CAN YOU FIND THE EDIT WHEN NOTHING LOOKS WRONG? A PAIRED BENCHMARK FOR TEXT-DRIVEN IMAGE MANIPULATION LOCALIZATION BEYOND SEMANTIC ABNORMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

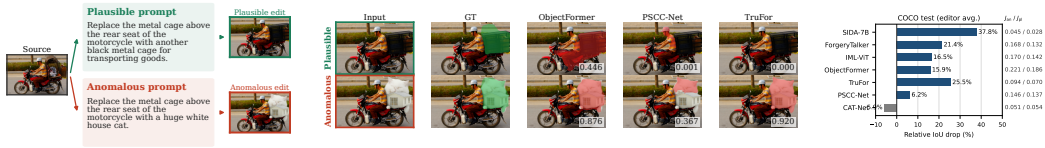
Instruction-driven image editing can produce manipulations that remain semantically coherent with the surrounding scene. Standard per-image evaluations do not reveal whether a localizer detects forensic evidence or relies on semantic abnormality that coincides with the edited region. We formalize this issue as the *semantic-plausibility shortcut*: a model may localize an anomalous replacement yet miss a plausible edit of the same target. We introduce Plausible–Anomalous Paired Edits (PAPE), a benchmark that makes intended semantic plausibility an explicit evaluation factor. Each triplet contains a source image and two edits of the same target, one plausible and one anomalous, produced with one editor and matched controlled editing settings, independently annotated masks, and overlap filtering for comparable spatial support. A two-stage construction pipeline yields ∼47K bilingual triplets from 6.5K photographs shared across four editors, three source domains, and six semantic violation categories. A pair-preserving protocol reports localization quality in each condition and the gap between them. Across seven public localizers, editor-averaged IoU is lower on plausible edits for every method on ImageNet and Places and for six of seven on COCO, with positive relative gaps from 2.5% to 42.2%. PAPE exposes a reliability dimension that aggregate scores conceal and provides paired records for analyzing its causes. Our anonymized code and dataset is available at `https://anonymous.4open.science/r/PAPEDataset/`.

1 INTRODUCTION

Instruction-driven image editing can alter what a photograph communicates while keeping the result coherent with the surrounding scene. Recent forensic resources increasingly reflect this shift toward diverse, realistic manipulations and diffusion-based edits (Qu et al., 2024; Wang et al., 2025; Zhuang et al., 2026). Such edits create a fundamental challenge for visual verification: a manipulated region may remain physically and semantically plausible, leaving no obvious reason to question its content. A reliable forensic localizer must therefore identify where an image was edited even when the edit does not look wrong.

A per-image localization score does not reveal whether a model found forensic evidence or followed a semantic cue. Figure 1(a) shows a matched PAPE pair: on a motorcycle photograph, the metal cage above the rear seat is replaced either with another black goods cage (plausible) or with a huge white house cat (anomalous), using the same source, target region, and editor. Both outputs are manipulated, but only the anomalous edit violates ordinary scene expectations. Figure 1(b) compares three representative localizers on this pair. ObjectFormer reaches IoU 0.876 on the anomalous edit but only 0.446 on the plausible edit, while PSCC-Net and TruFor collapse to 0.001 and 0.000 on the plausible output despite stronger overlap on the anomalous side. Figure 1(c) summarizes the corresponding average relative IoU drops on the COCO test split (Table 4), where six of seven methods remain positive and SIDA-7B reaches 37.8%. We refer to this dependence as the *semantic-plausibility shortcut*: semantic abnormality predicts the edited region in anomalous examples, but the cue weakens or disappears when the same target receives a plausible edit.

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.



(a) Controlled PAPE Pair (b) The Qualitative Results of Representative Localizers (c) The Quantitative Results of 7 Methods

Figure 1: (a) A matched plausible and anomalous edit pair from the same source, target, region, and editor. (b) Representative localization results on the anomalous and plausible edits. (c) Editor-averaged relative IoU drop on COCO test split across seven localizers.

Existing benchmarks cover diverse manipulation types, source content, and editing models, but aggregate scores mix plausible and anomalous edits. Strong performance on semantically conspicuous manipulations can consequently mask failures on coherent ones. Separating the two conditions after independently collecting images is insufficient because differences in sources, editors, targets, or edited regions can also produce a performance gap. Comparing plausibility sensitivity therefore calls for matched alternative edits that share the source, target, editor, and other editing configuration and retain comparable spatial support. This paired design makes it possible to compare localization performance as the intended semantic condition changes from anomalous to plausible under matched editing conditions.

To provide this comparison, we introduce **Plausible–Anomalous Paired Edits (PAPE)**. Each triplet contains a source image and two edits of the same target, one plausible and one anomalous. The two edits use the same editor and matched controlled editing settings, pass a spatial-overlap filter that retains comparable edited regions, and receive separate masks for their realized changes. A two-stage pipeline first generates and validates paired Chinese and English instructions and then constructs, annotates, and filters the edited outputs under the Six-Type Semantic Taxonomy of Object Anomalies (SSTOA). The resulting dataset contains ∼47K bilingual triplets (∼95K edited images) from 6.5K photographs shared across Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (Team et al., 2026), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), and Flux.2-dev (Black Forest Labs, 2025) on COCO (Lin et al., 2014), ImageNet (Russakovsky et al., 2015), and Places (Zhou et al., 2017). Separate manipulation and semantic labels ensure that both pair members remain manipulated examples, while a pair-preserving protocol reports localization quality for each condition and their relative gap.

Experiments with seven public localizers show why this distinction matters. Every evaluated method has lower editor-averaged IoU on plausible edits for ImageNet and Places, and six of seven exhibit the same pattern on COCO. The positive relative gaps range from 2.5% to 42.2% across domains and model families. These results establish the plausibility-associated gap as an observable evaluation outcome, while the paired records support further analysis of the forensic, geometric, and annotation factors that may contribute to it.

We make three contributions:

- We formulate a paired plausibility test for image manipulation localization and introduce a matched editing design that varies intended semantic plausibility while retaining output-specific masks and separate manipulation labels.
- We construct PAPE, a corpus of ∼47K bilingual triplets from 6.5K photographs spanning Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (Team et al., 2026), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), and Flux.2-dev (Black Forest Labs, 2025), three source domains, and six SSTOA categories.
- We provide a pair-preserving evaluation protocol and a diagnostic study of seven public localizers, revealing performance gaps that aggregate per-image scores obscure.

