresholds are

$$\tau_{\mathrm{in}} = Q_{0.20}(E[\mathcal{B}_{\mathrm{in}}]), \qquad \tau_{\mathrm{out}} = Q_{0.80}(E[\mathcal{B}_{\mathrm{out}}]). \tag{9}$$

Weakly supported inner pixels are removed, whereas strongly supported outer pixels are recovered:

$$A^- = \mathcal{B}_{\mathrm{in}} \cap \{x : E(x) \le \tau_{\mathrm{in}}\}, \qquad A^+ = \mathcal{B}_{\mathrm{out}} \cap \{x : E(x) \ge \tau_{\mathrm{out}}\}. \tag{10}$$

The provisional structural refinement is

$$\widetilde{M}^{\mathrm{prsr}} = (M^{\mathrm{sem}} \setminus A^-) \cup A^+. \tag{11}$$

This local asymmetric correction contracts unsupported semantic regions while recovering residual-supported structures beyond the original boundary, without allowing residual evidence to redefine the mask globally. We retain only connected components overlapping the original semantic proposal. If the original or refined mask is empty, or if the relative area change exceeds 12%, the correction is rejected and PRSR falls back to M^{sem}; otherwise, $\widetilde{M}^{\mathrm{prsr}}$ is accepted as M^{prsr}.

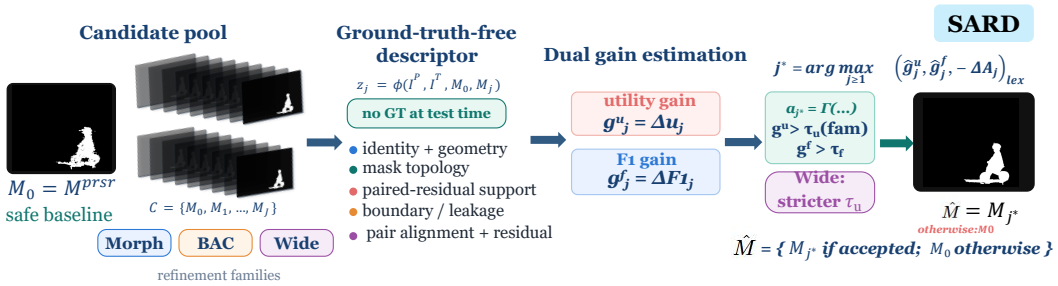


Figure 4: Illustration of Sample-Adaptive Refinement Decision (SARD). Starting from the PRSR baseline, SARD evaluates candidates from the Morph, BAC, and Wide refinement families using ground-truth-free sample-level descriptors, predicts candidate gains, ranks the alternatives, and accepts further refinement only when the evidence satisfies the decision thresholds; otherwise it retains the PRSR result.

3.3 Sample-Adaptive Refinement Decision

PRSR provides a structurally corrected mask, but no single additional refinement is consistently beneficial across samples. SARD therefore treats the final stage as evidence-gated refinement selection rather than unconditional post-processing. SARD introduces no new localization cue; instead, it predicts whether a candidate structural correction is likely to improve the current PRSR result. Figure 4 illustrates the candidate evaluation and refinement-decision process.

Let

$$M_0 = M^{\mathrm{prsr}}, \qquad \mathcal{C} = \{M_0, M_1, \ldots, M_J\}, \tag{12}$$

where the remaining candidates are generated by the fixed Morph, BAC, and Wide refinement families. For each M_j, SARD constructs a 23-D ground-truth-free descriptor

$$z_j = \phi(I^P, I^T, M_0, M_j), \tag{13}$$

summarizing candidate geometry and topology, paired-residual support, boundary/background-leakage statistics, and pair-quality evidence.

During training, candidate benefit is defined relative to M_0:

$$U_j = 0.5\,\Delta\mathrm{IoU}_j + 0.3\,\Delta\mathrm{BF1}_j + 0.2\,\Delta\mathrm{F1}_j, \tag{14}$$

where each metric difference is computed between candidate M_j and baseline M_0. Two Ridge regressors predict the utility gain and F1 gain,

$$\hat{g}^u_j = f_u(z_j), \qquad \hat{g}^f_j = f_f(z_j). \tag{15}$$

Ground-truth masks are used only to construct these training targets; descriptors, test-time gain prediction, ranking, and final selection do not access ground truth.

