
# Beyond Real or Fake: A Dual-Channel Authenticity and Reasoning Protocol for Photographic Assessment

Xiaoxiao Li Ruinan Jin Lili Meng
Miaosen Wang Athula Balachandran Pei Cao

Google YouTube DeepMind

**Abstract**

Forensic verification often uses a binary "real vs. fake" label that groups fully synthetic, tampered, and AI-retouched images despite their different consequences. We study these modifications through two complementary channels: a *camera channel* sensitive to capture and processing traces, and a *semantic channel* capturing scene content. The channels provide continuous evidence rather than deterministic signatures of manipulation history. We instantiate this perspective in **2CAP** (**2**-**C**hannel **A**uthenticity **P**rotocol), pairing a contrastively trained camera encoder with a frozen semantic encoder for **(i)** reference-free four-class classification through reliability-weighted cross-attention fusion and **(ii)** reference-based evidence generation. For the latter, query-reference channel similarities and a patch-level saliency map guide a frozen Vision Language Model (VLM) through an Observe–Generate–Refine loop, without forensic instruction tuning of the VLM. On the evaluated benchmark, 2CAP achieves overall AUC .963 and F1 .844; retouching F1 improves from .868 for the strongest compared baseline to .961. Across five VLM configurations, shared-parser paired evaluation shows model-dependent effects: evidence improves change-type accuracy over evidence-free refinement. These results support manipulation-type discrimination while delimiting the benefits of frozen-VLM refinement.

## 1 Introduction

When generative AI (GenAI) tools modify a photorealistic image, they do not all break it the same way. A portrait restyled by a lighting-enhancement model [1][2] may still depict the same person while altering the original capture traces; a news photograph with an inserted object [3] may retain camera-related cues in unedited regions while changing local semantic content; a fully synthesized scene [4] has no camera capture of the depicted scene, yet may remain semantically coherent. These scenarios can demand different responses, which a binary authenticity label alone does not express.

Cryptographic provenance (C2PA [5]) addresses authenticity at capture time but cannot reach legacy or stripped content. Computational forensics fills this gap, but the field has organized around binary classification, across both forgery localization [6][7][8] and AI-generated image detection [9][10][11][12][13], grouping AI-generated, edited, and retouched images under a single "fake" label regardless of how or why they were modified. This loses the structured evidence that journalism, legal proceedings, identity verification, and content moderation actually need: *what was changed, and does it matter under a given policy?* We argue that useful forensic systems must not only detect modifications, but characterize GenAI modification type by distinguishing (i) *retouching* (camera-parameter restyling intended to preserve scene content), (ii) *tampering* (localized semantic edits), and (iii) *full synthesis* (no camera capture).

We organize photographic assessment around two complementary evidence channels: a **camera channel** sensitive to capture and processing cues, and a **semantic channel** encoding scene content and relationships. We adopt camera-captured, GenAI-retouched, locally tampered, and fully synthetic images as four prototypical image-formation and editing regimes (Fig. 1(a)). They distinguish questions that binary real/fake detection conflates: whether an image

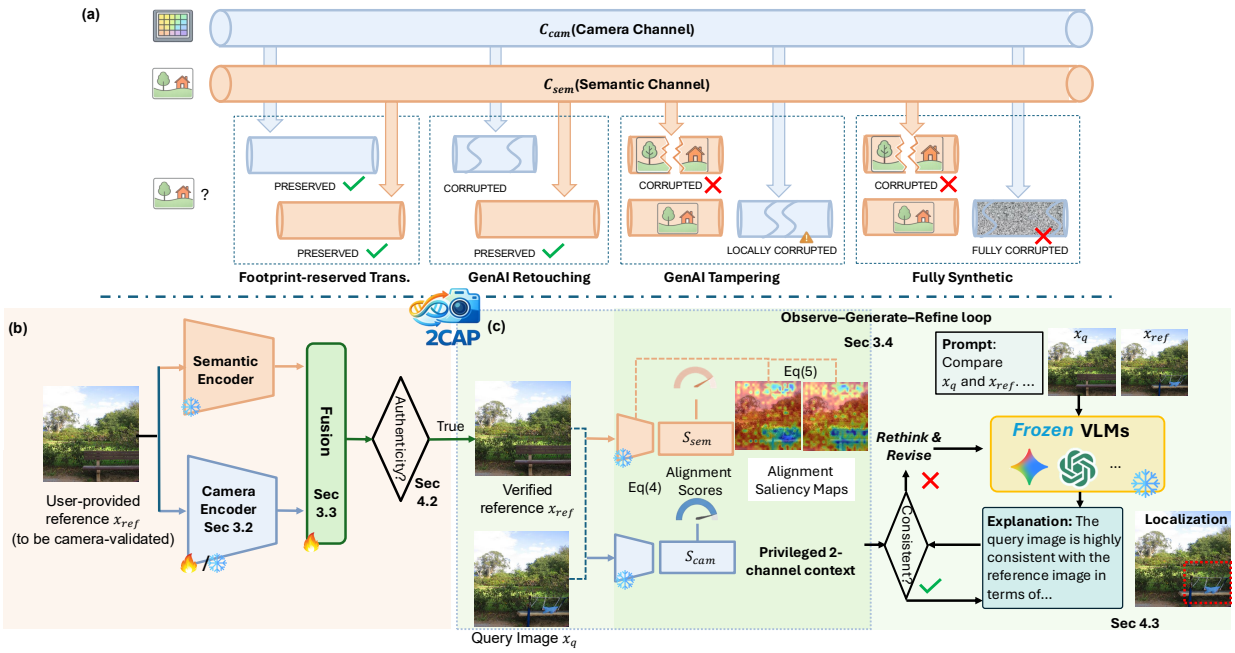


**Figure 1: 2CAP overview.** (a) *Prototypical two-channel evidence.* Camera-related processing traces and semantic content provide complementary cues; their responses can overlap across modification types. Benign processing may also alter camera traces, and synthesis need not introduce visible semantic inconsistencies. (b) *Reference-free authenticity classification* (Q1). The two encoders feed a reliability-weighted fusion module that predicts camera-captured, retouched, tampered, or fully synthetic. (c) *Reference-based evidence generation* (Q2). Query-reference alignment scores and a patch-level saliency map guide a frozen VLM through an Observe-Generate-Refine loop (Sec. 3.4).

originates from camera capture, whether its photographic appearance has been restyled, and whether its scene content has been modified or generated. These distinctions define an operational benchmark for evaluating the information contributed by the two channels across forms of image creation and editing.

Practical workflows also compose operations: local content editing followed by denoising and recompression can affect both semantic correspondence and camera-related traces. We therefore use the four regimes as reference cases for studying complementary evidence, rather than as an exhaustive partition of processing histories. Channel responses can overlap: benign processing can weaken capture cues, and synthesis can remain semantically coherent. Continuous query–reference alignment measurements describe such responses relative to a corresponding reference, while four-class prediction remains a dataset-defined task on joint image features.

We propose **2CAP** (**2**-**C**hannel **A**uthenticity **P**rotocol), which implements this evidence-based formulation with two encoders feeding two complementary applications. The shared backbone consists of a contrastively-trained camera encoder that maps images into a learned space of camera-related processing cues, paired with a frozen SigLIP2 [14] semantic encoder. **(1) Reference-free authenticity classification** (Fig. 1(b)): a cross-attention fusion module with learnable reliability weights adaptively combines the two channels per single image to produce a four-class verdict over camera, retouched, tampered, and fully synthetic. **(2) Reference-based evidence generation** (Fig. 1(c)): given a query image and a trusted corresponding reference, we treat per-channel alignment scores and a patch-level saliency map as *privileged context* that a frozen VLM cannot compute on its own, and feed them through a structured Observe-Generate-Refine (OGR) loop based on the iterative self-refinement paradigm of Madaan et al. [15] to obtain an explanation and a localized bounding box.

**Contributions. (1) A new problem formulation.** We move image forensics beyond binary real/fake classification by organizing assessment around camera-related and semantic evidence across four prototypical regimes. We evaluate four-class discrimination using the joint features and inspect pairwise channel scores for camera, retouched, and tampered images. **(2) A new detection capability.** We instantiate this formulation in 2CAP, a cross-attention fusion framework with learnable reliability weights that achieves the highest overall AUC and F1 among the compared methods on our benchmark. **(3) Forensic explanation without VLM fine-tuning.** We show that *frozen* VLMs can produce structured forensic evidence without any forensic SFT data: channel similarity scores and saliency maps, provided as privileged context through an OGR loop, yield model- and metric-dependent improvements over evidence-free

refinement; paired shared-parser evaluation also exposes classification degradation relative to a single pass.

## 2 Related Work

**Image forgery detection and localization.** Classical forensics exploit sensor fingerprints [16][17] and compression traces [18][19]. Learning-based methods fuse noise and RGB cues for end-to-end detection and localization [6][7][8][20]. In parallel, AI-generated image detection improves cross-generator generalization via pretrained representations [10][9][21][12][13][11][22]. These methods largely target binary detection or localization; our evaluation additionally distinguishes GenAI retouching as a separate category. Extended discussion is in Appendix B.

**Training data and representation controls.** B-Free [23] reduces content bias through aligned real/synthetic training data, while DRCT [24] uses diffusion-reconstructed hard examples with contrastive training. SSAFE [25] studies detection with frozen multimodal encoders. These methods motivate separating the effects of training data, representation choice, and fusion architecture when interpreting our results.

**VLM-based forensic reasoning.** Recent methods train VLMs for forensic detection, localization, and explanation, including FakeVLM [26], FakeShield [27], ForgeryGPT [28], and SIDA [29]. SIDA predicts tamper masks as well as explanations, while its 300K-image dataset should not be equated with 300K explanation-instruction pairs. More recent work explores expert-grounded reasoning through trained models, including REVEAL [30]. Our distinct evaluation concerns whether channel evidence improves a frozen VLM's reference-based assessment. We demonstrate this across five configurations from three model ecosystems, without forensic instruction tuning of the VLM; this does not imply that the encoders are training-free or that OGR outperforms fine-tuned forensic VLMs.

## 3 Method

### 3.1 Verification Protocol and Problem Formulation

We consider two practical forensic scenarios with complementary input requirements:

1. **Q1: Is a given image authentic?** Given a single image $\mathbf{x}$, predict its dataset-defined label: camera-captured, GenAI-retouched, locally tampered, or fully synthetic. This is a *reference-free* (single-image) task.
2. **Q2: What changed between two images?** Given a query $\mathbf{x}_q$ and a reference $\mathbf{x}_{\text{ref}}$ (e.g., a verified camera image), identify the nature, type, and location of modifications.

The reference-free classifier predicts a benchmark label from the joint encoder features; it does not infer an exhaustive processing history. For reference-based assessment, the continuous camera and semantic alignment scores $(s_{\text{cam}}, s_{\text{sem}})$ provide a two-dimensional description of evidence relative to a corresponding reference. This separates benchmark classification from the interpretation of combined processing effects: overlapping scores need not imply identical histories, and a composite workflow need not correspond uniquely to one of the four reference regimes.

**Two applications on a shared dual-channel foundation.** Both questions are answered by combining two complementary encoders: a camera encoder $\mathcal{E}_{\text{cam}}$ (Sec. 3.2) that learns camera-related processing cues, and a frozen semantic encoder $\mathcal{E}_{\text{sem}}$ (SigLIP2 [14]) that captures scene-level coherence. On top of these shared dual-channel features we build two applications, evaluated independently in Sec. 4. **(1) Authenticity classification** (Sec. 3.3; evaluated in Sec. 4.2): a cross-attention fusion module with learnable reliability weights adaptively combines the two channels per single input to produce a four-class verdict. **(2) VLM evidence generation** (Sec. 3.4; evaluated in Sec. 4.3): per-channel alignment scores between $\mathbf{x}_q$ and $\mathbf{x}_{\text{ref}}$, together with a patch-level saliency map, are supplied as privileged context to a frozen VLM through an Observe-Generate-Refine loop, producing interpretable evidence with localization.

### 3.2 Camera Fingerprint Encoder

**Limitation of existing pretrained encoders.** Pretrained semantic encoders provide useful forensic features [22][12], but their training objective does not explicitly target capture and processing traces. We therefore complement frozen semantic features with a camera-related encoder trained on camera/edit examples, and evaluate their contributions separately.