2 Related Work

Datasets for image manipulation localization. Manipulation-localization datasets traditionally pair each spliced or copy-moved image with one mask. Recent resources broaden the manipulation

Table 1: **Construction choices of recent image manipulation localization datasets.** Every listed resource provides pixel masks. Prior datasets evaluate independent images or other diagnostic axes and therefore cannot isolate a same-source plausibility contrast. PAPE makes that contrast the evaluation unit. In the last three columns, ✓ indicates that the resource satisfies the column criterion and × that it does not.

Resource	Forgery process	Masks	Matched pair	Paired plausibility
MIML (Qu et al., 2024)	Web / manual	✓	×	×
GIM (Chen et al., 2025)	Mask-guided inpaint	✓	×	×
OpenSDI (Wang et al., 2025)	Global / local diffusion	✓	×	×
LocateEdit (Wu et al., 2026)	Instruction edit	✓	×	×
SIGMA (Zhuang et al., 2026)	Annotates existing edits	✓	×	×
NeXT-IMDL (Li et al., 2025)	Diverse AIGC edits	✓	×	×
HSIM (Zhu et al., 2026)	Multi-step sequences	✓	×	×
PAPE	Instruction edit	✓	✓	✓

process while retaining this per-image evaluation unit. MIML (Qu et al., 2024) auto-annotates web-collected manual forgeries, whereas GIM (Chen et al., 2025), OpenSDI (Wang et al., 2025), and DEAL-300K (Zhang et al., 2025) extend mask supervision to generative and diffusion-based edits. Instruction-driven editors introduce a further challenge because they regenerate an image without a user-specified mask. SIGMA (Zhuang et al., 2026) derives localization labels from existing (source, instruction, output) records, while LocateEditBench (Wu et al., 2026) and PromptForge-350k (Wang et al., 2026a) construct pixel masks for instruction-based editing. Diagnostic suites expand the axes of evaluation without changing that unit. NeXT-IMDL (Li et al., 2025) organizes AIGC edits by editor, type, semantics, and granularity and tests unseen-condition generalization, yet each image is still scored independently: a semantic split does not yield a matched plausible and anomalous replacement of the same target under one editor and sampling configuration. HSIM (Zhu et al., 2026) annotates stepwise hierarchical trajectories and therefore isolates process order rather than the intended semantic plausibility of a single instruction-driven edit. LocateEditBench further filters toward coherent instruction edits, which can hide failures on ordinary replacements. Coverage, editor diversity, and process annotations therefore increase, but an aggregate or cross-split score cannot attribute a miss to the disappearance of semantic abnormality. PAPE changes the comparison unit by pairing plausible and anomalous edits under matched construction conditions (Table 1).

Forensic and semantic representations. Localization methods differ in the evidence they expose to the prediction head. CAT-Net (Kwon et al., 2021) uses compression information, PSCC-Net (Liu et al., 2022) models spatio-channel correlations, and TruFor (Guillaro et al., 2023) combines image and forensic clues. ObjectFormer (Wang et al., 2022) emphasizes object-level consistency, while IML-ViT (Ma et al., 2023) develops transformer-based localization. Foundation-model approaches include ForensicsSAM (Peng et al., 2025), the multimodal SIDA (Huang et al., 2025) and ForgeryTalker (Lian et al., 2026) systems, and FLAME (Wang et al., 2026b), which models local energy anomalies in AI-manipulated images. Recent architectures try to use both scales of evidence. FASA (Liang et al., 2026) injects CLIP semantic prototypes into a frequency pathway and still queries the decoder with a semantic prototype, so frequency modeling is combined with semantic reasoning rather than substituted for it. X-Edit (Bazyleva et al., 2025) localizes text-guided diffusion edits from inversion features, but each edited image is still scored alone. Architecture alone, however, does not establish whether a prediction follows low-level traces, semantic inconsistency, or both. Training data and deployment domain also affect the observed score. PAPE therefore measures plausibility sensitivity directly instead of assigning each model to a purely forensic or semantic category.

3 Defining the Paired Plausibility Test

The paired plausibility test asks whether localization quality changes when the same target is edited under two intended semantic conditions. One replacement is compatible with the scene, whereas the other is designed to violate a specified physical, biological, functional, spatial, interaction, or contextual expectation. The comparison is between a plausible edit and an anomalous edit, and the semantic condition is recorded separately from the manipulation label.

Table 2: **Six-Type Semantic Taxonomy of Object Anomalies (SSTOA): six sources of incompatibility.** Each row pairs a compatible replacement of the same target with a corresponding semantic violation.

ID	Locus	Category	Defining question	Contrast
C1	Object	Physical attribute	Does material or physical state violate ordinary object properties?	Rigid ceramic mug / mug made of unsupported liquid
C2	Object	Function or affordance	Does the object support its depicted use?	Functional watering can / watering can with a sealed spout being used to pour
C3	Object	Biological structure	Are anatomy, part count, or bodily configuration inconsistent?	Cat with ordinary anatomy / cat with an extra head
C4	Relation	Spatial relation	Are support, contact, containment, or scale inconsistent?	Book resting on a shelf / book intersecting the solid shelf
C5	Relation	Interaction relation	Is an action or relation between entities incoherent?	Hand grasping a cup handle / hand passing through the solid handle
C6	Scene	Scene context	Is an otherwise recognizable object incompatible with its surroundings?	Chair at a dining table / traffic barrier replacing the chair at a formal table

3.1 SSTOA: Structuring the Semantic Contrast

A useful plausibility contrast must specify which expectation the anomalous edit is meant to violate. A generic request to make an object "unusual" is underdetermined: it may change style, introduce an anatomical defect, or create a scene mismatch without stating which semantic relation the pair is intended to test.

SSTOA removes this ambiguity with six sources of incompatibility (Table 2). For a fixed source and target, the plausible instruction P_i^{P} specifies a scene-compatible replacement, whereas the anomalous instruction P_i^{A} specifies the sampled violation at the same target. Category c_i records the sampled row of the taxonomy for pair i. The six rows differ in *locus*: object-internal violations (C1–C3), relation-level violations (C4–C5), and scene-level violations (C6).