Candidates are prioritized by

$$j^* = \arg\max_{j\geq 1} \left(\hat{g}^u_j, \hat{g}^f_j, -\Delta A_j\right)_{\mathrm{lex}}, \tag{16}$$

where ΔA_j denotes the absolute area change of M_j relative to M_0. The top-ranked candidate is not adopted automatically. It is accepted only when

$$\hat{g}^u_{j^*} > 0.002, \qquad \hat{g}^f_{j^*} > -0.003, \tag{17}$$

with a stricter utility threshold of 0.010 for the Wide family. The final prediction is

$$\hat{M} = \begin{cases} M_{j^*}, & \text{if the acceptance conditions are satisfied,} \\ M_0, & \text{otherwise.} \end{cases} \tag{18}$$

SARD therefore converts refinement from a fixed post-processing operation into a sample-adaptive decision: candidate generation proposes structural alternatives, predicted gains estimate their expected benefit, and the acceptance gate retains the PRSR baseline whenever further modification is insufficiently supported. Figure 5 shows a representative sample-adaptive decision example.

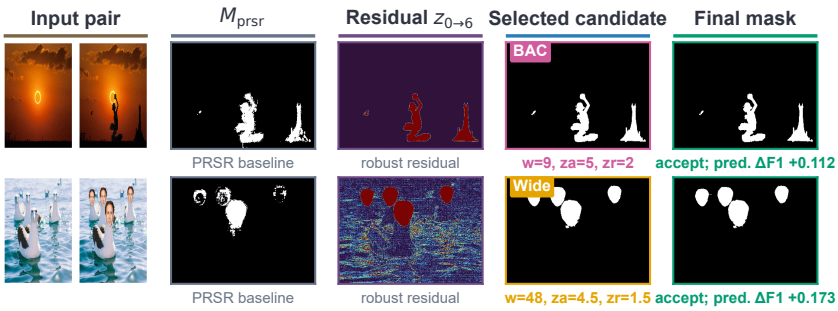


Figure 5: The top row shows an accepted BAC refinement, while the bottom row shows an accepted Wide refinement. The visualization shows the PRSR baseline, refinement candidates, sample-level evidence and predicted gains, the selected candidate, and the final accept-or-fallback decision leading to $\hat{M}$.

4 Experiments

4.1 Experimental Setup and Evaluation Protocol

We evaluate SPARR on five conventional image tampering benchmarks: IMD2020 Novozamsky et al. (2020) ($N = 2010$), CASIA1 Dong et al. (2013); Pham et al. (2019) ($N = 920$), Columbia Hsu & Chang (2006) ($N = 180$), NIST16 Guan et al. (2019) ($N = 496$), and Korus Korus & Huang (2017) ($N = 220$). For AIGC-based editing, we use the AGE-Set dataset Zhang et al. (2024) ($N = 1000$) and follow the evaluation framework of EditGuard Zhang et al. (2024) to evaluate four AIGC editing methods: SD-Inpaint Rombach et al. (2022), SDXL Podell et al. (2024), ControlNet Zhang et al. (2023), and RePaint Lugmayr et al. (2022). We further conduct component analysis to examine the contributions of SPARR's successive inference stages.

All SPARR evaluations follow the paired proactive setting defined in Sec. 3, where both the protected reference I^P and manipulated counterpart I^T are available, while the upstream Blank Canvas protection stage and pretrained SAM remain fixed.

The learned selection components are separated from test-time localization. In AMSP, ground-truth masks are used only to construct ranking supervision for the LambdaMART selector; in SARD, ground truth is used only to construct candidate-level refinement-gain targets for two Ridge regressors. Candidate descriptors, gain prediction, ranking, and all test-time selection decisions remain ground-truth free. The frozen SARD implementation uses five-fold Ridge averaging with $\alpha = 10.0$; the learned models and decision thresholds are fixed before evaluation.

For IMD2020, we explicitly distinguish a 1,000-sample development portion from the remaining 1,010 held-out samples. Predictions on the 1,000-sample portion are obtained out of fold, whereas the remaining 1,010 samples are evaluated using the frozen learned selectors. The full-dataset IMD2020 result reported later therefore aggregates these 1,000 out-of-fold predictions and 1,010 held-out predictions.

We report pixel-level F1 score and Intersection over Union (IoU) as the primary localization metrics. Boundary F1 (BF1) is additionally reported in the component analysis to evaluate structural boundary quality.

4.2 Performance Comparisons

Conventional Tampering Benchmarks. Table 1 compares localization performance on five conventional tampering benchmarks. SPARR achieves the strongest reported F1/IoU on CASIA1, IMD2020, Columbia, and Korus, while OmniGuard reports higher values on NIST16. The results indicate that progressively interpreting the exposed responses is effective across several manipulation distributions: AMSP reduces dependence on a single prompt response, while PRSR and SARD further correct local structural mismatch and sample-dependent refinement errors.