**Camera encoder architecture.** We adopt AIDE's DCT+SRM dual-branch backbone [11] and reprogram its training objective from cross-entropy classification to InfoNCE contrastive metric learning (Figure 2). The low-level feature extraction pipeline (DCT-based patch ranking and 30 SRM steganalysis filters) is inherited; our contribution is the training scheme, the cross-channel fusion, and the frozen-VLM evidence generation (a full architectural comparison is in Appendix C.3). Specifically, (i) we train with InfoNCE contrastive learning over camera-camera positives and paired or unpaired camera-edit negatives, producing a 128-d unit-norm embedding space rather than class logits; (ii) we use cosine similarity in this space as a continuous camera-channel score $s_{\text{cam}}$ rather than a binary decision; and (iii) we

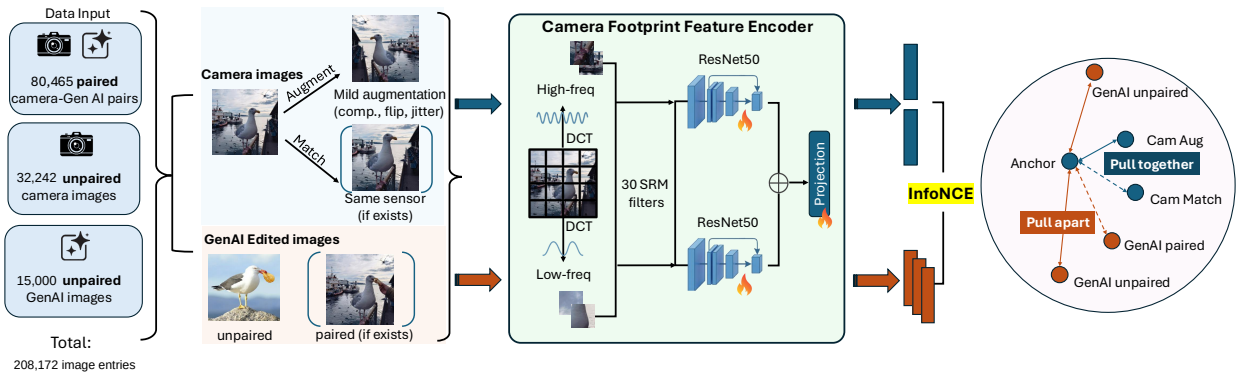


**Figure 2: Camera encoder pretraining.** *Left:* Training data composition: 80,465 pairs, 32,242 unpaired camera entries, and 15,000 unpaired edited entries, totaling 208,172 image entries (Table 4); unique images have not been deduplicated. *Center:* DCT-domain encoder with frequency-based patch selection and dual-branch feature extraction. *Right:* InfoNCE training where camera image embeddings are pulled together while edited image embeddings are pushed apart.

fuse camera features with features from a frozen semantic encoder (Sec. 3.3).

**Contrastive pretraining for generalizability.** A key design principle is that our camera encoder should learn *camera-specific traces*, not semantic shortcuts that fail to generalize across generators or editing tools. We encourage this through contrastive learning (Figure 2).

We construct two types of positives for each camera anchor: 1) *Camera Aug*: A mild augmentation of the same image (horizontal flip, mild color jitter, and mild JPEG compression). 2) *Camera Match*: Another image from the same camera sensor, when available. This encourages the encoder to learn sensor-consistent representations. The negative pool consists of all edited images in the batch, including 1) *GenAI paired*: The GenAI-edited version of the anchor's source image. This provides hard negatives that share scene content but lack authentic camera traces; and 2) *GenAI unpaired*: Other edited images in the batch, providing diverse negative examples across different editing styles and generators. The total loss combines paired and unpaired streams:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{paired}} + \lambda\, \mathcal{L}_{\text{unpaired}}, \tag{1}$$

where $\lambda = 0.5$. Each term is an InfoNCE loss with temperatures $\tau_p = 0.07$ (paired) and $\tau_u = 0.1$ (unpaired).

### 3.3 Cross-Channel Fusion for Authenticity Classification

The camera and semantic encoders provide features with different training objectives. We combine them using bidirectional cross-attention and input-dependent gating (Appendix C.2). The complementarity of the learned representations is evaluated through channel and fusion ablations, rather than assumed from a deterministic mapping between manipulation type and channel response.

**Cross-attention fusion.** We distinguish the token features used for classification from the global embeddings used for pairwise cosine scores. Let $Z_c \in \mathbb{R}^{B\times N_c\times d_c}$, for $c \in \{\text{cam}, \text{sem}\}$, denote the channel patch-token sequences, where $N_{\text{cam}}$ and $N_{\text{sem}}$ are the numbers of camera and semantic patches (the $p_1$ and $p_2$ axes in Fig. 8). Learned projections map them to $\tilde{Z}_c \in \mathbb{R}^{B\times N_c\times 256}$. For each attention head,

$$\text{Attn}(Q, K, V) = \text{softmax}_{\text{keys}}\left(QK^{\mathsf{T}}/\sqrt{d_h}\right) V, \tag{2}$$

with camera-to-semantic attention weights of shape $B \times N_{\text{cam}} \times N_{\text{sem}}$ and the reverse shape in the other direction. Residual cross-attention yields $Z'_{\text{cam}}$ and $Z'_{\text{sem}}$; channel-wise pooling produces the 256-d vectors $\mathbf{z}'_c$ used below. Attention is normalized over the key patches for each query patch, producing a patch-to-patch interaction matrix rather than a single attention coefficient.

**Reliability-weighted combination.** Not all channels are equally informative for every image. A two-layer gating MLP takes the concatenated channel vectors and produces two logits,

$$\boldsymbol{\ell} = g_{\text{gate}}([\mathbf{z}'_{\text{cam}}; \mathbf{z}'_{\text{sem}}]) \in \mathbb{R}^2, \qquad \alpha_c = \frac{\exp(\ell_c)}{\exp(\ell_{\text{cam}}) + \exp(\ell_{\text{sem}})}, \quad c \in \{\text{cam}, \text{sem}\}.$$

**Algorithm 1** OGR Loop: Observe-Generate-Refine

**Require:** Image pair $(\mathbf{x}_q, \mathbf{x}_{\text{ref}})$, scores $(s_{\text{cam}}, s_{\text{sem}})$, saliency $\mathcal{M}_{\text{sem}}$
1: **Observe:** VLM receives $(\mathbf{x}_q, \mathbf{x}_{\text{ref}})$ and performs visual inventory
2: **Generate:** VLM outputs initial assessment: change type, description, bounding box
3: **Refine:** VLM receives Pass-1 output with $(s_{\text{cam}}, s_{\text{sem}}, \mathcal{M}_{\text{sem}})$; revises if evidence conflicts
4: **return** Structured evidence: $\{\hat{y}, s_{\text{cam}}, s_{\text{sem}}, \mathcal{M}_{\text{sem}}, \text{explanation}\}$

Here, softmax is applied over the two channels for each image, so $\alpha_{\text{cam}} + \alpha_{\text{sem}} = 1$. This normalization is separate from the key-patch softmax inside cross-attention. We fuse the channel vectors using these input-dependent weights:

$$\mathbf{z}_{\text{fused}} = \alpha_{\text{cam}} \cdot \mathbf{z}'_{\text{cam}} + \alpha_{\text{sem}} \cdot \mathbf{z}'_{\text{sem}}, \quad \mathbf{p} = \text{softmax}(h_{\text{cls}}(\mathbf{z}_{\text{fused}})) \tag{3}$$

The classifier $h_{\text{cls}}$ has dimensions $256 \to 128 \to 4$ and is trained with four-class cross-entropy; its output categories are camera-captured, retouched, tampered, and fully synthetic.

### 3.4 VLM Evidence Synthesis with Privileged-Context Feedback

The dual-channel classifier (Sections 3.2-3.3) answers Question 1 (camera authenticity). To answer Question 2 (what changed) with interpretable evidence, we employ a structured Observe-Generate-Refine (OGR) loop with frozen VLMs, following the iterative self-refinement paradigm [15] and taking the *image pair* as input for direct comparison.

**Channel alignment scores and saliency.** We compute alignment scores via cosine similarity $\text{sim}(\cdot, \cdot)$:

$$s_{\text{cam}} = \text{sim}(\mathcal{E}^g_{\text{cam}}(\mathbf{x}_q),\, \mathcal{E}^g_{\text{cam}}(\mathbf{x}_{\text{ref}}))\,, \quad s_{\text{sem}} = \text{sim}(\mathcal{E}^g_{\text{sem}}(\mathbf{x}_q),\, \mathcal{E}^g_{\text{sem}}(\mathbf{x}_{\text{ref}})) \tag{4}$$

where $\mathcal{E}^g_c$ denotes the global image embedding of channel $c$, distinct from its patch-token output $Z_c = \mathcal{E}^{\text{tok}}_c(\mathbf{x})$ used for cross-attention. The camera global embedding is 128-d and unit-norm. These are learned feature similarities, not probabilities of authenticity or direct measurements of retained sensor information. Low $s_{\text{cam}}$ can reflect processing beyond GenAI, while high $s_{\text{sem}}$ indicates similar global content without excluding small local edits. For spatial localization, we compute a saliency map using patch-level features $\{p_i(x)\}$ from SigLIP2 [14]:

$$M_i = 1 - \max_j \text{sim}(p_i(\mathbf{x}_q),\, p_j(\mathbf{x}_{\text{ref}}))\,, \tag{5}$$

where high $M_i$ indicates query patches lacking close matches in the reference. The resulting map $\mathcal{M}_{\text{sem}} = \{M_i\}$ highlights query patches with weak reference matches. This directed map can miss removals or repeated-object changes when remaining query patches still find close matches.

**Motivation.** A frozen VLM comparing two images tends to describe visually salient differences first (lighting, color) and may miss two important cases: (i) retouching that changes camera traces without visible content differences, and (ii) localized edits in low-saliency regions. Channel scores quantify evidence the VLM cannot directly compute (e.g., sensor-trace similarity), so providing them as feedback after an initial assessment lets the VLM correct both failure modes.

**OGR protocol.** Unlike prior methods that train VLMs for forensic tasks [26][29][28], we keep the VLM frozen and supply pre-computed signals ($s_{\text{cam}}$, $s_{\text{sem}}$, $\mathcal{M}_{\text{sem}}$) through a two-pass Observe-Generate-Refine loop (Algorithm 1). *Pass 1 (Generate):* the VLM receives the image pair and outputs an initial assessment including change type (addition/removal/replacement), description, and bounding box. *Pass 2 (Refine):* the VLM receives its own Pass-1 output together with the channel scores and saliency map, and revises if the evidence conflicts with its initial assessment. The VLM is not forced to agree with the scores; it performs independent reasoning and can confirm, refine, or override. Revisions can correct prior errors or introduce new ones; accuracy contrasts alone do not identify the two transition counts. We use a single prompt template across the tested VLM configurations (Appendix C); portability across proprietary and open-weights VLMs (Table 2) shows that the evidence contribution depends on both model and metric.

## 4 Experiments

Our experiments evaluate the two applications of 2CAP's dual-channel features, mirroring the questions posed in Sec. 1. **(i) Dual-channel classifier** (Sec. 4.2): a reference-free four-class classifier with cross-attention fusion and learnable reliability weights, addressing Q1 (*is the image authentic?*). **(ii) OGR feedback loop** (Sec. 4.3): a privileged-context protocol that supplies similarity scores and saliency map to frozen VLMs for evidence generation, addressing Q2 (*what changed?*).

### 4.1 Dataset Construction and Implementation Details

We construct a multi-source benchmark for distinguishing *camera-captured* images from *edited* images across three modification categories (see Appendix A.3 for full details):

- **Fully synthetic**: Images generated entirely by GenAI from text prompts, lacking camera fingerprints, including Defactify Image (MS COCOAI) [31] (SD2.1 [1], SDXL [2], SD3 [32], DALL-E 3 [33], Midjourney 6 [1]) and the Flickr8K-derived subset of RU-AI [34] (SDv1.5 [1], SDAbsoluteReality [2], SDepiCRealism [3], SDXL [2], SDv6.0 [35]).
- **Tampered**: Localized GenAI editing (object insertion/removal/inpainting) that corrupts semantic content in specific regions. Sources: SID-Set [29], SAGI-D [36], and MagicBrush [37].
- **GenAI-retouched**: Generative Photography [38] dataset with global camera-parameter restyling (bokeh, color temperature, focal length, motion blur) using AnimateDiff [39] intended to preserve scene content.