Annotated pair unit. Let I_i^{R} denote source photograph i. For semantic condition $z \in \{\mathrm{P}, \mathrm{A}\}$, editor $\mathcal{G}_{e_i}$ applies instruction P_i^z under shared controlled editing settings θ_i:

$$I_i^z = \mathcal{G}_{e_i}(I_i^{\mathrm{R}}, P_i^z; \theta_i). \tag{1}$$

The two instructions replace the same target, while θ_i fixes the editor-side editing configuration. The resulting pair is the annotated unit

$$u_i = (I_i^{\mathrm{R}}, I_i^{\mathrm{P}}, I_i^{\mathrm{A}}, M_i^{\mathrm{P}}, M_i^{\mathrm{A}}, P_i^{\mathrm{P}}, P_i^{\mathrm{A}}, c_i, e_i), \tag{2}$$

where M_i^z is the mask of the realized edit on output I_i^z, c_i is the SSTOA category sampled for pair i, and e_i identifies the editor. Both edited outputs receive manipulation label $y = 1$, with intended anomaly labels $a^{\mathrm{P}} = 0$ and $a^{\mathrm{A}} = 1$. The source I_i^{R} is labeled $y = 0$. The primary observable on u_i is the *plausibility-associated localization gap*: how much localization quality changes between its plausible and anomalous members.

4 PAPE: Design and Construction

This section constructs the paired unit u_i from Section 3. Figure 2 shows the two stages: Stage A writes and validates the prompt pair, and Stage B edits, annotates, and filters it. We then summarize the retained collection and illustrate one triplet per SSTOA category in Figure 5.

4.1 Stage A: Prompt Pair Generation

As shown in Figure 2, Stage A samples a category c_i from SSTOA and conditions on the source photograph I_i^{R}. The first two modules share the same core MLLM (Qwen3.5-9B (Qwen Team, 2026)) but use distinct structured templates (Appendix A). The perception module applies a fixed perception instruction to I_i^{R} together with the sampled c_i and records the target object, target region, and a short description (Appendix A.1). The prompt generation module reads I_i^{R}, the sampled c_i, and that perception record to produce bilingual replace-style P_i^{P} and P_i^{A} with semantic rationales and pair-control notes (Appendix A.2). The prompt validation module runs SPPV (Appendix A.3), a deterministic rule engine that accepts a pair only when complete grounding, edit-form consistency,

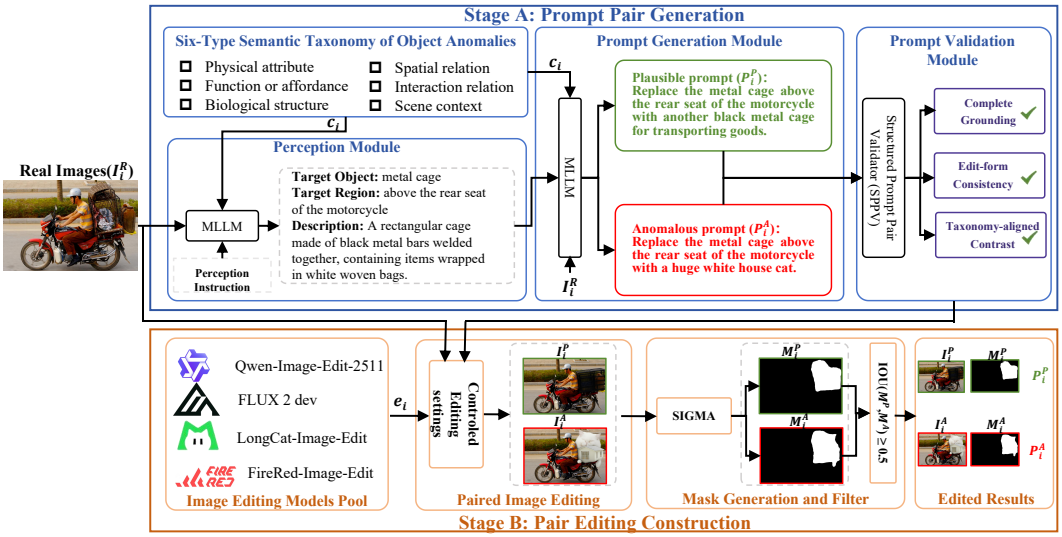


Figure 2: **Two-stage construction of a matched plausible and anomalous pair.** Stage A samples a SSTOA category for the source I_i^{R}, then runs perception, prompt generation, and SPPV to produce P_i^{P} and P_i^{A}. Stage B applies that pair under shared editing settings, annotates each output with SIGMA, and retains the pair when the two masks overlap.

Table 3: **PAPE common-source release.** Every source image is retained for all four editors after prompt validation, paired editing, mask-overlap filtering, and cross-editor alignment. Triplet counts include Chinese and English instruction variants.

Statistic	COCO	ImageNet	Places
Common source images	2,000	2,500	2,000
Triplets	14,668	18,488	14,591
Qwen-Image-Edit (Wu et al., 2025)	3,754	4,725	3,719
FireRed-Image-Edit (Team et al., 2026)	3,832	4,806	3,803
LongCat-Image-Edit (Meituan LongCat Team et al., 2025)	3,511	4,477	3,534
Flux.2-dev (Black Forest Labs, 2025)	3,571	4,480	3,535
Total	47,747 triplets from 6,500 sources / 4 editors		

and taxonomy-aligned contrast all hold, without calling a separate judging LLM to save the cost. Finally, accepted pairs are released in Chinese and English and the pairs that fail SPPV are rejected before Stage B.

4.2 Stage B: Pair Editing Construction

Stage B in Figure 2 materializes each validated prompt pair as two edited images with independent masks. *Paired image editing* draws one editor $\mathcal{G}_{e_i}$ from a pool of four models, Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (Team et al., 2026), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), and Flux.2-dev (Black Forest Labs, 2025), and applies P_i^{P} and P_i^{A} to the same source I_i^{R} under shared controlled editing settings θ_i, yielding outputs I_i^{P} and I_i^{A}. *Mask generation and filtering* invokes SIGMA (Zhuang et al., 2026) on each edited image with its instruction to estimate M_i^{P} and M_i^{A}. Empty or failed masks are discarded. A record is kept only when the two mask supports align:

$$q_i = \mathrm{IoU}(M_i^{\mathrm{P}}, M_i^{\mathrm{A}}) = \frac{|M_i^{\mathrm{P}} \cap M_i^{\mathrm{A}}|}{|M_i^{\mathrm{P}} \cup M_i^{\mathrm{A}}|} \geq 0.5. \tag{3}$$

The resulting release unit contains $(I_i^{\mathrm{P}}, M_i^{\mathrm{P}})$ and $(I_i^{\mathrm{A}}, M_i^{\mathrm{A}})$ for the same source, editor, and θ_i, so downstream localization compares plausible and anomalous conditions on comparable regions.