Table 1: Quantitative comparison on conventional image tampering benchmarks. Best results are shown in bold and second-best results are underlined. For shared baselines, we use the published results reported by Blank Canvas Song et al. (2026).

		CASIA1		IMD2020		Columbia		NIST16		Korus	
Method	Venue	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑
ManTraNet Wu et al. (2019)	CVPR	0.130	0.090	0.160	0.100	0.070	0.040	0.200	0.140	0.050	0.020
SPAN Hu et al. (2020)	ECCV	0.140	0.110	0.140	0.090	0.200	0.140	0.210	0.160	0.100	0.060
CAT-Net Kwon et al. (2021)	WACV	0.510	0.440	0.190	0.140	0.130	0.080	0.190	0.140	0.060	0.040
OSN Wu et al. (2022)	CVPR	0.510	0.470	0.470	0.380	0.690	0.580	0.330	0.250	0.190	0.140
PSCC-Net Liu et al. (2022)	TCSVT	0.460	0.360	0.320	0.220	0.740	0.640	0.260	0.180	0.220	0.150
MVSS-Net Dong et al. (2023)	TPAMI	0.480	0.400	0.310	0.230	0.610	0.480	0.290	0.240	0.170	0.120
HiFi-Net Guo et al. (2023)	CVPR	0.180	0.130	0.140	0.090	0.110	0.060	0.130	0.090	0.020	0.010
EditGuard Zhang et al. (2024)	CVPR	0.670	0.600	0.620	0.550	0.780	0.700	0.400	0.350	0.280	0.220
FakeShield Xu et al. (2025)	ICLR	0.620	0.560	0.580	0.520	0.760	0.680	0.390	0.340	0.260	0.220
OmniGuard Zhang et al. (2025)	CVPR	0.585	0.514	0.717	0.625	0.890	0.845	**0.762**	**0.698**	0.777	0.709
Blank Canvas Song et al. (2026)	AAAI	0.670	0.620	0.660	0.580	0.810	0.740	0.450	0.310	0.310	0.270
SPARR (Ours)	–	**0.893**	**0.831**	**0.854**	**0.790**	**0.921**	**0.872**	0.728	0.634	**0.865**	**0.787**

Table 2: Quantitative comparison across four generative-AI image editing settings. Best results are shown in bold and second-best results are underlined. † denotes MVSS-Net fine-tuned with the AIGC-edited dataset. For shared baselines, we use the published results reported by Blank Canvas Song et al. (2026).

		SD-Inpaint		SDXL		ControlNet		RePaint	
Method	Venue	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑	F1↑	IoU↑
PSCC-Net Liu et al. (2022)	TCSVT	0.166	0.112	0.189	0.115	0.177	0.116	0.140	0.109
MVSS-Net Dong et al. (2023)	TPAMI	0.178	0.103	0.037	0.028	0.178	0.103	0.104	0.082
MVSS-Net† Dong et al. (2023)	TPAMI	0.694	0.575	0.482	0.359	0.678	0.558	0.185	0.111
HiFi-Net Guo et al. (2023)	CVPR	0.547	0.128	0.828	0.261	0.542	0.123	0.681	0.339
EditGuard Zhang et al. (2024)	CVPR	0.966	0.936	0.965	0.936	0.968	0.940	0.967	0.938
OmniGuard Zhang et al. (2025)	CVPR	0.947	0.916	0.753	0.687	0.971	0.947	0.629	0.544
Blank Canvas Song et al. (2026)	AAAI	0.972	0.958	0.970	0.958	**0.973**	0.938	0.961	**0.957**
SPARR (Ours)	–	**0.974**	**0.959**	**0.975**	**0.960**	0.972	**0.956**	**0.971**	0.953

Generative-AI Editing. Table 2 further evaluates four generative-AI editing settings. SPARR remains tightly clustered across SD-Inpaint, SDXL, ControlNet, and RePaint, with only small variation across editors. Overall stability suggests that paired semantic responses and residual evidence still hold reference value in the four evaluated generative editing mechanisms.

Qualitative Comparison. Figure 6 complements the quantitative results with six challenging examples. Across the selected cases, SPARR produces more spatially coherent manipulated regions and suppresses scattered background responses that appear in several comparison methods. The examples also show that the remaining errors are mainly associated with local boundary mismatch or incomplete structural coverage, which is consistent with the motivation for residual-guided refinement rather than simply increasing the strength of the semantic response.

4.3 Ablation Studies

To isolate the contribution of each stage under a consistent protocol, Table 3 reports a progressive ablation on the 1,010 held-out IMD2020 samples. All configurations use the same upstream protection, pretrained SAM, paired inputs, and evaluation samples, and all learned selectors remain frozen; only the downstream response-interpretation stages are changed.