Camera images are aggregated from Dresden [40], VISION [41], Defactify Image [31], and SID-Set [29]. This multi-source design broadens coverage, but unequal source and tool distributions across classes can introduce confounding. Shared sources across some classes do not eliminate that risk.

**Implementation Details.** Our architecture comprises a camera encoder $\mathcal{E}_{\text{cam}}$ (DCT-domain dual-branch ResNet-50 [42][11] with SRM filters, outputting 128-d embeddings), a frozen semantic encoder $\mathcal{E}_{\text{sem}}$ (SigLIP2 [14], 768-d), and a cross-attention fusion module with learnable reliability weights. The pretraining on $\mathcal{E}_{\text{cam}}$ is on 208K images via InfoNCE contrastive learning (The details of the training data are in Table 4).

### 4.2 Camera Authenticity: Classification on Camera vs. Fully Synthetic vs. Tampered vs. GenAI-retouched Images

This section evaluates the *first* application: the dual-channel classifier with cross-attention fusion and learnable reliability weights (Sec. 3.3), trained to assign each image to one of four classes: *camera-captured*, *fully synthetic*, *tampered*, or *GenAI-retouched*, the categories corresponding to distinct forensic scenarios requiring different responses.

#### 4.2.1 Baselines and Setup

We compare against eight methods spanning three categories: (i) *Training-free*: TruFor [8]. (ii) *Pretrained feature*: D3 [22], CNNSpot [10], AIDE [11]. (iii) *End-to-end*: VIBAIGCD [13], FreqNet [21], FatFormer [12], UnivFD [9].

**Training protocol.** The adapted baseline rows use the historical 12,800-image downstream training pool. For the controlled 2CAP ablation, we exclude 468 training rows whose source-reference groups overlap the historical test split, deduplicate four repeated training paths, and then partition the remainder into 9,877 training and 2,451 validation images by source-reference group. All six 2CAP variants use these same examples, frozen encoder embeddings, minibatch order, optimizer settings, and 20-epoch budget; validation macro-F1 selects each checkpoint before one evaluation on the common 3,200-image test split. The full-model row uses a separate rerun in which all weights shared with the no-gate control start identically and the gate starts at equal weights; the no-gate control is the previously completed run under the same seed and recipe. Thus the method and baseline rows share the test population but *not* the exact downstream training population or adaptation budget. 2CAP also uses separate task-specific contrastive pretraining, so Table 1 should not be read as a fully matched baseline comparison. The historical test split was inspected during prior development and is not a pristine blind holdout.

**Evaluation protocol.** VISION and the Flickr8K-derived RU-AI subset are held out from both of our training stages. This tests dataset and camera-domain transfer. Dataset holdout does not itself establish generator-family holdout: for example, SDXL appears in both the listed training-source and held-out-source generator sets. We distinguish these settings from the stricter evaluation in which an entire generator or editing-tool family is excluded from all training and model selection.

#### 4.2.2 Results

**Quantitative classification results.** A key contribution of 2CAP is distinguishing *among* modification types. All evaluations in this section are *reference-free* (single-image). Table 1 shows that the paired-initialization 2CAP run

[1] https://www.midjourney.com/home
[2] https://huggingface.co/Lykon/AbsoluteReality
[3] https://huggingface.co/emilianJR/epiCRealism

**Table 1: Four-class authenticity classification.** Overall and per-class performance on the common 3,200-image test set, plus held-out datasets (*). Per-class metrics report one-vs-rest AUC (OvR) and F1. Best in **bold**, second underlined (ranking computed over baselines and 2CAP only). The bottom block uses a common group-filtered training/validation protocol for all six variants and is excluded from the ranking; the full row uses paired, neutral gate initialization and baseline training populations differ.

| Method | Overall | | | | Camera | | Tampered | | Synthetic | | Retouched | | Flickr8k* | VISION* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | AUC | OvR | F1 | OvR | F1 | OvR | F1 | OvR | F1 | AUC | Brier ↓ |
| TruFor [8] | .100 | .254 | .123 | .559 | .618 | .410 | .724 | .000 | .497 | .000 | .396 | .082 | .637 | .123 |
| D3 [22] | .755 | .754 | .750 | .928 | .864 | .591 | .899 | .680 | .973 | .863 | .975 | .868 | .671 | .179 |
| CNNSpot [10] | .370 | .361 | .359 | .631 | .601 | .349 | .565 | .316 | .685 | .430 | .672 | .344 | .583 | .237 |
| AIDE [11] | .767 | .770 | .767 | .934 | .872 | .638 | .946 | .777 | .963 | .835 | .952 | .816 | .870 | .228 |
| VIBAIGCD [13] | .752 | .740 | .730 | .925 | .863 | .523 | .890 | .673 | .975 | .871 | .974 | .851 | .879 | .253 |
| FreqNet [21] | .530 | .518 | .484 | .768 | .663 | .258 | .791 | .569 | .886 | .688 | .731 | .422 | .840 | .329 |
| FatFormer [12] | .539 | .547 | .537 | .783 | .695 | .422 | .727 | .452 | .851 | .621 | .858 | .653 | .938 | .159 |
| UnivFD [9] | .380 | .323 | .269 | .615 | .576 | .415 | .559 | .061 | .600 | .252 | .727 | .347 | .818 | .165 |
| SAFE [43] | .591 | .594 | .583 | .838 | .776 | .464 | .891 | .669 | .888 | .675 | .797 | .523 | .937 | .136 |
| GAPL [44] | .455 | .445 | .387 | .695 | .796 | .548 | .637 | .396 | .727 | .524 | .620 | .081 | **.996** | **.095** |
| **2CAP** | **.908** | **.907** | **.906** | **.982** | **.957** | **.836** | **.987** | **.899** | **.987** | **.926** | **.998** | **.964** | .859 | .394 |
| *Ablation (ours): channels* | | | | | | | | | | | | | | |
| Semantic only ($\mathcal{E}_{\text{sem}}$) | .779 | .784 | .781 | .938 | .875 | .628 | .915 | .707 | .973 | .874 | .990 | .914 | .850 | .634 |
| Camera only ($\mathcal{E}_{\text{cam}}$) | .782 | .776 | .775 | .939 | .905 | .706 | .960 | .811 | .960 | .815 | .933 | .767 | .944 | .270 |
| *Ablation (ours): fusion strategy* | | | | | | | | | | | | | | |
| Concat + MLP | .890 | .890 | .890 | .977 | .947 | .821 | .982 | .886 | .985 | .914 | .994 | .938 | .914 | .464 |
| Addition + MLP | .888 | .888 | .888 | .978 | .948 | .818 | .983 | .881 | .986 | .915 | .995 | .939 | .928 | .390 |
| Cross-Attn (no rel. wts) | .908 | .908 | .908 | .983 | .958 | .834 | .987 | .899 | .989 | .931 | .998 | .968 | .914 | .281 |

achieves overall AUC .982 and F1 .906 on the common four-class test split. These exceed the reported baseline values, subject to the downstream-training differences described above. For retouching on GP, 2CAP reaches OvR AUC .998 and F1 .964, compared with F1 .868 for D3, the strongest compared baseline. The semantic-only variant reaches retouching F1 .914; the full model is 5.0 points higher and exceeds semantic-only overall F1 by 12.5 points. These differences support combining both channels on this test population, but do not isolate a benefit from reliability gating. Generalization is setting-dependent: Flickr8K AUC is .859 for the full model, below FatFormer (.938) and GAPL (.996).

**Ablation.** Table 1 (bottom) compares the six variants under the common training and selection protocol. *Channels.* Semantic-only obtains AUC .938 / F1 .781 and camera-only .939 / .775, below the full model (.982 / .906). *Fusion strategy.* Concat-MLP reaches .977 / .890 and addition-MLP .978 / .888. Cross-attention with fixed equal channel weights reaches .983 / .908, slightly *above* the reliability-gated model (.982 / .906). The full and no-gate variants have matched initial shared weights; in this single-seed comparison, the learned gate has no demonstrated performance benefit. Their test-set F1 difference (full minus no-gate) is $-0.0017$, with a source-group bootstrap 95% interval of $[-0.0103, 0.0065]$. The full model also trails several variants on Flickr8K and VISION, so the four-class gain does not imply uniform transfer.

**Learned reliability weights.** Fig. 3 (left) shows category-dependent gate distributions: the camera-side representation has higher weight for tampered images, whereas the semantic-side representation is favored for retouching and synthesis. These are descriptive diagnostics, not calibrated uncertainty or causal feature attributions. Each representation has already received information from the other channel through cross-attention; gate magnitudes alone therefore do not establish which original forensic cue caused a prediction.

### 4.2.3 Source and Processing Diagnostics

Within-source evaluation partially addresses whether performance reflects the dataset name alone. On SID-Set, camera-versus-tampered AUC is 95.5%; on Defactify, camera-versus-synthetic accuracy is 90.4% (Appendix D.2). These results show discrimination within the respective collections, but do not control all underlying image sources, class

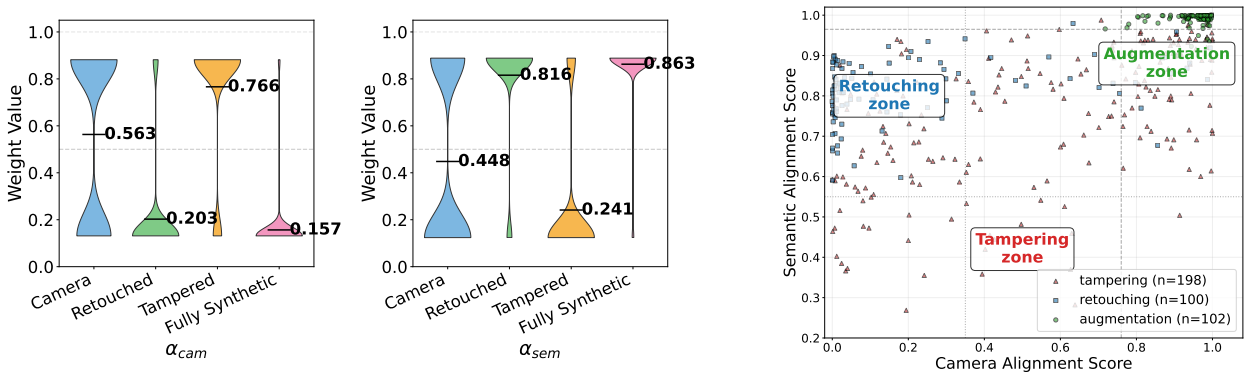


**Figure 3: Left: learned reliability weights** $\alpha_{\text{cam}}$ and $\alpha_{\text{sem}}$ averaged per category, showing clear channel specialization. **Right: score distribution.** Retouching: low $s_{\text{cam}}$, high $s_{\text{sem}}$; camera: high/high; tampering: higher variance.

balance, or editing and export pipelines. They also do not test retouching transfer beyond GP.

Two additional diagnostics examine sensitivity to processing. OpenCV examples are evaluation-only and mapped exclusively to the GP evaluation subset. Classical OpenCV versions of four photographic effects receive the retouched prediction in 81.2-92.7% of cases (Appendix D.3). This measures response to non-neural processing, not successful identification of an unseen GenAI editor. In a separate mild-processing diagnostic, the retouched prediction rate remains 1.5-2.0%, while camera recall falls from 62.5% to at least 58.5% (Appendix D.4). Thus, the small change in retouching false alarms coexists with substantial camera-class errors. The two diagnostics use different processing conditions and should not be pooled into a single robustness claim.

### 4.3 Evaluation on VLM Evidence Generation

This section evaluates the second application: reference-based assessment of what changed. Qualitative examples are in Fig. 9 and Appendix C.5.

**Channel score space as VLM context.** Fig. 3 (right) shows raw query-reference cosine similarities computed without the classification fusion module. Retouching examples cluster at low camera similarity and high semantic similarity; camera-derived pairs tend to score highly on both axes, while tampered examples show greater variation. These observed distributions motivate evidence conditioning for the evaluated pairs, but do not establish universal class boundaries. Pairwise scores and post-attention classification gates serve different tasks and are not equivalent measurements.