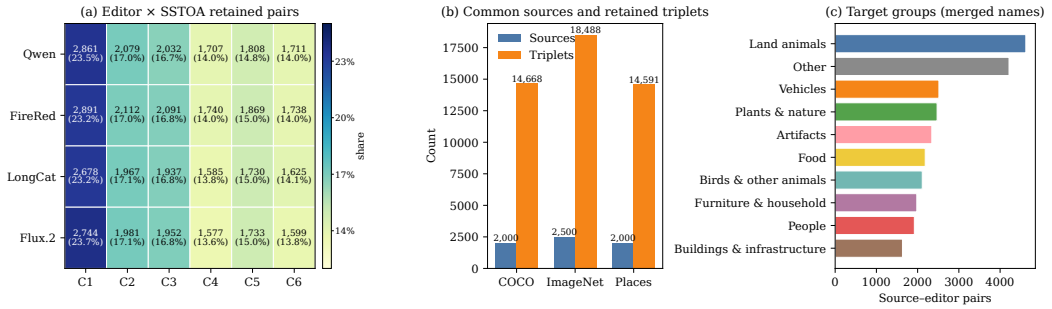


Figure 3: **Coverage of the PAPE release.** (a) SSTOA counts and within-editor percentages. (b) Distinct common sources and retained triplets by source domain. (c) Target-object coverage after merging 3,483 raw names into semantic groups (people, animals, vehicles, food, furniture, plants, buildings, artifacts, and unmatched other). Counts are unique source–editor pairs.

4.3 Dataset Composition and Traceability

Table 3 summarizes the common-source release: 47,747 bilingual triplets derived from 6,500 photographs that are represented for all four editors, Qwen-Image-Edit (Wu et al., 2025), FireRed-Image-Edit (Team et al., 2026), LongCat-Image-Edit (Meituan LongCat Team et al., 2025), and Flux.2-dev (Black Forest Labs, 2025). Each triplet contributes one plausible output, one anomalous output, and two independently estimated masks, for 95,494 edited images and 95,494 masks. Counts differ across editors because a source–editor record may retain the Chinese instruction, the English instruction, or both after overlap filtering. The 6,500 photographs therefore yield 26,000 source–editor records and 47,747 language-specific triplets. Appendix B.1 reports editor-by-category counts and language coverage. Traceability follows the same paired structure as evaluation. Stable pair and source-group identifiers link instructions, target metadata, outputs, masks, generation settings, and filtering decisions. The source-group identifier also keeps repeated generations and editor variants of one photograph together for source-aware analysis.

The retained category mixture remains stable across editors after prompt validation, editing, mask annotation, and overlap filtering. C1 accounts for 23.2%–23.7% of each editor's pairs and C4 for 13.6%–14.0%, with the same ordering of all six categories (Figure 3(a)). Category sampling precedes editor application, and the six-type shares remain aligned after overlap filtering. Domain counts reflect the common-source allocation: 2,000 COCO, 2,500 ImageNet, and 2,000 Places photographs yield 14,668, 18,488, and 14,591 triplets (Figure 3(b)). The 3,483 raw target names form a long tail. After synonym merging, land animals form the largest named group, and Other collects unmatched names (Figure 3(c)).

All 47,747 retained pairs satisfy $q_i \geq 0.5$, with mean overlap 0.770. Among them, 67.6% reach $q_i \geq 0.7$ and 20.1% reach $q_i \geq 0.9$ (Figure 4(a)). In a stratified sample of 1,920 pairs, median mask-area similarity is 0.87, while anomalous masks occupy a mean foreground fraction of 0.349 compared with 0.324 for plausible masks (Figure 4(b)–(c)).

4.4 Illustrative SSTOA Contrasts

Figure 5 shows how the construction pipeline realizes each SSTOA category on retained data. The display uses one English COCO triplet per category C1–C6 (Table 2). Each row shares a single source I_i^{R} and arranges three columns: the source photograph, the plausible edit I_i^{P} with overlay of its mask M_i^{P}, and the anomalous edit I_i^{A} with overlay of M_i^{A}. The plausible member replaces the same target with a scene-compatible alternative, whereas the anomalous member encodes the row-specific incompatibility named in the panel title (physical attribute, affordance, biological structure, spatial relation, interaction relation, or scene context). SIGMA (Zhuang et al., 2026) annotates each output separately, so the two masks trace the realized edit boundary on that image rather than a shared template.

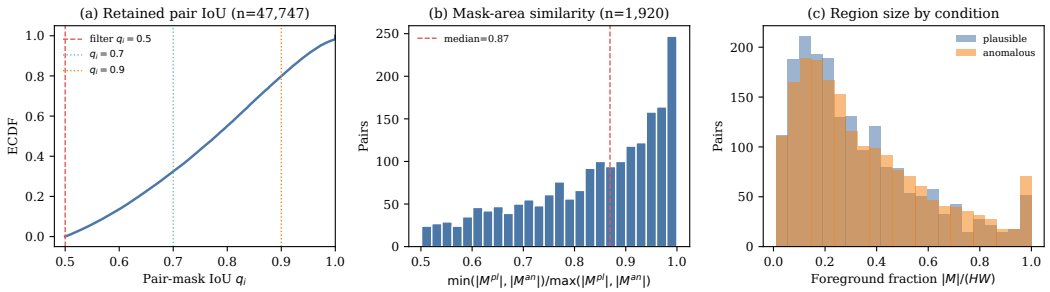


Figure 4: **Pair comparability after overlap filtering.** (a) Empirical cumulative distribution function (ECDF) of pair-mask IoU q_i over all 47,747 retained triplets (mean 0.770). Among them, 67.6% have $q_i \geq 0.7$ and 20.1% have $q_i \geq 0.9$. (b) Mask-area similarity and (c) foreground fraction for a stratified sample of 1,920 pairs (80 per editor–category stratum). Median area similarity is 0.87, and anomalous masks have mean foreground fraction 0.349 compared with 0.324 for plausible masks.

5 Benchmark Protocol

The benchmark scores each pair member separately and reports condition-specific means together with their gap. A localizer processes one edited image at a time and predicts a manipulation mask from that image alone. PAPE is evaluated as a full test-only collection.

Let B_i^z be the binary mask predicted by model f using a threshold fixed before evaluation and shared across both semantic conditions, and let $J_i^z = \mathrm{IoU}(B_i^z, M_i^z)$. For N pairs, the protocol reports

$$\bar{J}_{\mathrm{P}} = \frac{1}{N}\sum_{i=1}^{N} J_i^{\mathrm{P}}, \qquad \bar{J}_{\mathrm{A}} = \frac{1}{N}\sum_{i=1}^{N} J_i^{\mathrm{A}}, \tag{4}$$

$$\Delta_{\mathrm{rel}} = \frac{\bar{J}_{\mathrm{A}} - \bar{J}_{\mathrm{P}}}{\bar{J}_{\mathrm{A}}}. \tag{5}$$

A positive Δ_{rel} means that plausible edits receive lower mean IoU than their anomalous counterparts. The ratio is undefined when $\bar{J}_{\mathrm{A}} = 0$.