Starting from AMSP, PRSR improves F1, IoU, and BF1 by approximately 0.035, 0.052, and 0.026, respectively. The gains are moderate but consistent across region- and boundary-level metrics, supporting the intended role of paired RGB residuals as local structural evidence: PRSR does not

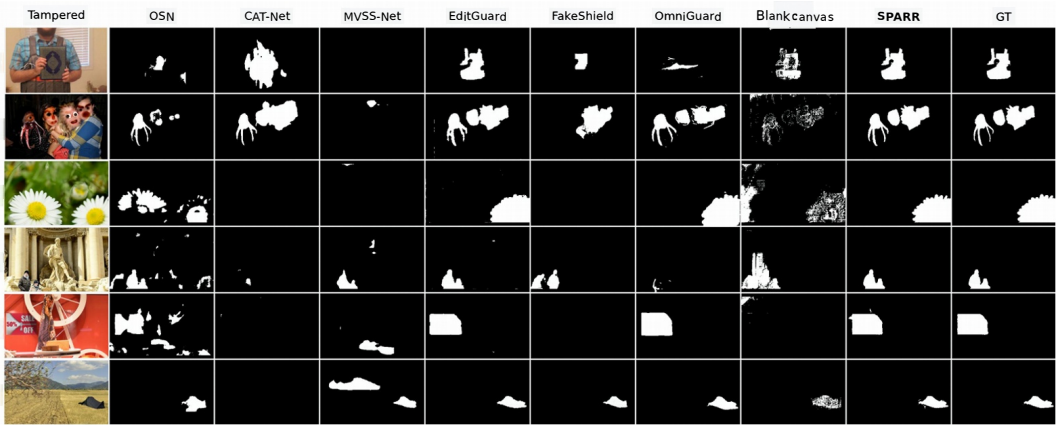


Figure 6: Qualitative comparison on six challenging manipulation examples. Columns show the tampered image, Comparison method and the ground-truth mask.

Table 3: Progressive ablation on the 1,010 held-out IMD2020 samples.

Configuration	AMSP	PRSR	SARD	F1↑	IoU↑	BF1↑
AMSP	✓			0.728	0.614	0.302
AMSP + PRSR	✓	✓		0.763	0.666	0.329
SPARR	✓	✓	✓	**0.840**	**0.774**	**0.558**

replace the semantic proposal, but corrects pixels near its boundary when the image-space evidence disagrees with the semantic extent.

Adding SARD produces a larger improvement of approximately 0.077 F1, 0.108 IoU, and 0.229 BF1 over AMSP+PRSR. The particularly large BF1 gain indicates that the final quality is strongly affected by whether an additional structural correction is appropriate for the current sample. This supports the use of sample-adaptive refinement selection instead of applying one fixed post-processing rule to every prediction. Importantly, the gain is aggregate rather than sample-wise monotonic: a minority of held-out samples still degrade relative to PRSR, showing that the predicted refinement gains are not perfectly calibrated and that refinement-risk estimation remains a limitation.

Overall, the ablation supports the progressive design of SPARR. AMSP first establishes manipulation-sensitive semantic support, PRSR then corrects local semantic–structural mismatch, and SARD finally decides whether further refinement is sufficiently supported. The three stages therefore contribute sequentially rather than as interchangeable or parallel components.

5 Conclusion

This paper presents SPARR for the *observability-to-localization gap* in proactive image protection. With the upstream protection mechanism and pretrained SAM fixed, AMSP builds semantic proposals from paired multi-prompt responses, PRSR uses paired RGB residuals for local correction, and SARD makes a ground-truth-free refinement decision at inference. Experiments on five conventional benchmarks, four generative-AI editing settings, and controlled component analyses show complementary contributions from the three stages.

The method still depends on the upstream mechanism exposing responses that distinguish manipulated from unchanged regions, and weak or misaligned responses remain difficult to localize. The evaluation also uses fixed protection and vision backbones. Future work will test broader protection settings and study tighter coupling between protection and localization to improve stability across editing conditions.

REFERENCES

Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. MaLP: Manipulation localization using a proactive scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12343–12352, June 2023. doi: 10.1109/CVPR52729.2023.01188.

Christopher J. C. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-82, Microsoft Research, June 2010.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. 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.

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 (ChinaSIP)*, pp. 422–426, July 2013. doi: 10.1109/ChinaSIP.2013.6625374.

Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel F. Zhou, Timothee N. Kheyrkhah, Jeff Smith, and Jonathan G. Fiscus. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 63–72, 2019. doi: 10.1109/WACVW.2019.00018.