We evaluate whether dual-channel scores and saliency maps improve the outputs of the same frozen VLM on query-reference pairs. The comparison isolates the utility of evidence conditioning within the tested VLMs; it does not establish superiority to forensic VLMs trained for single-image detection and localization, such as FakeVLM, SIDA, and ForgeryGPT [26][29][28]. Such methods remain relevant practical comparators under explicitly documented input and supervision differences.

**Setup.** We evaluate 400 query–reference pairs: 102 benign camera-derived augmentations, 100 retouched, and 198 tampered. B is a single VLM pass given the image pair. N refines the same first-pass output without channel scores or heatmap; F refines it with both scores and the heatmap. All five models have B/N/F results: GPT-4o and GPT-4o-mini [45], Gemini 2.5 Flash and Pro [46], and Qwen3-VL-32B [47]. We average each pair's metric over three seeds and apply the same `vlm_only` change-type reading across conditions, without an F-only numerical-score override.

**Metrics and uncertainty.** Change-type accuracy uses the three benchmark labels. Object Recall and StrictHit measure noun coverage and action-compatible hits on 198 tampered pairs. Localization reports mean LocIoU and ACC@0.25 separately on 180 pairs with text-derived proxy locations; these are coarse spatial-agreement metrics, not pixel-level localization. Mean SalIoU is reported separately on its 197-pair subset. Paired 95% intervals resample original-image groups after averaging the three seeds per pair; they quantify photograph-sampling uncertainty conditional on those runs, not run-to-run standard deviation. Appendix C.6.2 states remaining metric-audit requirements.

The shared-parser results do not support a uniform classification gain across models. Full OGR's change-type accuracy ranges from 64.2% to 83.9%. Relative to B, the differences are $+4.4$, $+0.7$, $+3.0$, $-0.3$, and $-5.4$ points for GPT-4o-mini, GPT-4o, Flash, Pro, and Qwen, respectively; only Qwen's interval excludes zero, in the negative direction. Grounding and classification must be assessed separately: for example, GPT-4o-mini improves mean LocIoU over B by 4.1 points [2.6, 5.6] while its change-type interval includes zero.

**Table 2: VLM evidence generation with shared scoring.** B: single pass; N: two passes without evidence; F: full OGR. Per-pair metrics are averaged over three seeds, then over pairs. Change-type accuracy (400 pairs) uses the same `vlm_only` reading in every condition. Object Recall/StrictHit use 198 pairs; mean LocIoU and ACC@0.25 use 180 text-proxy pairs; mean SalIoU uses 197 pairs. All values are percentages (IoU means multiplied by 100). Run standard deviations are not provided by the paired summary.

| Model | Cond. | ChgType | ObjRecall | StrictHit | LocIoU | ACC@0.25 | SalIoU |
|---|---|---|---|---|---|---|---|
| GPT-4o-mini | B | 62.7 | 21.3 | 32.5 | 8.9 | 12.0 | 8.4 |
| | N | 68.6 | 26.5 | 39.6 | 12.5 | 18.9 | 10.7 |
| | F | 67.1 | 29.3 | 39.7 | 12.9 | 19.6 | 11.9 |
| GPT-4o | B | 83.2 | 35.6 | 59.4 | 23.6 | 32.8 | 21.7 |
| | N | 82.4 | 36.5 | 60.3 | 23.5 | 31.1 | 22.1 |
| | F | 83.9 | 36.4 | 60.3 | 23.7 | 31.1 | 22.0 |
| Gemini-2.5-Flash | B | 68.1 | 41.8 | 64.0 | 31.3 | 51.1 | 28.3 |
| | N | 67.5 | 43.4 | 67.8 | 30.6 | 50.4 | 29.1 |
| | F | 71.1 | 46.8 | 72.4 | 33.3 | 55.7 | 31.9 |
| Gemini-2.5-Pro | B | 74.2 | 57.0 | 80.6 | 30.2 | 52.6 | 36.1 |
| | N | 72.6 | 55.2 | 77.1 | 31.6 | 52.6 | 35.0 |
| | F | 73.9 | 55.3 | 79.0 | 32.0 | 55.0 | 36.4 |
| Qwen3-VL-32B | B | 69.6 | 49.1 | 73.6 | 29.0 | 53.0 | 42.3 |
| | N | 58.2 | 50.7 | 78.5 | 30.6 | 58.3 | 42.6 |
| | F | 64.2 | 51.0 | 80.5 | 30.0 | 56.1 | 42.7 |

#### 4.3.1 What does evidence add beyond a second look?

The matched-call contrast F−N is the primary measure of the added evidence. Change-type accuracy improves by 3.6 points [1.1, 6.1] for Flash and 5.9 [2.6, 9.4] for Qwen; intervals include zero for GPT-4o, Pro, and GPT-4o-mini (Table 3). Qwen's positive F−N coexists with a negative F−B: evidence partly offsets the deterioration caused by evidence-free refinement, but does not recover the single-pass baseline. Flash's mean LocIoU also improves over N by 2.7 points [1.3, 4.3]. Full metric contrasts appear in Appendix D.1.

**Table 3: Paired change-type contrasts under shared `vlm_only` scoring.** B/N/F are defined in Table 2. Means are percentages; differences and 95% CIs are percentage points. Each pair's metric is averaged over three seeds before differencing; the bootstrap resamples original-image groups (`original_path`) with paired conditions. *: interval excludes zero (pointwise, not a multiple-comparison claim). Contrasts are reported from unrounded results, so rounded columns need not subtract exactly.

| Model | B | N | F | F−N [95% CI] | F−B [95% CI] |
|---|---|---|---|---|---|
| GPT-4o-mini | 62.7 | 68.6 | 67.1 | $-1.5$ $[-6.3, +3.3]$ | $+4.4$ $[-0.7, +9.7]$ |
| GPT-4o | 83.2 | 82.4 | 83.9 | $+1.5$ $[-0.5, +3.5]$ | $+0.7$ $[-2.1, +3.5]$ |
| Gemini-2.5-Flash | 68.1 | 67.5 | 71.1 | $+3.6$ $[+1.1, +6.1]^*$ | $+3.0$ $[-0.3, +6.4]$ |
| Gemini-2.5-Pro | 74.2 | 72.6 | 73.9 | $+1.3$ $[-0.9, +3.6]$ | $-0.3$ $[-3.0, +2.4]$ |
| Qwen3-VL-32B | 69.6 | 58.2 | 64.2 | $+5.9$ $[+2.6, +9.4]^*$ | $-5.4$ $[-8.7, -2.2]^*$ |

The two scores and heatmap are added jointly in F−N; this contrast does not isolate their individual contributions. Earlier leave-one-out results used the superseded scoring pipeline and are withheld pending shared-parser rescoring (Appendix D.5). A score-only logistic classifier also requires independent training/validation pairs and is deferred. Aggregate accuracies do not reveal how many initially correct predictions become incorrect; a paired correction/degradation table requires per-pair predictions.

## 5 Scope and Limitations

**Source and tool dependence.** Image overlap and source/tool confounding are different concerns. An original photograph and its edited derivatives must be separated as a group across training, validation, and testing; exact-image hashes alone cannot establish this separation. Even without overlap, a model can exploit class-dependent source or editing-pipeline cues. The within-source results in Sec. 4.2.3 provide partial evidence against collection-name shortcuts, but do not eliminate these finer confounds.

**Retouching scope.** Our GenAI-retouching results are limited to the GP pipeline and its four photographic effects. We have not evaluated an independent GenAI retouching generator. The OpenCV diagnostic (Appendix D.3) measures response to conventional processing and cannot establish such transfer. An additional evaluation must use documented camera originals and generative photographic adjustments that preserve scene content; outputs that add, remove, or replace foreground or background content are not suitable retouching examples under this definition. Appendix A.5 describes additional controls using existing data, explicitly as proposed experiments. These controls can address particular sources of bias, but cannot substitute for an independent-generator test.

**Ambiguous inputs and explanations.** Our labels describe dominant benchmark modifications; mixed editing and benign processing can overlap in channel space. The evidence formulation accommodates these overlaps, but the current benchmark does not establish recovery of composed processing histories. Low camera similarity is not specific to GenAI, and the model has no evaluated open-set rejection mechanism. Reference-based assessment assumes a corresponding trusted image. Reference robustness, semantic-backbone sensitivity, and explanation faithfulness beyond coarse localization proxies remain open tests.

## 6 Conclusion

We presented 2CAP, combining camera-related and semantic representations for four-class photographic assessment and reference-based evidence generation. Reliability-weighted fusion improves discrimination on the evaluated benchmark, while the benefits of OGR without forensic instruction tuning depend on the model and metric. Under shared scoring, evidence improves change-type accuracy over evidence-free refinement for Flash and Qwen, but Qwen remains worse than its single-pass baseline. These findings motivate structured evidence that distinguishes *what changed* from *whether it matters under a given policy*.

## References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

[2] 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. *arXiv preprint arXiv:2307.01952*, 2023.

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *CVPR*, 2023.

[4] Black Forest Labs. Flux.1: State-of-the-art text-to-image generation. https://blackforestlabs.ai, 2024.

[5] Coalition for Content Provenance and Authenticity. C2pa technical specification. https://c2pa.org, 2024.

[6] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*, 2019.

[7] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. In *IEEE TPAMI*, 2022.

[8] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *CVPR*, 2023.

[9] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

[10] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

[11] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

[12] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024.

[13] Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23828–23837, 2025.

[14] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[15] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2023.

[16] Jan Lukas, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE TIFS*, 1(2):205–214, 2006.

[17] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukáš. Determining image origin and integrity using sensor noise. *IEEE TIFS*, 3(1):74–90, 2008.

[18] Tiziano Bianchi and Alessandro Piva. Image forgery localization via block-grained analysis of jpeg artifacts. In *IEEE TIFS*, 2012.

[19] Irene Amerini, Rudy Becarelli, Roberto Caldelli, and Andrea Del Mastio. Splicing forgeries localization through the use of first digit features. In *IEEE WIFS*, 2014.

[20] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE TIFS*, 2020.

[21] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[22] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23850–23859, 2025.

[23] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18685–18694, 2025.

[24] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 7621–7639, 2024. URL https://proceedings.mlr.press/v235/chen24ay.html.

[25] Seunghyun Lee, Byoungkwon Kim, Jaehyun Nam, Kyungmin Lee, and Jinwoo Shin. SSAFE: Simple and strong ai-generated image detection via frozen vision encoders. *arXiv preprint arXiv:2606.08634*, 2026. URL https://arxiv.org/abs/2606.08634.

[26] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025.

[27] 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. *arXiv preprint arXiv:2410.02761*, 2024.

[28] Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.

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

[30] Huangsen Cao, Qin Mei, Zhiheng Li, Yuxi Li, Zhan Meng, Ying Zhang, Chen Li, Zhimeng Zhang, Xin Ding, Yongwei Wang, Jing Lyu, and Fei Wu. REVEAL: Reasoning-enhanced forensic evidence analysis for explainable ai-generated image detection. *arXiv preprint arXiv:2511.23158*, 2025. URL https://arxiv.org/abs/2511.23158.

[31] Rajarshi Roy, Nasrin Imanpour, Ashhar Aziz, Shashwat Bajpai, Gurpreet Singh, Shwetangshu Biswas, Kapil Wanaskar, Parth Patwa, Subhankar Ghosh, Shreyas Dixit, et al. A comprehensive dataset for human vs. ai generated image detection. *arXiv preprint arXiv:2601.00553*, 2026.

[32] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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

[34] Liting Huang, Zhihao Zhang, Yiran Zhang, Xiyue Zhou, and Shoujin Wang. Ru-ai: A large multimodal dataset for machine-generated content detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 733–736, 2025.

[35] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in neural information processing systems*, 37:117340–117362, 2024.

[36] Paschalis Giakoumoglou, Dimitrios Karageorgiou, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Sagi: Semantically aligned and uncertainty guided ai image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16090–16101, 2025.

[37] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[38] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. *CVPR*, 2025.

[39] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[40] Thomas Gloe and Rainer Böhme. The dresden image database for benchmarking digital image forensics. In *ACM SAC*, 2010.