6 Experiments

We use the paired protocol in Equation 5 to evaluate seven public localizers on the released PAPE collection. The analysis groups checkpoints by their published design. SIDA-7B (Huang et al., 2025) and ForgeryTalker (Lian et al., 2026) are multimodal MLLM forensic systems. ObjectFormer (Wang et al., 2022) and IML-ViT (Ma et al., 2023) are dedicated localizers built around object-level or transformer features. CAT-Net (Kwon et al., 2021), PSCC-Net (Liu et al., 2022), and TruFor (Guillaro et al., 2023) use compression traces, spatio-channel correlations, or complementary forensic clues. All evaluations use publicly released weights and a prediction threshold fixed (0.5) before evaluation and shared across conditions. Tables 4–6 report condition-specific IoU and the relative gap for every source domain and editor. All values in Tables 4–6 are reported to three decimals, and each Δ_{rel} entry is computed from the corresponding $\bar{J}_{\mathrm{P}}$ and $\bar{J}_{\mathrm{A}}$ using Equation 5. Avg columns average the four editor-specific values within each metric.

Three regularities in Tables 4–6 show how localization changes when the same target moves from an anomalous edit to a plausible one:

- **Plausible-edit deficit.** Editor-averaged Δ_{rel} is positive for every method on ImageNet and Places and for six of seven on COCO. Under matched source, target, editor, and controlled editing settings, localizers overlap the anomalous replacement more often than the plausible one. This is consistent with the *semantic-plausibility shortcut* from Section 3: many models localize a region whose content looks wrong in context, and that cue weakens when the replacement remains scene-coherent even though both images are edited.
- **Weak plausible overlap despite usable anomalous scores.** SIDA-7B shows the largest average relative drops (0.378–0.422) while $\bar{J}_{\mathrm{P}}$ stays near the floor ($\leq$ 0.028).

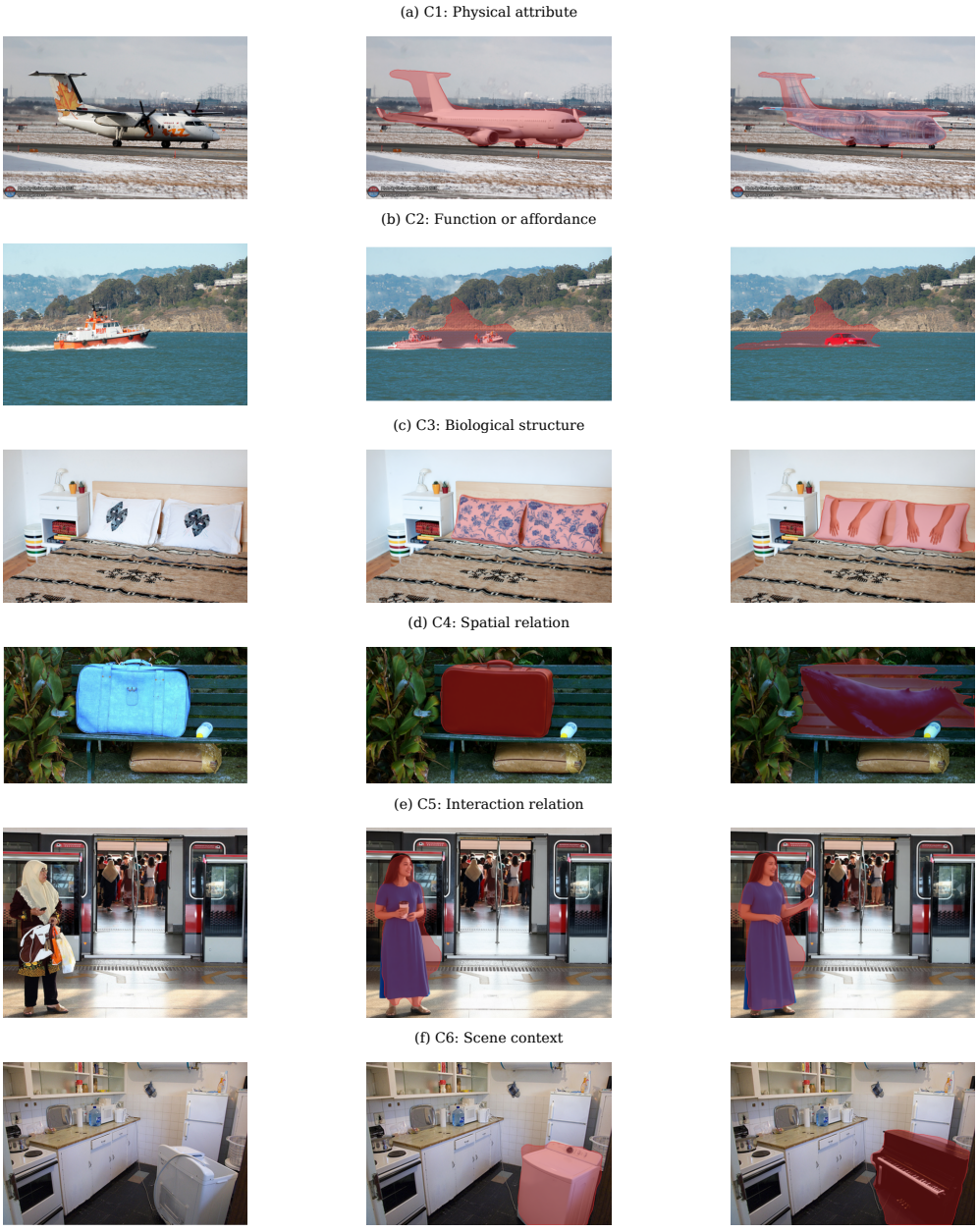


Figure 5: **SSTOA contrasts from the PAPE release.** Columns show the source, the plausible edit with its own mask overlay, and the anomalous edit with its own mask overlay.

ForgeryTalker and ObjectFormer reach moderate $\bar{J}_{\mathrm{P}}$ (0.132–0.194 and 0.145–0.186) yet still lose 0.108–0.275 of their anomalous-condition IoU on plausible edits. Strong $\bar{J}_{\mathrm{A}}$ therefore does not imply reliable plausible-edit localization. Multimodal systems may respond to conspicuous semantic mismatches, while object- or transformer-based localizers can track inconsistency on the anomalous side without transferring the same sensitivity once the edit looks ordinary.