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. doi: 10.1109/CVPR52729.2023.01974.

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 (CVPR)*, pp. 3155–3165, June 2023. doi: 10.1109/CVPR52729.2023.00308.

Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 549–552, July 2006. doi: 10.1109/ICME.2006.262447.

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*, pp. 312–328, 2020. doi: 10.1007/978-3-030-58589-1_19.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023. doi: 10.1109/ICCV51070.2023.00371.

Pawel Korus and Jiwu Huang. Multi-scale analysis strategies in PRNU-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4):809–824, 2017. doi: 10.1109/TIFS.2016.2636089.

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 375–384, January 2021. doi: 10.1109/WACV48630.2021.00042.

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. doi: 10.1109/TCSVT.2022.3189545.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, June 2022. doi: 10.1109/CVPR52688.2022.01117.

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. doi: 10.1109/WACVW50321.2020.9096940.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Nam Thanh Pham, Jong-Weon Lee, Goo-Rak Kwon, and Chun-Su Park. Hybrid image-retrieval method for image-splicing validation. *Symmetry*, 11(1):83, 2019. doi: 10.3390/sym11010083.

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 *International Conference on Learning Representations (ICLR)*, 2024.

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. doi: 10.1109/CVPR52688.2022.01042.

Qi Song, Ziyuan Luo, and Renjie Wan. Creating blank canvas against AI-enabled image forgery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):9088–9096, 2026. doi: 10.1609/aaai.v40i11.37865.

Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13440–13449, June 2022. doi: 10.1109/CVPR52688.2022.01308.

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 (CVPR)*, pp. 9543–9552, June 2019. doi: 10.1109/CVPR.2019.00977.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023. doi: 10.1109/ICCV51070.2023.00355.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. EditGuard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11964–11974, June 2024. doi: 10.1109/CVPR52733.2024.01137.

Xuanyu Zhang, Zecheng Tang, Zhipei Xu, Runyi Li, Youmin Xu, Bin Chen, Feng Gao, and Jian Zhang. OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3008–3018, June 2025. doi: 10.1109/CVPR52734.2025.00286.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1053–1061, June 2018. doi: 10.1109/CVPR.2018.00116.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation and code organization. Their primary uses included language polishing and expression refinement, suggestions for manuscript structure and content organization, assistance with code inspection and debugging, consistency checking between the manuscript and implementation, and preparation of anonymized code and submission materials. Generative AI was also used to assist in checking experimental procedures, implementation

details, and technical descriptions in the manuscript, but it was not used as a source of experimental results, evaluation metrics, or research conclusions.

All AI-assisted text, code modification suggestions, and analytical outputs were manually reviewed, verified, and revised by the authors as necessary. The authors were responsible for confirming the method design, experimental setup, experimental data, quantitative results, and final research conclusions. The authors take full responsibility for the entire manuscript, all research claims, the reported experimental results, and the submitted code.

A Additional Experimental Protocol and Ground-Truth Boundary

All SPARR evaluations follow the paired proactive setting used in the main paper: the protected reference I^P and manipulated counterpart I^T are both available, while the upstream protection mechanism and pretrained SAM remain fixed. The supplementary analyses below do not alter the upstream protection condition, paired inputs, frozen selectors, or held-out evaluation set.

A.1 Development, Out-of-Fold Prediction, and Held-Out Evaluation

For IMD2020, 1,000 samples form the development portion and the remaining 1,010 samples form the held-out portion. Predictions for the 1,000 development samples are obtained out of fold (OOF). The 1,010 held-out samples are evaluated only after the learned selectors and decision thresholds are frozen. The full IMD2020 result reported in the main paper therefore aggregates 1,000 OOF predictions and 1,010 held-out predictions; all per-sample analyses in Sections C–E use only the 1,010 held-out samples.

Table 4: IMD2020 development and held-out protocol.

Split	Samples	Prediction protocol
Development	1,000	5-fold OOF selector predictions
Held-out	1,010	Frozen selector inference
Full IMD2020	2,010	1,000 OOF + 1,010 held-out

A.2 Ground-Truth Usage Boundary

Ground truth is permitted for offline supervision and evaluation, but it is not an input to frozen test-time localization. In AMSP, GT is used to construct ranking supervision for the learned selector. In SARD, GT is used to construct candidate-level utility and F1-gain targets. PRSR is deterministic and GT-free. Candidate descriptors, frozen gain prediction, candidate ranking, acceptance/fallback decisions, and final test-time mask generation do not access GT.

Table 5: Ground-truth usage boundary. "Offline" includes training-target construction and metric computation.