[41] Dasara Shullani, Marco Fontani, Massimo Iuliani, Omar Al Shaya, and Alessandro Piva. Vision: A video and image dataset for source identification. *EURASIP Journal on Information Security*, 2017.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[43] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pages 2405–2414, 2025.

[44] Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 43008–43017, 2026.

[45] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[46] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[47] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[48] Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

[49] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *IEEE WIFS*, 2019.

[50] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020.

[51] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *IEEE MIPR*, 2019.

[52] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019.

[53] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *ICCV*, 2023.

[54] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. Detecting facial retouching using supervised deep learning. *IEEE TIFS*, 11(9):1903–1913, 2016.

[55] Aparna Bharati, Mayank Vatsa, Richa Singh, Kevin W Bowyer, and Xin Tong. Demography-based facial retouching detection using subclass supervised sparse autoencoder. In *IEEE IJCB*, 2017.

[56] Qichao Ying, Jiaxin Liu, Sheng Li, Haisheng Xu, Zhenxing Qian, and Xinpeng Zhang. Retouchingffhq: A large-scale dataset for fine-grained face retouching detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 737–746, 2023.

[57] Fengchuang Xing, Xiaowen Shi, Yuan-Gen Wang, and Chunsheng Yang. Frrffusion: Unveiling authenticity with diffusion-based face retouching reversal. *arXiv preprint arXiv:2405.07582*, 2024.

[58] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

[59] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[60] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. 2010.

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[62] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

# Appendix

## A Dataset Details

We curate a multi-source dataset comprising four categories: camera-captured images, fully synthetic images, tampered images, and GenAI-retouched images. This section provides full details on data composition, construction methodology, and the measures taken to ensure evaluation integrity.

To prevent data leakage, we enforce strict image-level disjointness: no image (camera-captured or edited) appearing in any evaluation split is used during either Stage A (contrastive pretraining) or Stage B (fusion training). For datasets that contribute to both training and evaluation (e.g., SID-Set [29], Defactify [31]), we use the official train/test splits. For held-out datasets (VISION [41], Flickr8k [34]), zero training images are drawn from these sources. We check exact-image duplication using image hashes. This check does not establish that an evaluation image's original photograph, crop, or other edited version is absent from training; the parent-image audit described in Appendix A.5 is an additional requirement.

### A.1 Training Data for Contrastive Learning

We focus on datasets whose real images are verifiably camera-captured photographs, excluding common image-editing benchmarks (*e.g.*, ImgEdit [48]) that contain specialized content such as remote sensing or document images. Such non-photographic imagery would weaken the camera-fingerprint signal our encoder aims to learn.

Table 4 details the training data composition for contrastive learning.

**Table 4: Training data for contrastive learning.** Paired sources provide camera-edit correspondences; unpaired sources contribute additional camera and edited images. The total counts image entries ($2 \times 80{,}465 + 17{,}242 + 15{,}000 + 15{,}000$), not deduplicated unique images. The unique-image count requires a per-image manifest.

| Source | Type | Size |
|---|---|---|
| MagicBrush [37] + SAGID [36] | Paired (camera-edit) | 80,465 pairs |
| Dresden [40] | Unpaired camera | 17,242 |
| SID-Set [29] | Unpaired camera | 15,000 |
| SID-Set [29] | Unpaired edited | 15,000 |
| **Total image entries** | | **208,172** |

### A.2 Evaluation Data Statistics

Table 5 summarizes the dataset splits used for supervised training and evaluation.

**Table 5: Data statistics for supervised training.** Training is class-balanced at 3,200 images per class (camera, retouched, tampered, fully synthetic) for a total of 12,800. * indicates not seen in contrastive learning.

| Dataset | Total training data | Total testing data |
|---|---|---|
| SAGID [36] | 2,240 | 3,838 |
| Defactify* [31] | 3,840 | 22,000 |
| SID-Set [29] | 640 | 16,000 |
| Magic Brush [37] | 2,240 | 0 |
| Dresden [40] | 640 | 0 |
| Generative Photography* [38] | 3,200 | 800 |
| flickr8k* (*held-out*) [34] | 0 | 3,156 |
| Vision* (*held-out*) [41] | 0 | 6,000 |

### A.3 Dataset Construction Details

**Camera images.** We aggregate authentic photographs from complementary sources. To capture classic device-forensics conditions, we include images from the Dresden Image Database (Dresden) [40], which is widely used for camera model identification, and VISION [41], which includes native captures as well as versions processed by online sharing

pipelines. To better reflect modern content distributions encountered in real-vs-synthetic detection, we additionally use the authentic splits of Defactify [31] and SID-Set [29].

**Fully synthetic images.** We define *full synthesis* as images generated entirely by GenAI from text prompts, with no camera capture involved. These images lack any authentic camera fingerprints and may exhibit subtle semantic inconsistencies (lighting, geometry, physics). We compile images from diverse text-to-image generators: Defactify [31] (multiple popular models including SD2.1, SDXL, SD3, DALL-E 3, Midjourney 6) and Flickr8k-derived generations [34] (controlled text-conditioned synthesis). The absence of camera capture does not determine either learned channel response: synthesis can be semantically coherent, and learned camera-related features can respond to processing artifacts. In particular, reference-based cosine scores depend on the chosen query–reference pair and are not fixed by the synthesis label.

**Tampered images.** We define *tampering* as localized GenAI editing that modifies semantic content in specific image regions while potentially preserving camera traces elsewhere. This includes object insertion, removal, replacement, and inpainting operations. We include three complementary families: SID-Set [29] (social-media motivated edits), SAGI-D [36] (AI-based inpainting and region-level modifications), and MagicBrush [37] (instruction-guided editing triplets). Tampering can change semantic content locally while leaving camera-related cues elsewhere; small or coherent edits need not produce low global semantic similarity. Channel responses can overlap with those of other modification types.

**GenAI-retouched images.** We define *retouching* here as GenAI-enabled camera-parameter restyling intended to preserve scene content. We use the Generative Photography dataset [38], covering bokeh, color temperature, focal length, and shutter speed. These edits can alter capture-related processing cues, but their channel scores need not be distinct from every benign or tampered image. Focal-length restyling can also change visible content at crop boundaries, so ambiguous cases require explicit annotation rules.

### A.4 Retouching Dataset Scope

The retouching evaluation uses GP examples covering bokeh, color temperature, focal length, and shutter speed (motion blur). The intended operation changes photographic appearance while preserving scene content; this intention does not guarantee that every generated output meets the definition. Object identity, object presence, and scene layout must be checked independently of detector predictions. Photometric changes affecting foreground or background can be retouching; replacing foreground or background content cannot. Focal-length changes that reveal or remove content at crop boundaries require an explicit annotation policy.

The current experiments do not establish transfer to an independent GenAI retouching generator. Different effects or checkpoints within GP are not independent generator families. The classical OpenCV diagnostic is informative about processing sensitivity but does not close this gap. Broader portrait, computational-photography, and mixed-operation settings also remain outside the validated scope. Suitable external data would require genuine camera originals, documented generative editing provenance, and verified content preservation. Availability of a general image-editing model or a physically simulated retouching target set is not sufficient by itself.

### A.5 Proposed Controls for Source and Processing Confounds

**Status.** The following controls are proposed additional analyses; no results from them are reported in this paper. They are designed to use existing images where provenance and original-edit links can be recovered. The within-source, OpenCV, and mild-processing results in Sec. 4.2.3 are separate, already reported diagnostics.

**Parent-image separation.** Create a joint manifest for Stage A, Stage B, validation, and evaluation. For every image, record the source collection, original-photo identifier, derivative links, operation, tool/checkpoint, and split. Group each original with all crops, recompressions, and edited descendants before splitting. Check both provenance identifiers and exact/near-duplicate image matches across datasets; related source collections can contain the same original under different identifiers. All evaluation groups must be absent from training and model selection. If overlap is found, retrain from the earliest affected stage and rerun the affected comparisons. Report this task-specific audit separately from unknown foundation-model pretraining exposure.

**Matched-original comparison.** Where genuine camera originals for GP outputs are available, evaluate the original and its retouched version using the same frozen checkpoint, restricted to original-photo groups excluded from both training stages. Report camera recall, retouched recall, the four-way confusion matrix, and the paired change in retouched probability. Preserve all four output classes; do not silently discard predictions assigned to tampered or

synthetic. This comparison reduces content/source differences between the two classes, but remains a within-GP evaluation. If original-camera provenance cannot be recovered, this control cannot be claimed.

**Processing controls.** Evaluate native files and a separately reported common-export condition. Apply the same specified resizing and encoding policy to originals and derivatives, and record original dimensions, codec, and export settings. Report the change in performance rather than treating normalization as guaranteed to preserve forensic traces. Use the existing mild-processing and OpenCV examples only after resolving parent overlap and documenting operation parameters and labels. Under a GenAI-specific definition, conventional processing is not positive evidence of GenAI provenance; examine retouched false alarms separately from all non-camera predictions. A source/format-only diagnostic can additionally test how predictive these nuisance variables are of the labels.

**Leave-one-effect-out evaluation.** If the existing GP samples have reliable effect labels, construct four folds: exclude bokeh, color temperature, focal length, or shutter-speed examples in turn from all task-specific training and validation, while maintaining parent-image separation. Retrain the affected components and test the excluded effect on unseen original-photo groups. Ensure that another derivative of a test original does not appear among the three training effects. Keep the remaining-class evaluation protocol fixed and report each fold separately, alongside a control trained with all four effects using a matched training budget. This tests generalization to a held-out photographic effect within GP; it is not a test of an unseen generator.

**Comparison and uncertainty.** Use the same evaluation groups for full 2CAP, semantic-only and camera-only variants, and applicable baselines. Report effect- and source-specific results, sample counts in original-photo groups as well as derivative images, and confidence intervals resampled at the original-photo-group level. Select thresholds and processing settings on validation data. Freeze these choices before testing, and distinguish training-seed variability from uncertainty over photographs.

**Remaining external-data requirement.** None of these controls removes the need for an independent GenAI retouching pipeline if claiming cross-generator transfer. Such a test requires camera-original/generated-retouch pairs with documented provenance and content preservation. An editor that replaces scene content, a set of conventional simulated targets, or a new checkpoint within the same family does not automatically meet that requirement. Until suitable data become available, the retouching claim remains limited to GP and the reported processing diagnostics.

### A.6 Metric Choice Rationale

We select metrics to match the evaluation goal of each experiment:

**Camera authenticity detection.** We report AUC, precision, recall, and F1 for the classification task. Per-class AUC is one-vs-rest. For a four-way softmax, the standard single-label decision is $\arg\max_k p_k$, rather than four independent 0.5 thresholds; the aggregation and decision rules must match those used to produce Table 1.

**Cross-generator generalization.** We report **AUC only**, as the goal is to assess ranking performance across diverse generators rather than to fix a specific operating point. Per-generator AUC isolates how well each method generalizes to individual generator families, while the **Average** column summarizes overall robustness.

**Held-out camera-only datasets (VISION).** For datasets containing only authentic camera images (a single class), we report the **Brier score**:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2,$$

where $p_i$ is the predicted camera probability and $y_i = 1$ for every VISION image. This camera-only score measures squared error on authentic images, but cannot by itself establish calibration across authentic and manipulated classes. AUC requires both positive and negative classes and is therefore not applicable to single-class datasets.

**VLM evidence generation (Sec. 4.3).** We evaluate 400 query–reference pairs (198 tampered, 100 retouched, 102 augmented) under B (single pass), N (two passes without evidence), and F (full OGR). N and F refine the same first-pass output; only F receives the two channel scores and heatmap. Change-type accuracy uses the shared `vlm_only` reading for all three conditions. Object Recall and StrictHit use 198 tampered pairs; mean LocIoU and ACC@0.25 use 180 text-proxy pairs; mean SalIoU uses 197 pairs. Metrics are averaged over three seeds per pair before aggregation. Paired confidence intervals resample original-image groups. Detailed definitions and outstanding implementation checks appear in Appendix C.6. Mock-policy and revision-rate statistics are not included in the updated paired report and are not mixed into the rescored comparison.

## B Extended Related Work

Several concurrent works address related problems. We discuss key differences and clarify 2CAP's positioning.