- **Domain and editor effects dominate architecture.** On Places, every average Δ_{rel} exceeds the COCO and ImageNet values (0.239–0.422), plausibly because scene-heavy violations and cluttered backgrounds make anomalous edits easier to localize. PSCC-Net stays near zero on COCO and ImageNet (0.062, 0.025) but reaches 0.260 on Places, so compression-focused features still vary with domain difficulty. CAT-Net shifts from -0.059 on COCO to 0.333 on Places, where low $\bar{J}_{\mathrm{A}}$ and $\bar{J}_{\mathrm{P}}$ make the ratio unstable. In addition, on COCO,

Table 4: The pair localization performance on the official COCO test split.

	Qwen			FireRed			LongCat			Flux.2			Avg		
Method	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}
SIDA-7B	0.036	0.066	0.455	0.026	0.045	0.422	0.034	0.054	0.370	0.016	0.016	0.000	0.028	0.045	0.378
ForgeryTalker	0.119	0.167	0.287	0.126	0.164	0.232	0.199	0.224	0.112	0.084	0.118	0.288	0.132	0.168	0.214
IML-ViT	0.155	0.215	0.279	0.163	0.200	0.185	0.094	0.121	0.223	0.155	0.143	-0.084	0.142	0.170	0.165
ObjectFormer	0.207	0.255	0.188	0.190	0.223	0.148	0.185	0.210	0.119	0.160	0.196	0.184	0.186	0.221	0.159
PSCC-Net	0.144	0.150	0.040	0.157	0.161	0.025	0.092	0.119	0.227	0.156	0.153	-0.020	0.137	0.146	0.062
CAT-Net	0.072	0.060	-0.200	0.046	0.053	0.132	0.032	0.032	0.000	0.065	0.058	-0.121	0.054	0.051	-0.059
TruFor	0.101	0.144	0.299	0.101	0.143	0.294	0.027	0.034	0.206	0.052	0.057	0.088	0.070	0.094	0.255

Table 5: The pair localization performance on the official ImageNet test split.

	Qwen			FireRed			LongCat			Flux.2			Avg		
Method	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}
SIDA-7B	0.022	0.042	0.476	0.024	0.034	0.294	0.033	0.045	0.267	0.009	0.013	0.308	0.022	0.034	0.353
ForgeryTalker	0.195	0.252	0.226	0.190	0.241	0.212	0.272	0.301	0.096	0.120	0.159	0.245	0.194	0.238	0.185
IML-ViT	0.174	0.228	0.237	0.194	0.208	0.067	0.111	0.131	0.153	0.145	0.138	-0.051	0.156	0.176	0.114
ObjectFormer	0.203	0.228	0.110	0.170	0.192	0.115	0.175	0.199	0.121	0.146	0.161	0.093	0.174	0.195	0.108
PSCC-Net	0.163	0.175	0.069	0.193	0.181	-0.066	0.120	0.131	0.084	0.159	0.166	0.042	0.159	0.163	0.025
CAT-Net	0.072	0.079	0.089	0.065	0.069	0.058	0.045	0.048	0.063	0.056	0.066	0.152	0.059	0.066	0.106
TruFor	0.135	0.166	0.187	0.114	0.128	0.109	0.033	0.045	0.267	0.037	0.046	0.196	0.080	0.096	0.167

Table 6: The pair localization performance on the official Places test split.

	Qwen			FireRed			LongCat			Flux.2			Avg		
Method	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}	$\bar{J}_{\mathrm{P}}$	$\bar{J}_{\mathrm{A}}$	Δ_{rel}
SIDA-7B	0.033	0.067	0.507	0.032	0.051	0.373	0.027	0.043	0.372	0.012	0.019	0.368	0.026	0.045	0.422
ForgeryTalker	0.120	0.170	0.294	0.140	0.182	0.231	0.182	0.221	0.176	0.095	0.131	0.275	0.134	0.176	0.239
IML-ViT	0.169	0.273	0.381	0.155	0.187	0.171	0.079	0.116	0.319	0.117	0.139	0.158	0.130	0.179	0.274
ObjectFormer	0.161	0.231	0.303	0.132	0.185	0.286	0.151	0.192	0.214	0.134	0.191	0.298	0.145	0.200	0.275
PSCC-Net	0.117	0.196	0.403	0.130	0.164	0.207	0.072	0.100	0.280	0.125	0.140	0.107	0.111	0.150	0.260
CAT-Net	0.072	0.127	0.433	0.063	0.086	0.267	0.030	0.047	0.362	0.057	0.077	0.260	0.056	0.084	0.333
TruFor	0.217	0.322	0.326	0.140	0.183	0.235	0.027	0.053	0.491	0.037	0.067	0.448	0.105	0.156	0.327

Flux.2 (Black Forest Labs, 2025) can flip the sign of Δ_{rel} for IML-ViT and CAT-Net (-0.084 and -0.121 vs. positive on Qwen (Wu et al., 2025)) and yields lower $\bar{J}_{\mathrm{A}}$ than Qwen for most methods, so unequal generative traces across conditions can contribute to the gap alongside plausibility.

7 Conclusion

Instruction-driven image editing can change a photograph while keeping the result coherent with the surrounding scene, so a reliable localizer must find the edited region even when the edit does not look wrong. Standard per-image scores cannot reveal whether a model detects forensic evidence or follows semantic abnormality that coincides with the edited region. We formalize this failure mode as the *semantic-plausibility shortcut*, in which a model may localize an anomalous replacement of a target but miss a plausible edit of the same target. Aggregate benchmarks mix the two conditions, and strong scores on conspicuous manipulations can conceal failures on coherent ones. Collecting the two conditions separately is also insufficient, because differences in sources, editors, targets, or edited regions can themselves produce a performance gap.

We address this gap with a paired plausibility test that varies intended semantic plausibility while matching the source, target, editor, and spatial support. PAPE realizes this test with 47,747 bilingual triplets from 6,500 photographs, four editors, three source domains, and six semantic categories. Each triplet provides output-specific masks and separate manipulation and semantic labels, and a pair-preserving protocol reports localization quality in each condition and the gap between them. Across seven public localizers, editor-averaged IoU is lower on plausible edits for every method on ImageNet and Places and for six of seven on COCO, with positive relative gaps from 2.5% to 42.2%. PAPE therefore makes semantic plausibility an explicit evaluation factor and supplies paired records for analyzing why localization fails when nothing looks wrong.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing: to edit the manuscript for readability, terminology consistency, and grammar. We have reviewed all AI-assisted text. The authors verified every claim, citation, and number against the source records and take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Valentina Bazyleva, Nicolo Bonettini, and Gaurav Bharaj. X-edit: Detecting and localizing edits in images altered by text-guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2720–2730, 2025.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Yirui Chen, Xudong Huang, Quan Zhang, Wei Li, Mingjian Zhu, Qiangyu Yan, Simiao Li, Hanting Chen, Hailin Hu, Jie Yang, Wei Liu, and Jie Hu. GIM: A million-scale benchmark for generative image manipulation detection and localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2311–2319, 2025.