Operation	Offline GT use	Frozen test-time GT input
AMSP ranking supervision	Yes	No
AMSP candidate descriptors / frozen ranking	No	No
PRSR structural refinement	No	No
SARD refinement-gain targets	Yes	No
SARD 23-D descriptor / frozen routing	No	No
Offline F1 / IoU / BF1 evaluation	Yes	Not part of prediction

Ground-truth masks shown in the qualitative figures are therefore visualization references only and are not used to produce the displayed frozen predictions.

B Additional Implementation Details

AMSP. The shared base prompt bank is a 4×4 Grid-16 applied identically to I^P and I^T. For each prompt, AMSP takes the absolute difference between paired SAM multi-mask logits, computes an ℓ_2 response across mask channels, normalizes each response by its 1st and 99th percentiles, and

forms the coarse response with a pixel-wise median. High-response connected components and local peaks generate at most 12 spatially separated refined prompt candidates. The frozen candidate pool contains base-only, prefix, sparse, response-quality-guided, and non-prefix subsets; every candidate retains all Grid-16 base responses and adds the selected refined responses. The frozen AMSP selection stack contains a five-fold 71-feature anchor ranker and a five-fold 105-feature structured LambdaMART ranker. Both operate on GT-free deployment features. The selected response is median-fused, robustly normalized, and binarized by Otsu thresholding to obtain M^{sem}.

PRSR. PRSR computes the channel-mean absolute RGB difference between I^P and I^T, applies Gaussian smoothing with $\sigma = 1.2$, and robustly normalizes the residual using the 2nd and 98th percentiles. Refinement is restricted to 5-pixel inner and outer bands around M^{sem}. The inner removal threshold is the 20th percentile of residual evidence in the inner band, while the outer addition threshold is the 80th percentile in the outer band. After removal/addition, PRSR keeps only connected components overlapping the original semantic proposal. If a valid refined mask cannot be formed or its relative area change exceeds 12%, the method falls back to M^{sem}.

SARD candidate families. SARD starts from $M_0 = M^{\mathrm{prsr}}$ and generates fixed Morph, BAC, and Wide alternatives. Exact duplicate candidate masks are removed before scoring. The frozen candidate families are summarized in Table 6.

Table 6: Frozen SARD candidate families and validity gates.

Family	Candidate settings	Max. rel. area change	Max. topology excess
Morph	erode/dilate $r \in \{1, 2, 3\}$; open/close $r \in \{1, 2\}$	0.20	1
BAC	$w \in \{3, 5, 9\}$; $(z_a, z_r) \in \{(2.5, 1.0), (3.5, 1.5), (5.0, 2.0)\}$	0.25	1
Wide	$(w, z_a, z_r) \in \{(24, 3.0, 1.0), (24, 4.5, 1.5), (48, 3.0, 1.0), (48, 4.5, 1.5)\}$	0.55	2

For each candidate M_j, SARD constructs a 23-D GT-free descriptor containing 10 geometry/topology quantities, 9 residual-support and boundary/leakage quantities, and 4 pair-quality quantities. During training, candidate utility is

$$U_j = 0.5\,\Delta\mathrm{IoU}_j + 0.3\,\Delta\mathrm{BF1}_j + 0.2\,\Delta\mathrm{F1}_j. \tag{19}$$

Two Ridge regressors predict utility gain and F1 gain, and the frozen model uses five-fold averaging with $\alpha = 10$. Candidate ranking and the final accept-or-fallback decision follow the procedure in Sec. 3.3 of the main paper. The frozen acceptance thresholds are 0.002 for predicted utility gain and -0.003 for predicted F1 gain, with the stricter utility threshold 0.010 for the Wide family. If the acceptance conditions are not satisfied, SARD retains the PRSR baseline M_0.

Reviewer artifact scope. The reviewer artifact package contains the frozen learned components and lightweight verification and inspection utilities needed to inspect their structure and GT independence. Historical research and evaluation drivers may read GT when constructing supervision, oracle diagnostics, or reported metrics; this is distinct from the frozen deployment decisions above. The artifact is not presented as a redistribution of the benchmark datasets, SAM checkpoint, or the complete historical upstream preparation pipeline.

C Extended Quantitative Analysis

C.1 Held-Out Distribution Statistics

Table 7 reports means together with medians and interquartile ranges (IQRs) on the 1,010 held-out IMD2020 samples. The mean values reproduce the progressive ablation in the main paper, while the median/IQR statistics show that the final stage also narrows the central F1 and IoU spread.

Table 7: Distribution statistics on the 1,010 held-out IMD2020 samples.