**Classical forensic cues.** Earlier image forensics focused on hand-crafted traces left by image formation and post-processing pipelines. Sensor-pattern-noise methods exploit device-specific PRNU fingerprints for source attribution and integrity analysis [16][17], while JPEG-based analysis detects recompression inconsistencies and local block artifacts introduced by splicing or resaving [18]. Related statistical approaches also localize tampered regions from distributional irregularities in transform-domain statistics, such as first-digit features [19]. These approaches are physically grounded and often interpretable, but their effectiveness can degrade under resizing, heavy post-processing, or when manipulations do not leave stable low-level traces. In contrast, 2CAP retains access to low-level forensic evidence while embedding it into a learned reference-based comparison framework.

**Learning-based image forgery detection and localization.** Deep IFDL methods replace hand-crafted cues with end-to-end feature learning for tamper detection and localization. ManTra-Net [6] models local anomaly patterns across a wide range of manipulations, MVSS-Net [7] combines multiview and multiscale supervision for manipulation segmentation, and TruFor [8] integrates complementary clues including RGB content and forensic residuals to improve trustworthy localization, building on learned camera fingerprints such as Noiseprint [20]. These methods have substantially improved robustness over classical pipelines, but they remain primarily single-image predictors trained for fixed output spaces (labels or masks). 2CAP is complementary: rather than predicting authenticity from a single image alone, it explicitly compares a query against a reference and returns evidence in a metric space that supports verification, retrieval, and downstream reasoning.

**Early AI-generated image detection.** A parallel line of work studies artifacts specific to synthetic image generation. Early methods analyzed spectral and checkerboard artifacts in GAN outputs [49][50], while others showed that generative models leave characteristic fingerprints that can support both detection and attribution [51][52]. CNNSpot [10] demonstrated that CNN-generated images were surprisingly easy to distinguish under in-domain conditions, helping establish cross-model generalization as a central evaluation criterion. This literature highlights the importance of low-level generator traces, but most methods still formulate the task as direct classification. By contrast, 2CAP uses low-level forensic cues as one component of a reference-based embedding, which better supports fine-grained comparison across images and manipulation settings.

**Generalizable detection across generators.** More recent work emphasizes robustness to unseen synthesis models and post-processing. UnivFD [9] shows that pretrained representations can provide strong transfer across diverse generators with minimal task-specific adaptation, while DIRE [53] exploits reconstruction discrepancies induced by diffusion inversion for diffusion-image detection. Frequency-oriented methods continue this trend by targeting generator-dependent spectral artifacts, including AIDE [11]. D3 [22] further scales multi-generator discrepancy learning with shallow detection heads over pretrained features. Relative to these approaches, 2CAP is less tied to a fixed discriminative decision rule: its contrastive objective and reference-conditioned inference allow the same representation to support verification, nearest-neighbor evidence retrieval, and compatibility with frozen VLM-based reasoning modules.

**Specialized manipulation domains.** Some related work focuses on narrower but practically important manipulation families, especially facial retouching. Earlier methods study supervised retouching detection from facial appearance changes [54][55], and more recent datasets such as RetouchingFFHQ [56] enable fine-grained benchmarking at larger

scale. Diffusion-based restoration or reversal has also been explored for authenticity analysis in this setting [57]. These methods are valuable for domain-specific benchmarks, but they are typically tailored to a particular manipulation family. 2CAP instead targets a broader verification setting spanning AI-generated imagery and image-level forensic comparison without assuming a single manipulation type.

**Model provenance and metadata-based authenticity.** Content provenance standards such as C2PA [5] provide a complementary route to authenticity by attaching cryptographically verifiable provenance metadata at creation or editing time. Such approaches are important for deployment ecosystems, but they depend on adoption across capture and editing pipelines and may be unavailable for legacy or redistributed content. 2CAP is orthogonal to provenance standards: it operates directly on image evidence and remains applicable when trusted metadata is absent, stripped, or incomplete.

**Positioning of 2CAP.** The literature spans (i) classical low-level forensic analysis, (ii) end-to-end image forgery detection and localization, (iii) generalizable AI-generated image detection, and (iv) fine-tuned VLM-based explainable detectors. 2CAP occupies a different point in this design space. We decompose GenAI manipulations along two complementary information channels: a *camera channel* that encodes sensor-level traces, and a *semantic channel* that encodes scene-level coherence. On top of this shared dual-channel foundation we build two applications. **(1) Authenticity classification:** a four-class classifier that distinguishes camera-captured, fully synthetic, tampered, and GenAI-retouched images via cross-attention fusion with learnable reliability weights, moving beyond the binary real/fake formulation that dominates prior work in (i)-(iii). **(2) VLM evidence generation:** dual-channel alignment scores and patch-level saliency are supplied to a *frozen* VLM through an Observe-Generate-Refine loop, eliminating forensic instruction tuning used by FakeVLM [26], SIDA [29], and ForgeryGPT [28] in (iv). The dual-channel decomposition is the unifying primitive that lets a single framework address both authenticity classification and interpretable evidence generation, while the frozen-VLM design avoids the annotation and retraining costs of supervised forensic VLM pipelines.

# C Experimental Details

## C.1 Visualization of Data

We visualize representative examples from each of the four modification categories to illustrate the diversity and difficulty of the evaluation data.

**Camera images.** Figure 4 shows authentic camera-captured photographs from diverse sources (Dresden, VISION, Defactify, SAGID, MagicBrush). These span a wide range of scenes, lighting conditions, and camera models.


**Figure 4: Visualization of camera-captured images** from our evaluation datasets. For Dresden and VISION, we show images from diverse camera models (Agfa, Canon, Nikon, Samsung, Huawei, iPhone, Sony, etc). These authentic photographs serve as both references and positive samples for camera-channel verification.

**Fully synthetic images.** Figure 5 shows text-to-image generations from multiple generators (SD2.1, SDXL, SD3, DALL-E 3, Midjourney 6). These images lack camera fingerprints entirely and may exhibit subtle semantic artifacts.

**Tampered images.** Figure 6 shows localized GenAI edits (object insertion, removal, replacement, inpainting) from SAGID, MagicBrush, and SID-Set. Each example is shown alongside its camera-captured reference to highlight the semantic modifications.


**Figure 5: Visualization of fully synthetic images** generated by different text-to-image models. These images contain no camera traces and may show subtle semantic inconsistencies in lighting, geometry, or physics.


**Figure 6: Visualization of tampered images** with their camera-captured references. Tampered images contain localized semantic edits; their effects on semantic similarity and camera-related cues depend on the edit and can overlap with other categories.

**GenAI-retouched images.** Figure 7 shows global camera-parameter restyling from the Generative Photography dataset, covering bokeh adjustment, color temperature shift, focal length change, and shutter speed modification. Each example is paired with its original camera capture.

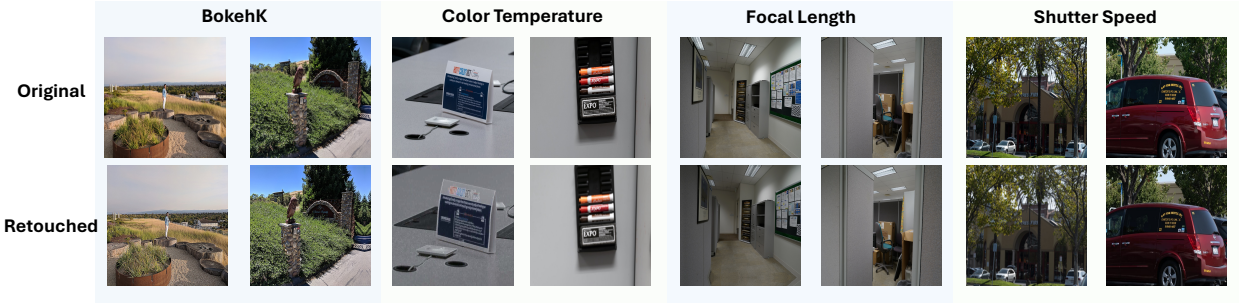


**Figure 7: Visualization of GenAI-retouched images** with their original camera captures. These examples show retouching with low camera-feature similarity and high semantic similarity; actual responses may overlap with tampering.

### C.2 Model Architecture and Training Details

We provide complete implementation details for reproducibility.

**Camera encoder $\mathcal{E}_{\text{cam}}$.** We use a DCT-domain dual-branch pipeline with patch-level frequency analysis [58]. Images are split into $32 \times 32$ patches, ranked by DCT frequency energy, and selected patches are processed with 30 SRM steganalysis filters [11] to extract noise residuals. Two ResNet-50 backbones [42] separately encode high-frequency (textured) and low-frequency (smooth) patches; branch features are fused by element-wise addition and projected to 128-d unit-norm embeddings.

**Semantic encoder $\mathcal{E}_{\text{sem}}$.** We use frozen SigLIP2 [14] (ViT-SO400M) to extract 768-d global semantic embeddings. For the saliency map, we extract $27 \times 27$ patch-level features and compute pairwise cosine similarities between query and reference patches.

**Cross-attention fusion.** The camera and semantic patch-token sequences retain their respective patch counts when

projected to 256-d via learned linear layers (Sec. 3.3). Bidirectional 4-head cross-attention (2 layers) normalizes over key patches in each direction. After channel-wise pooling, a 2-layer gating MLP takes the concatenated vectors and produces two logits; a softmax over channels converts these into input-dependent weights that sum to one. A 2-layer classification head (256→128→4) outputs probabilities over camera-captured, retouched, tampered, and fully synthetic images. The cross-attention fusion architecture is depicted in Fig. 8.

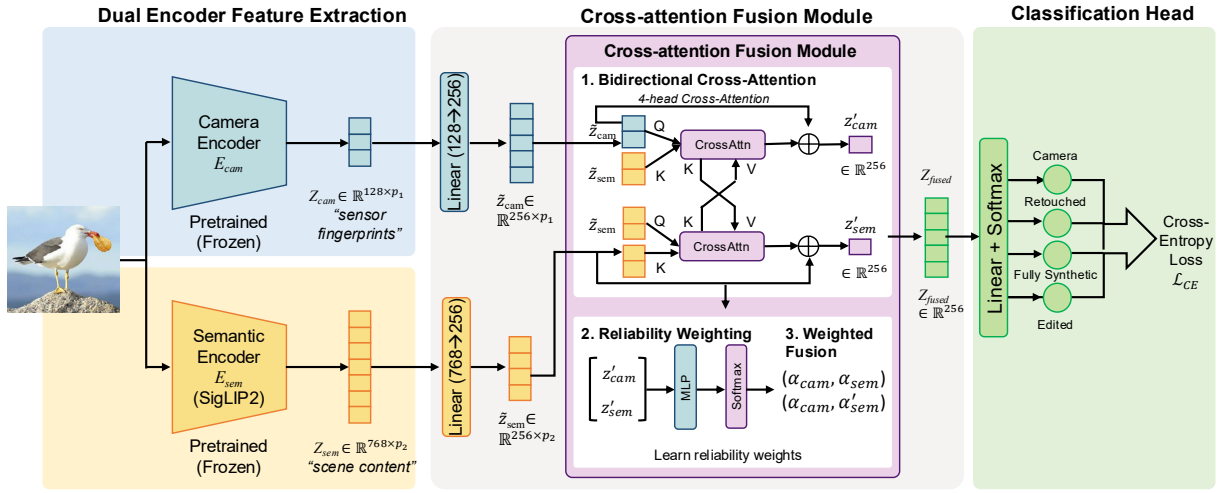


**Figure 8: Cross-channel attention fusion for authenticity classification.** Camera patch tokens $Z_{\text{cam}}$ and semantic patch tokens $Z_{\text{sem}}$ are projected to a shared 256-d space, then refined via bidirectional 4-head cross-attention. A gating MLP followed by a two-channel softmax produces weights $(\alpha_{\text{cam}}, \alpha_{\text{sem}})$ that sum to one and adaptively combine the pooled channel representations. The fused representation feeds a four-class head trained with cross-entropy. The token-sequence axes are retained through attention before the channel representations are combined.