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*, pp. 20606–20615, 2023.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841, 2025.

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, 2021.

Yifei Li, Haoyuan He, Yu Zheng, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. NeXT-IMDL: Build benchmark for NeXT-generation image manipulation detection & localization, 2025.

Jingchun Lian, Lingyu Liu, Yaxiong Wang, Yujiao Wu, Lianwei Wu, Li Zhu, and Zhedong Zheng. Generating attribution reports for manipulated facial images: A dataset and baseline. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 30455–30473, 2026.

Xiaojie Liang, Zhimin Chen, Ziqi Sheng, and Wei Lu. Bridging the micro–macro gap: Frequency-aware semantic alignment for image manipulation localization. *arXiv preprint arXiv:2604.12341*, 2026.

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, 2014.

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.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Xia Du, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image technical report. *arXiv preprint arXiv.2512.07584*, 2025.

Rongxuan Peng, Shunquan Tan, Chenqi Kong, Anwei Luo, Alex C Kot, and Jiwu Huang. Forensic-ssam: Toward robust and unified image forgery detection and localization resisting to adversarial attack. *arXiv preprint arXiv:2508.07402*, 2025.

Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Jianpeng Wang, Haoyu Wang, Baoying Chen, Jishen Zeng, Yiming Qin, Yiqi Yang, and Zhongjie Ba. Promptforge-350k: A large-scale dataset and contrastive framework for prompt-based ai image forgery localization. *arXiv preprint arXiv:2603.29386*, 2026a.

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.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Opensdi: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4291–4301, 2025.

Yiming Wang, Baiqi Wu, Qingming Li, Jiahao Chen, Tong Zhang, and Shouling Ji. Order within chaos: Capturing intrinsic energy anomalies for AI-manipulated image forgery localization. In *Forty-third International Conference on Machine Learning*, 2026b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Shiyu Wu, Shuyan Li, Jing Li, Jing Liu, and Yequan Wang. Locateedit-bench: A benchmark for instruction-based editing localization. *arXiv preprint arXiv:2602.05577*, 2026.

Rui Zhang, Hongxia Wang, Hangqing Liu, Yang Zhou, and Qiang Zeng. Deal-300k: Diffusion-based editing area localization with a 300k-scale dataset and frequency-prompted baseline. *arXiv preprint arXiv:2511.23377*, 2025.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Xuekang Zhu, Ji-Zhe Zhou, Kaiwen Feng, Chenfan Qu, Xiwen Wang, Yunfei Wang, Liting Zhou, and Jian Liu. Revisiting image manipulation localization under realistic manipulation scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2026.

Peiyu Zhuang, Jianquan Yang, Haodong Li, Zhuoying Cai, Ruitao Xie, Jishen Zeng, Baoying Chen, Jiwu Huang, and Xiaochun Cao. Sigma: Semantic-difference instruction-grounding mask annotator for text-driven image manipulation localization. *arXiv preprint arXiv:2605.27924*, 2026.

A STAGE A PROMPT TEMPLATES AND VALIDATION

This appendix lists the English-facing Stage A prompt templates and the deterministic validation rules used in PAPE construction. Placeholders such as { objects_block } and {anomaly_block} are filled at runtime. The production pipeline returns strict JSON with paired Chinese (_zh) and English (_en) fields. The field names match the release schema.

A.1 PERCEPTION MODULE TEMPLATE

The perception module queries the source image with a fixed multimodal template to enumerate editable objects before pair generation.

System.

```
You are a rigorous assistant for constructing image-editing datasets for image manipulation
     localization research.
You must first identify visible objects in the image, and later generate paired edit
     instructions for the same object based on a replace operation.
Return strict JSON only. Do not output markdown, code fences, reasoning traces, or any text
     outside JSON.
The first character of your reply must be "{" and the last character must be "}".
```

User.

```
Carefully inspect the input image and list all clearly visible objects that can be edited
     independently.

Requirements:
- List only genuinely visible objects; do not invent off-screen or fully occluded objects.
- For each object, provide a short name, its spatial region, and a brief appearance
     description.
- Order objects from large to small, or from foreground to background.
- List at least 3 objects; if fewer are visible, list all of them.

Return JSON only, in the following format:
{
  "detected_objects": [
    {
      "object": "object",
      "region": "location of the object in the image",
      "description": "one-sentence description of the object appearance or state"
    }
  ]
}
```

A.2 PROMPT GENERATION MODULE TEMPLATE

Conditioned on the perception record, sampled category c_i, and source image, the prompt generation module selects one target and emits a bilingual plausible–anomalous replace pair with explicit semantic rationales and pair-control notes.

System.

```
You are a rigorous assistant for constructing image-editing datasets for image manipulation
     localization research.
Based on the already identified objects, generate both a normal (plausible) and an abnormal (
     anomalous) replace-style edit instruction for the same target object.
Fields ending with _zh must be written in Simplified Chinese; fields ending with _en must be
     accurate, natural English.
Return strict JSON only. Do not output markdown, code fences, reasoning traces, or any text
     outside JSON.
The first character of your reply must be "{" and the last character must be "}".
```

User (core).

```
Below is the list of objects detected in the image:
{objects_block}

Select exactly 1 object from the list that is most suitable for a local replace edit, and
     generate a bilingual pair of edit instructions.

Editing requirements:
1. Both the normal and abnormal instructions must be replace-style: Chinese must contain the
     release replace keyword, and English must use "Replace ...".
2. normal_prompt_*: after replacement, the result must be semantically plausible and
     consistent with common sense and physical laws.
3. abnormal_prompt_*: the replacement target must still be visually renderable, but must
     introduce a logical contradiction, physical implausibility, or common-sense violation.
4. Both instructions must edit the same object in the same region, and should preserve scale,
     viewpoint, lighting, and style as much as possible.
5. Edit only one target object; do not modify multiple unrelated objects or redraw the whole
     image.
6. Avoid violent, sexual, hateful, or gory content.
7. The English version must faithfully convey the Chinese version and must not introduce any
     new edit target absent from the Chinese version.
{anomaly_block}

Return JSON only, with the following fields:
{
  "target_object_zh": "...",
  "target_object_en": "...",
  "target_region_zh": "...",
  "target_region_en": "...",
  "target_object_description_zh": "...",
  "target_object_description_en": "...",
  "anomaly_type": {anomaly_type_field},
  "normal_prompt_zh": "...",
  "normal_prompt_en": "...",
  "abnormal_prompt_zh": "...",
  "abnormal_prompt_en": "...",
  "normal_semantic_rationale_zh": "...",
  "normal_semantic_rationale_en": "...",
  "abnormal_semantic_rationale_zh": "...",
  "abnormal_semantic_rationale_en": "...",
  "pair_control_notes_zh": "...",
  "pair_control_notes_en": "..."
}
```

When a SSTOA category is sampled, {anomaly_block} expands to the category id, bilingual name, definition, in-category examples, and guardrails. In auto mode it instead instructs the model to choose an anomaly type from the selected object and scene.