	F1			IoU			BF1		
Stage	Mean	Median	IQR	Mean	Median	IQR	Mean	Median	IQR
AMSP	0.7277	0.8017	0.2383	0.6136	0.6690	0.3157	0.3023	0.2907	0.2893
AMSP + PRSR	0.7631	0.8599	0.2525	0.6660	0.7543	0.3568	0.3286	0.2890	0.4711
SPARR	0.8402	0.9371	0.1401	0.7741	0.8817	0.2334	0.5577	0.6146	0.5134

C.2 PRSR Sample-Wise Effects

Table 8: Per-sample PRSR effects, categorized by the sign of ΔF1 from AMSP to PRSR.

Category	Samples	Fraction	Mean ΔF1	Mean ΔBF1
Improved	659	0.6525	+0.0553	+0.0516
Unchanged	304	0.3010	+0.0000	+0.0000
Degraded	47	0.0465	-0.0143	-0.1585
Overall	1,010	1.0000	+0.0354	+0.0263

PRSR improves F1 on 659 samples, leaves 304 unchanged, and degrades 47. The net held-out gain is therefore moderate but broadly distributed rather than being driven by only a few extreme samples.

C.3 SARD Decision and Candidate-Family Analysis

Table 9: Frozen SARD decisions by the final selected option. Metric changes are recomputed from the final held-out evaluation relative to the PRSR baseline.

Selected option	Samples	Fraction	Mean ΔF1	Mean ΔBF1
PRSR baseline / fallback	19	0.0188	+0.0000	+0.0000
Morph	370	0.3663	+0.0196	+0.2082
BAC	363	0.3594	+0.0767	+0.2600
Wide	258	0.2554	+0.1661	+0.2323

SARD accepts a non-baseline refinement for 991/1,010 samples (98.12%) and falls back to PRSR for 19/1,010 (1.88%). Relative to PRSR, final SPARR F1 improves on 892/1,010 samples (88.32%), is unchanged on 19/1,010 (1.88%), and degrades on 99/1,010 (9.80%). Among the 991 accepted refinements, the mean changes are ΔF1=+0.0787 and ΔBF1=+0.2335. Thus, the aggregate gain is substantial, but the learned refinement-risk estimate is not sample-wise monotonic.

D Additional Module-Wise Qualitative Results

All examples in this section are held-out IMD2020 samples. The displayed intermediate responses, masks, and SARD decisions are frozen inference artifacts. Ground truth is included only for post-hoc visual comparison.

D.1 AMSP: Adaptive Semantic Proposal Construction

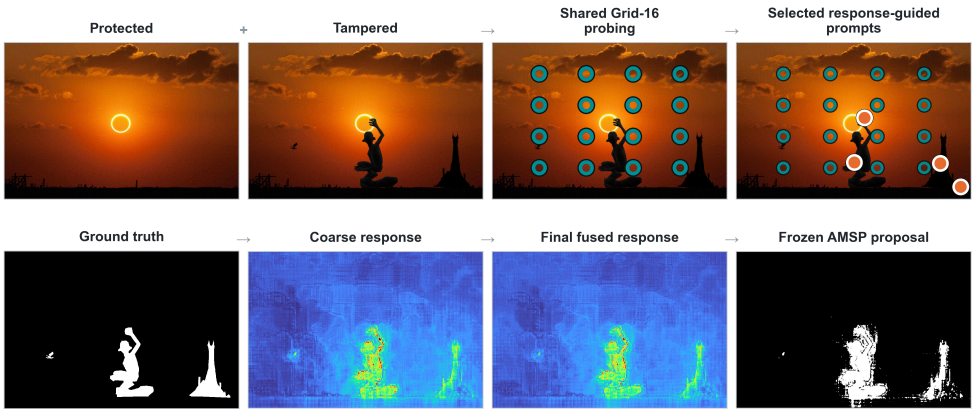


Figure 7: AMSP qualitative analysis. Top: protected/tampered pair, shared Grid-16 probing, and selected response-guided prompts. Bottom: GT, coarse response, fused response, and the frozen semantic proposal.