**Contrastive pretraining for $\mathcal{E}_{\text{cam}}$.** We pretrain $\mathcal{E}_{\text{cam}}$ on 208,172 image entries (not a verified unique-image count) with InfoNCE loss (data composition in Table 4). Positives include mild augmentations (horizontal flip, mild color jitter $\pm 0.1$, JPEG quality $\geq 90$) and, when available, same-camera images; negatives include paired edits and batch negatives. Training uses AdamW [59] optimizer (lr $3 \times 10^{-4}$, weight decay 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$) for 20 epochs on 4× NVIDIA A100 80GB GPUs with batch size 512. Temperature-weighted sampling balances dataset sizes. Augmentations avoid blur and heavy JPEG recompression to limit processing changes; this does not guarantee preservation of sensor fingerprints.

**Supervised fusion training.** With both encoders frozen, we train only the fusion module (cross-attention layers + reliability MLP + classification head, ∼2M parameters) on the supervised four-class authenticity task. Training uses AdamW (lr $10^{-3}$) with cosine annealing schedule, batch size 2048, for 50 epochs, selecting the best checkpoint by validation AUC. Cross-entropy loss with label smoothing (0.1) is used.

### C.3 Architectural Comparison with AIDE

Table 6 explicitly delineates the architectural differences between 2CAP's camera encoder and AIDE [11], which inspired the frequency-aware patch selection design.

**Summary.** The low-level feature extraction pipeline is shared with AIDE; 2CAP's novelty lies in (1) contrastive pretraining for a metric embedding space enabling reference-based verification, (2) dual-channel decomposition with cross-attention fusion, and (3) VLM evidence synthesis via OGR.

### C.4 Inference Pipeline

Algorithm 2 summarizes the inference procedure of 2CAP. For each channel $c \in \{\text{cam}, \text{sem}\}$, $z_c^v = \mathcal{E}_c^g(\mathbf{x}_v)$ denotes a global embedding, whereas $Z_c^v = \mathcal{E}_c^{\text{tok}}(\mathbf{x}_v)$ is a patch-token sequence, for $v \in \{q, r\}$. The camera global embedding is the 128-d unit-norm contrastive representation. The superscripts $g$ and $\text{tok}$ denote distinct outputs of the same encoder, not additional encoders. Q1 uses only query tokens: cross-attention precedes pooling, gating, and four-class classification. Q2 compares global query/reference embeddings for cosine scores and semantic patch tokens for saliency;

**Table 6: Architectural comparison: 2CAP camera encoder vs. AIDE.**

| Component | AIDE | 2CAP |
|---|---|---|
| Patch selection | DCT energy ranking | DCT energy ranking (same) |
| Filter bank | 30 SRM filters | 30 SRM filters (same) |
| Backbone | Dual-branch ResNet-50 | Dual-branch ResNet-50 (same) |
| Training objective | Cross-entropy (binary) | InfoNCE contrastive |
| Training data | Single-image labels | Paired camera-edit + unpaired |
| Embedding space | Class logits | 128-d unit-norm (metric space) |
| Downstream use | Binary classifier | Cosine similarity $s_{\text{cam}}$ |
| Semantic channel | CLIP + linear proj. | Frozen SigLIP2 + cross-attn |
| VLM integration | None | OGR privileged-context loop |

these signals guide the frozen VLM. Q1's four-class prediction and Q2's three-class change assessment are distinct outputs; the reference is required only for Q2.

**Algorithm 2** 2CAP Inference Pipeline

**Require:** Query image $\mathbf{x}_q$, reference image $\mathbf{x}_{\text{ref}}$
**Ensure:** Q1 four-class prediction, Q2 forensic explanation, saliency map

$z_c^v \leftarrow \mathcal{E}_c^g(\mathbf{x}_v)$ for $c \in \{\text{cam}, \text{sem}\}$, $v \in \{q, r\}$ *// Global embeddings for Q2*
$Z_c^q \leftarrow \mathcal{E}_c^{\text{tok}}(\mathbf{x}_q)$ for $c \in \{\text{cam}, \text{sem}\}$ *// Query patch tokens for Q1*
$Z_{\text{sem}}^r \leftarrow \mathcal{E}_{\text{sem}}^{\text{tok}}(\mathbf{x}_{\text{ref}})$ *// Reference patch tokens for Q2*
$s_{\text{cam}} \leftarrow \cos(z_{\text{cam}}^q, z_{\text{cam}}^r)$; $s_{\text{sem}} \leftarrow \cos(z_{\text{sem}}^q, z_{\text{sem}}^r)$
$M_i \leftarrow 1 - \max_j \cos(Z_{\text{sem},i}^q, Z_{\text{sem},j}^r)$ *// Eq. 5*
$\mathbf{z}_{\text{fused}} \leftarrow \text{CrossAttnFusion}(Z_{\text{cam}}^q, Z_{\text{sem}}^q)$ *// Attention, pooling, gating; Q1 only*
$\mathbf{p} \leftarrow \text{softmax}(h_{\text{cls}}(z_{\text{fused}}))$; $\hat{y} \leftarrow \arg\max_k p_k$ *// Classification*
*// OGR Loop:*
obs $\leftarrow$ VLM.Observe$(\mathbf{x}_q, \mathbf{x}_{\text{ref}})$
draft $\leftarrow$ VLM.Generate(obs) *// Initial explanation*
output $\leftarrow$ VLM.Refine(draft, $s_{\text{cam}}, s_{\text{sem}}, M$) *// Score-guided refinement*
**return** $\hat{y}$, output, $M$

### C.5 Additional VLM Evidence Generation Examples

We present additional qualitative examples of the Stage 3 OGR pipeline operating on different modification types. Figures 10-12 illustrate how the VLM adapts its reasoning based on the dual-channel scores and saliency maps.

### C.6 VLM Evaluation Setup Details

This section provides detailed definitions of the metrics used in Section 4.3 for evaluating VLM-generated evidence. The VLM output examples and formats can refer to App. C.5.

#### C.6.1 Change Type Classification

We measure three-class accuracy over tampered, retouched, and augmented. Ground-truth addition/removal/replacement/multiple labels map to tampered, retouching to retouched, and augmentation to augmented. The updated B/N/F comparison uses the same `vlm_only` reading throughout; the historical condition-specific numerical camera-score override is not used. The paired summary does not expose the full parser implementation or invalid-output handling, which require the evaluation script and saved predictions to audit.

#### C.6.2 Metric Reconciliation and Shared Scoring

The current B/N/F tables replace the legacy whole-pipeline aggregates with the supplied shared-reading paired evaluation. The older localization column is replaced by separately reported mean LocIoU and ACC@0.25. The report contains 180 localization pairs and 197 SalIoU pairs. The underlying code must still establish the exact box-matching aggregation,

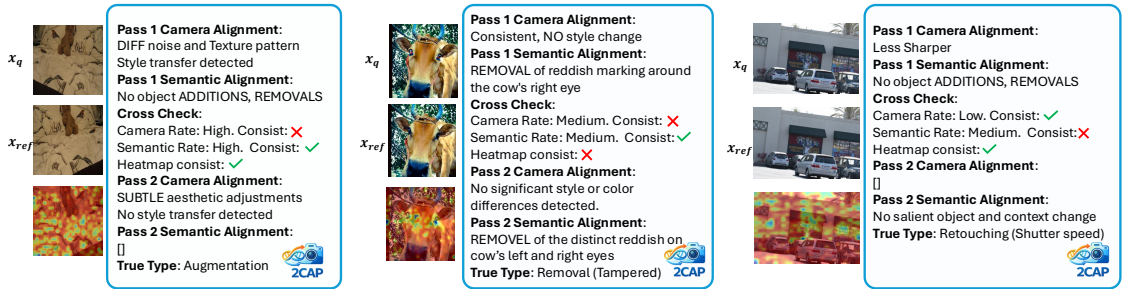


**Figure 9: 2CAP evidence generation case study.** For each query-reference pair, the VLM first assesses the images, then receives channel scores and a saliency map in a refinement pass. The examples illustrate augmentation, tampered removal, and retouching.

whether ACC@0.25 uses $>$ or $\geq$, the SalIoU target construction, and the missing-output policy; we do not infer these implementation details from metric names.

**Uncertainty and unresolved outputs.** Three seeds are averaged within each pair before contrasts are computed; confidence intervals resample original-image clusters using `original_path`. They do not estimate variability across new VLM runs. Macro-F1, per-class recall, invalid-output rates, per-run standard deviations, and the four correction/degradation counts are absent from the paired summary and cannot be reconstructed from aggregate accuracy. In particular, F$-$N is the net of incorrect-to-correct and correct-to-incorrect rates, not the count of corrected predictions. These statistics require the paired outputs and run identities. A score-only logistic regression remains deferred until independent training/validation pairs are available.

### C.6.3 Object Recall

Object recall measures how completely the VLM identifies changed objects described in the ground truth, which is evaluated on tampered pairs only. Specifically, our computation follows the following steps:

1. Extract GT object nouns from the annotation by tokenizing, stemming (plurals, -ing forms), and filtering stopwords, adjectives, colors, and positional words.
2. Concatenate all predicted `object_changes[].description` strings.
3. Match GT nouns to predictions using substring matching, stemming, and a synonym dictionary (e.g., man/woman $\rightarrow$ person, sofa $\rightarrow$ couch).
4. Compute: ObjRecall $=$ (# GT nouns found)$/$(# total GT nouns).

We report the averaged Recall across all tampered pairs.

### C.6.4 Object Hit

It is a stricter object detection metric requiring both *sufficient noun overlap* and *action type agreement*. Evaluated on tampered pairs only.
**Requirements for a hit:**

1. *Noun threshold*: At least $\min(2, |\text{GT nouns}|)$ GT nouns must appear in predictions.
2. *Action match*: The predicted action type must be compatible with the GT action type.

**Action compatibility rules:**

- Same action (addition/removal/replacement) $\rightarrow$ match.
- Replacement is compatible with both addition and removal.
- If either action is undetectable $\rightarrow$ match (no penalty).
- Addition vs. removal (without replacement) $\rightarrow$ no match.

### C.6.5 Localization Proxy

Localization is evaluated against proxy GT boxes parsed from free-text annotations on tampered pairs with parseable locations. These coarse spatial targets are separate from mask-derived boxes and pixel-level masks.
**Location parsing.** Free-text location descriptions (e.g., "bottom-right", "left side") are mapped to approximate bboxes on a 0-1000 coordinate system:

- Horizontal: `left` → [0, 500], `right` → [500, 1000], `center` → [250, 750].
- Vertical: `top/upper` → [0, 500], `bottom/lower/foreground` → [500, 1000].
- Corners are tightened to one-third regions (e.g., "top-left corner" → [0, 0, 333, 333]).

Object-relative descriptions (e.g., "on the dog's face") and "entire image" are skipped.

## C.7 Policy Compliance Evaluation on Mock Policy

A key application of forensic evidence is automated content policy enforcement. We implement a rule-based policy checker that evaluates the VLM-generated explanations for policy violations. The checker uses regex pattern matching to identify specific keywords and outputs a **Pass/Fail** status based on the following criteria:

- **Biometric Modification (Fail):** Any alteration of physical identity markers such as skin tone, eye color, hair texture, facial features, or facial geometry.
- **Sensitive Addition/Removal (Fail):** The addition or removal of protected entities, including people, weapons, illegal substances, and hate symbols.
- **Permitted Edits (Pass):**
  - *Aesthetic Adjustments:* Gen AI retouching, compression artifacts, and camera parameter restyling.
  - *Non-Sensitive Content:* Color changes, additions, or removals of animals and inanimate objects not included in the sensitive list.
  - *Environment:* General modifications to the image background.

**Implementation details.** The policy checker examines two sections of the VLM output:

1. *Biometric check*: Scan `identity_biometric` for change phrases ("facial geometry changed", "identity swap", "face morph") unless overridden by safe phrases ("no identity manipulation detected", "facial geometry is consistent"). This section is unchanged between passes.
2. *Sensitive entity check*: For each `object_changes[]` entry with `change_type` $\in$ {addition, removal, replacement}, check the `description` for person keywords (29 terms: person, man, woman, child, pedestrian, etc.) or sensitive keywords (14 terms: weapon, gun, knife, drug, hate symbol, etc.). In the +OGR configuration, `revised_semantic_alignment` is used when present, providing the revised object-change list from Pass 2.

**Metrics.** We evaluate image-level detection with *reject* as the positive class: Precision (TP/(TP+FP)), Recall (TP/(TP+FN)), Accuracy, and False Positive Rate (FP/(FP+TN)). Ground truth verdicts (365 accept, 35 reject) are human-annotated.