A.3 PROMPT VALIDATION MODULE (SPPV)

Rather than calling an additional LLM to judge prompt quality, Stage A inserts a lightweight *structured prompt pair validator* (SPPV) after JSON parsing. SPPV is a deterministic rule module aligned with Figure 2. It accepts a candidate pair only when **complete grounding**, **edit-form consistency**, and **taxonomy-aligned contrast** all hold. Failed candidates are logged and discarded.

Complete grounding. Every bilingual schema field must be present and non-empty: target object, target region, pre-edit description, `anomaly_type`, normal/abnormal prompts, semantic rationales, and pair-control notes (Chinese and English counterparts). This ties both edit instructions to the same explicit object–region record from the perception module.

Edit-form consistency. Both Chinese prompts contain the release replace keyword and both English prompts contain "replace" (case-insensitive), locking the pair to a shared replace-style edit operator so P_i^{P} and P_i^{A} differ in semantic plausibility rather than in edit operator.

Taxonomy-aligned contrast. When taxonomy mode is active, `anomaly_type` must equal the sampled c_i and both semantic rationales must be non-empty, so the anomalous prompt is intended to instantiate c_i while the plausible prompt remains scene-coherent on the same target.

```
SPPV(pair) -> ACCEPT iff all checks pass; otherwise REJECT

Complete grounding:
  all required bilingual fields present and non-empty

Edit-form consistency:
  release replace keyword in normal_prompt_zh and abnormal_prompt_zh
  "replace" in normal_prompt_en and abnormal_prompt_en (case-insensitive)

Taxonomy-aligned contrast:
  anomaly_type == sampled SSTOA category c_i  (taxonomy mode)
  normal_semantic_rationale_* and abnormal_semantic_rationale_* non-empty
```

B Construction and Release Specification

B.1 Release Composition

Chinese and English variants of one `sample_id` are counted as separate triplets: each variant produces its own edited outputs and masks. Table 7 traces the Stage B population from usable-mask generation through overlap filtering, cross-editor source alignment, and final source selection. Table 8 reports the retained triplets by source dataset and editor. Tables 9 and 10 report the SSTOA mixture by source dataset and by editor. Table 11 reports instruction-language coverage. Across 26,000 source–editor `sample_ids`, 21,747 (83.6%) retain both language variants, 1,829 retain only Chinese, and 2,424 retain only English.

Table 7: **Construction funnel recovered from generation and export logs.** Each numeric cell is a count. Percentages are relative to the preceding stage in the same column. The release is the common-source subset after $q_i \geq 0.5$ filtering, four-editor source alignment, and balanced source selection.

Stage or accounting item	Qwen	FireRed	LongCat	Flux.2
Generated pairs with two usable masks	29,614	29,616	29,616	29,600
Retained after $q_i \geq 0.5$	18,551 (62.6%)	20,611 (69.6%)	18,357 (62.0%)	16,784 (56.7%)
Common-source aligned (6,967 images)	13,071 (70.5%)	13,324 (64.6%)	12,358 (67.3%)	12,413 (74.0%)
Release after selecting 6,500 sources	12,198 (93.3%)	12,441 (93.4%)	11,522 (93.2%)	11,586 (93.3%)

Table 8: **Common-source triplets by dataset and editor.** Sources are unique photographs shared by all four editors.

Dataset	Sources	Qwen	FireRed	LongCat	Flux.2	Total pairs
COCO	2,000	3,754	3,832	3,511	3,571	14,668
ImageNet	2,500	4,725	4,806	4,477	4,480	18,488
Places	2,000	3,719	3,803	3,534	3,535	14,591
Total	6,500	12,198	12,441	11,522	11,586	47,747

Table 9: **SSTOA distribution by source dataset** (count; share within dataset).

Category	COCO	ImageNet	Places
C1 Physical attribute	3,566 (24.3%)	4,126 (22.3%)	3,482 (23.9%)
C2 Function / affordance	2,477 (16.9%)	3,150 (17.0%)	2,512 (17.2%)
C3 Biological structure	2,392 (16.3%)	3,216 (17.4%)	2,404 (16.5%)
C4 Spatial relation	1,965 (13.4%)	2,542 (13.7%)	2,102 (14.4%)
C5 Interaction relation	2,260 (15.4%)	2,836 (15.3%)	2,044 (14.0%)
C6 Scene context	2,008 (13.7%)	2,618 (14.2%)	2,047 (14.0%)
Total	14,668	18,488	14,591

Table 10: **SSTOA distribution by editor** (count; share within editor).

Category	Qwen	FireRed	LongCat	Flux.2
C1 Physical attribute	2,861 (23.5%)	2,891 (23.2%)	2,678 (23.2%)	2,744 (23.7%)
C2 Function / affordance	2,079 (17.0%)	2,112 (17.0%)	1,967 (17.1%)	1,981 (17.1%)
C3 Biological structure	2,032 (16.7%)	2,091 (16.8%)	1,937 (16.8%)	1,952 (16.8%)
C4 Spatial relation	1,707 (14.0%)	1,740 (14.0%)	1,585 (13.8%)	1,577 (13.6%)
C5 Interaction relation	1,808 (14.8%)	1,869 (15.0%)	1,730 (15.0%)	1,733 (15.0%)
C6 Scene context	1,711 (14.0%)	1,738 (14.0%)	1,625 (14.1%)	1,599 (13.8%)
Total	12,198	12,441	11,522	11,586

Table 11: **Instruction-language coverage.** A source–editor pair may retain Chinese, English, or both after overlap filtering, which is why editor totals are unequal.

	Chinese	English	Total triplets
COCO	7,229	7,439	14,668
ImageNet	9,181	9,307	18,488
Places	7,166	7,425	14,591
Total	23,576	24,171	47,747