D.2 PRSR: Residual-Guided Structural Correction

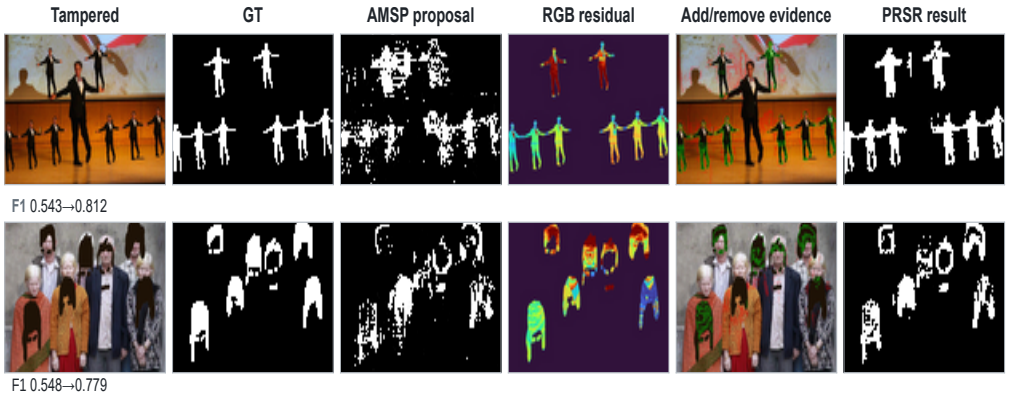


Figure 8: PRSR qualitative analysis. Columns show the tampered image, GT, AMSP proposal, paired RGB residual, add/remove evidence, and PRSR result. F1 improves from 0.543 to 0.812 in the first example and from 0.548 to 0.779 in the second.

D.3 SARD: Accept/Fallback Decisions Across Families

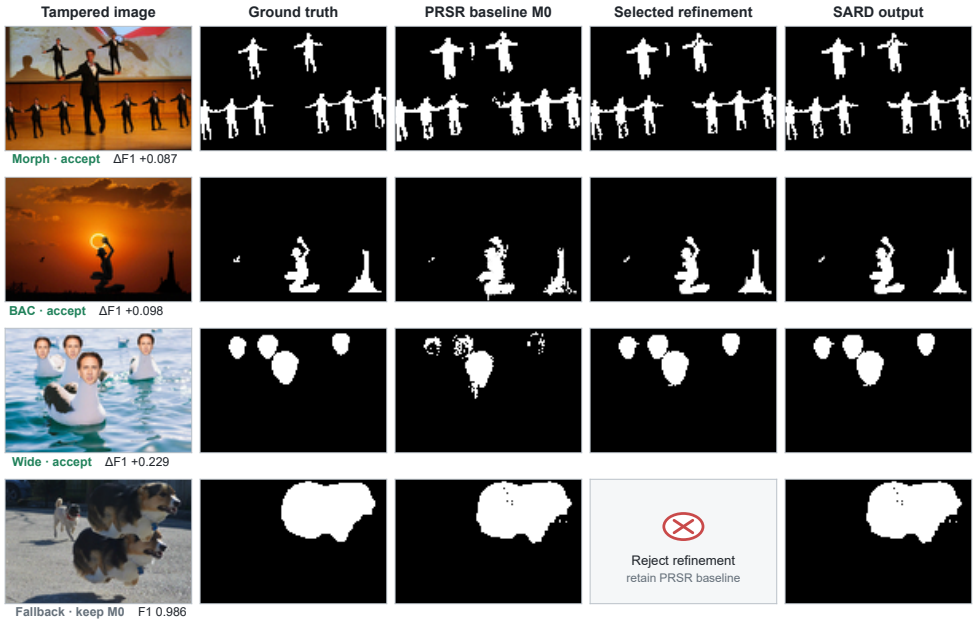


Figure 9: SARD decisions across Morph, BAC, and Wide, plus fallback. The accepted refinements yield F1 changes of +0.087, +0.098, and +0.229 relative to PRSR; the fallback case retains the PRSR baseline exactly (F1=0.986).

E FAILURE CASES AND LIMITATIONS



Figure 10: Representative held-out failure cases. Columns show the tampered image, GT, AMSP, PRSR, and final SPARR prediction.

Over-localization. The first case has high recall (0.969) but substantial background, reducing precision to 0.156 and F1 to 0.269.

Under-localization. The second remains highly precise (0.994) but covers only part of the manipulated region (recall 0.214; F1=0.352). F1 rises from AMSP (0.289) to PRSR (0.327) and SPARR (0.352), indicating incomplete semantic support rather than harmful refinement.

Refinement-risk miscalibration. In the third case, SARD accepts a Wide candidate ($w = 24$, $z_a = 4.5$, $z_r = 1.5$). Predicted utility/F1 gains are +0.056/+0.026, but F1 drops from 0.752 to 0.619 and BF1 from 0.371 to 0.181.

For display, masks whose native resolution differs from GT are nearest-neighbor aligned. Metrics are taken from the frozen held-out evaluation and not recomputed from aligned masks.

Overall, errors reflect insufficient/excessive semantic support and imperfect refinement-risk estimation. The 99 F1 degradations versus PRSR motivate uncertainty- or calibration-aware gating.