# D Additional Experiments

**Run-variability reconciliation.** Shared `vlm_only` rescoring replaces the legacy Flash change-type mean of 78.7% with 71.1% for F. The old 0.6-versus-2.6 standard-deviation conflict cannot be resolved from seed-averaged paired contrasts. We report paired confidence intervals and withhold run standard deviations until the three per-run values and identifiers are available; the old spreads must not be attached to the new means.

## D.1 Evidence-Free Two-Pass Self-Refinement Control

B receives the image pair in one pass; N and F refine the same first-pass output, respectively without and with the two scores and heatmap. Table 7 gives the evidence contrast F−N and total two-pass contrast F−B for all five models. The estimator averages each metric over three seeds per pair before taking paired differences. Bootstrap resampling uses `original_path` clusters. The intervals describe sampling uncertainty over original images conditional on the observed runs; their relative widths are empirical, not guaranteed by the contrast definition.

**Table 7: Paired evidence and total-refinement effects.** Differences and pointwise 95% CIs in percentage points (IoU differences multiplied by 100). * denotes a CI excluding zero. Metric denominators: change type 400; object recall/StrictHit 198; LocIoU/ACC@0.25 180; SalIoU 197. B/N/F means are in Table 2.

| Model | Metric | $n$ | F−N [95% CI] | F−B [95% CI] |
|---|---|---|---|---|
| GPT-4o-mini | ChangeType | 400 | $-1.5$ $[-6.3, +3.3]$ | $+4.4$ $[-0.7, +9.7]$ |
| | ObjRecall | 198 | $+2.8$ $[+1.2, +4.5]^*$ | $+7.9$ $[+5.9, +10.2]^*$ |
| | StrictHit | 198 | $+0.2$ $[-3.2, +3.6]$ | $+7.2$ $[+3.0, +11.7]^*$ |
| | LocIoU mean | 180 | $+0.5$ $[-0.9, +1.9]$ | $+4.1$ $[+2.6, +5.6]^*$ |
| | LocIoU ACC.25 | 180 | $+0.7$ $[-2.4, +3.9]$ | $+7.6$ $[+3.7, +11.7]^*$ |
| | SalIoU mean | 197 | $+1.3$ $[+0.2, +2.4]^*$ | $+3.5$ $[+2.0, +5.2]^*$ |
| GPT-4o | ChangeType | 400 | $+1.5$ $[-0.5, +3.5]$ | $+0.7$ $[-2.1, +3.5]$ |
| | ObjRecall | 198 | $-0.1$ $[-0.8, +0.6]$ | $+0.8$ $[-0.4, +2.0]$ |
| | StrictHit | 198 | $-0.0$ $[-1.8, +1.7]$ | $+0.8$ $[-2.7, +4.4]$ |
| | LocIoU mean | 180 | $+0.2$ $[-0.1, +0.5]$ | $+0.1$ $[-1.6, +2.0]$ |
| | LocIoU ACC.25 | 180 | $+0.0$ $[+0.0, +0.0]$ | $-1.7$ $[-5.2, +1.7]$ |
| | SalIoU mean | 197 | $-0.1$ $[-0.3, +0.1]$ | $+0.3$ $[-1.2, +1.8]$ |
| Gemini-2.5-Flash | ChangeType | 400 | $+3.6$ $[+1.1, +6.1]^*$ | $+3.0$ $[-0.3, +6.4]$ |
| | ObjRecall | 198 | $+3.5$ $[+1.9, +5.2]^*$ | $+5.0$ $[+2.4, +7.8]^*$ |
| | StrictHit | 198 | $+4.5$ $[+1.9, +7.4]^*$ | $+8.4$ $[+3.0, +13.9]^*$ |
| | LocIoU mean | 180 | $+2.7$ $[+1.3, +4.3]^*$ | $+2.0$ $[-0.3, +4.4]$ |
| | LocIoU ACC.25 | 180 | $+5.4$ $[+2.4, +8.5]^*$ | $+4.6$ $[-0.4, +9.8]$ |
| | SalIoU mean | 197 | $+2.9$ $[+1.5, +4.4]^*$ | $+3.6$ $[+1.3, +6.1]^*$ |
| Gemini-2.5-Pro | ChangeType | 400 | $+1.3$ $[-0.9, +3.6]$ | $-0.3$ $[-3.0, +2.4]$ |
| | ObjRecall | 198 | $+0.2$ $[-1.0, +1.4]$ | $-1.7$ $[-3.7, +0.4]$ |
| | StrictHit | 198 | $+1.9$ $[-0.5, +4.2]$ | $-1.7$ $[-6.1, +2.7]$ |
| | LocIoU mean | 180 | $+0.4$ $[-0.6, +1.5]$ | $+1.8$ $[+0.4, +3.3]^*$ |
| | LocIoU ACC.25 | 180 | $+2.4$ $[+0.4, +4.4]^*$ | $+2.4$ $[-1.5, +6.5]$ |
| | SalIoU mean | 197 | $+1.4$ $[+0.4, +2.5]^*$ | $+0.3$ $[-1.3, +1.9]$ |
| Qwen3-VL-32B | ChangeType | 400 | $+5.9$ $[+2.6, +9.4]^*$ | $-5.4$ $[-8.7, -2.2]^*$ |
| | ObjRecall | 198 | $+0.4$ $[-1.4, +2.3]$ | $+2.0$ $[+0.1, +3.9]^*$ |
| | StrictHit | 198 | $+2.0$ $[-1.0, +5.2]$ | $+6.9$ $[+3.4, +10.7]^*$ |
| | LocIoU mean | 180 | $-0.6$ $[-1.8, +0.5]$ | $+1.0$ $[-0.3, +2.4]$ |
| | LocIoU ACC.25 | 180 | $-2.2$ $[-4.8, +0.2]$ | $+3.1$ $[-0.6, +7.0]$ |
| | SalIoU mean | 197 | $+0.1$ $[-1.1, +1.2]$ | $+0.4$ $[-1.0, +2.0]$ |

### D.2 Within-Source Classification Diagnostics

Table 8 reports two comparisons restricted to a named source collection. This restriction removes variation in the collection name within each comparison, but does not necessarily match underlying original-photo sources, compression, resolution, or generator signatures across classes. The SID-Set AUC indicates ranking discrimination within that collection. The Defactify accuracy additionally depends on class proportions and the decision rule. These metrics therefore cannot be averaged or interpreted as proof that source confounding is absent. Neither comparison includes the GP retouching class.

**Table 8: Within-source diagnostics.** Values are retained in the metric originally reported for each comparison. These are not matched-original or independent-generator evaluations.

| Collection | Comparison | Metric | Result |
|---|---|---|---|
| SID-Set | Camera vs. tampered | AUC | 95.5% |
| Defactify | Camera vs. fully synthetic | Accuracy | 90.4% |

### D.3 Cross-Pipeline Retouching Diagnostic

The OpenCV data are used only for evaluation and are mapped exclusively to the GP evaluation subset; they are not used in either training stage. Table 9 reports the predicted-class distributions for GP effects and corresponding effects recreated using classical OpenCV operations on the associated evaluation source images. This establishes the intended evaluation split; auditing underlying parent identities across other source collections is a separate check. The OpenCV outputs are assigned the retouched label in 81.2-92.7% of cases. This is evidence of sensitivity to related visual processing beyond the GP implementation; it is not a test of attribution to GenAI, because the OpenCV operations are non-neural. It also does not replace evaluation on an entirely unseen GenAI retouching tool. Under the GenAI-specific retouching definition, a conventional operation alone does not establish membership in the GenAI-retouched class, regardless of its visual strength. The predictions therefore expose a potential provenance ambiguity. Operation severity and the policy for conventional edits must be specified before this diagnostic can be scored as a classification task.

**Table 9: Cross-pipeline diagnostic.** Row-normalized predicted-class percentages. "Retouched" is a model prediction, not proof of GenAI processing or an established correctness label for OpenCV operations.

| Pipeline | Effect | Camera | Retouched | Tampered | Synthetic |
|---|---|---|---|---|---|
| GP | Color temperature | 1.0 | 97.0 | 2.0 | 0.0 |
| GP | Bokeh | 1.5 | 94.5 | 4.0 | 0.0 |
| GP | Focal length | 5.5 | 84.5 | 10.0 | 0.0 |
| GP | Shutter speed | 1.0 | 97.0 | 1.5 | 0.5 |
| OpenCV | Color temperature | 3.2 | 90.6 | 5.4 | 0.8 |
| OpenCV | Bokeh | 2.7 | 81.2 | 15.3 | 0.8 |
| OpenCV | Focal length | 4.4 | 90.0 | 5.4 | 0.2 |
| OpenCV | Shutter speed | 0.3 | 92.7 | 6.7 | 0.3 |

### D.4 Mild Benign-Processing Diagnostic

Table 10 reports a limited diagnostic under mild settings. Camera recall decreases by at most 4.0 points relative to the unprocessed images. However, the unprocessed camera recall is only 62.5%, corresponding to a 37.5% non-camera prediction rate before any transformation; denoising raises that rate to 41.5%. Small relative changes therefore do not establish a low false-positive rate. This diagnostic does not cover operation severity, real smartphone capture pipelines, or compound processing.

### D.5 OGR Component Ablation Status

The historical leave-one-out component table is excluded from the current comparison pending rescoring with the shared `vlm_only` rule. F−N tests the joint contribution of both scores and the heatmap; it does not identify a camera-only, semantic-only, or heatmap-only effect. No score-only classifier is reported without independent training/validation pairs.

**Table 10: Mild benign-processing diagnostic.** Predicted-class percentages on camera-derived images. All rows should retain the camera label under this diagnostic's policy.

| Transformation | Camera | Tampered | Synthetic | Retouched |
|---|---|---|---|---|
| No processing | 62.5 | 26.5 | 9.5 | 1.5 |
| Resize | 60.5 | 29.0 | 9.0 | 1.5 |
| Denoise | 58.5 | 29.5 | 10.0 | 2.0 |
| HDR tonemap | 59.5 | 27.5 | 11.0 | 2.0 |
| Social recompress | 60.5 | 27.0 | 11.0 | 1.5 |
| Computational photo | 62.0 | 26.0 | 10.5 | 1.5 |

### D.6 Per-Generator Classification Results

Table 11 provides a per-generator breakdown of camera-vs-edited AUC across the Defactify (in-distribution) and Flickr8k* (held-out) datasets, complementing the dataset-level summary reported elsewhere. 2CAP achieves the best AUC on every Defactify generator (.892-.923), showing strong within-source performance across the listed diffusion families (SD2.1/SDXL/SD3), DALL-E 3, and Midjourney rather than overfitting to any one generator. On the held-out Flickr8k* set, FatFormer leads on all five generators while 2CAP remains competitive (.772-.929), and the two methods are essentially tied on the ten-column average (2CAP: .893; FatFormer: .894).

**Table 11: Camera authenticity across generators (AUC).** Defactify and Flickr8k*, each with five generators. *Flickr8k** is held-out. Best in **bold**, second underlined.

| **Method / AUC** | **Defactify Images** | | | | | **Flickr8k*** | | | | | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **SD2.1** | **SDXL** | **SD3** | **DALL-E 3** | **Midjourney 6** | **SDAbso.** | **SDEpic.** | **SDv1.5** | **SDv6.0** | **SDXL** | |
| TruFor [8] | .455 | .407 | .382 | .452 | .516 | .656 | .668 | .660 | .550 | .689 | .544 |
| D3 [22] | .613 | .715 | .618 | .778 | .762 | .723 | .693 | .668 | .643 | .635 | .685 |
| CNNSpot [10] | .561 | .547 | .542 | .606 | .560 | .625 | .574 | .598 | .550 | .533 | .570 |
| AIDE [11] | .626 | .701 | .633 | .567 | .862 | .929 | .905 | .895 | .835 | .796 | .775 |
| VIBAIGCD [13] | .757 | .808 | .740 | .844 | .820 | .932 | .924 | .922 | .903 | .698 | .835 |
| FreqNet [21] | .541 | .584 | .522 | .544 | .579 | .851 | .868 | .837 | .843 | .823 | .699 |
| FatFormer [12] | .836 | .860 | .812 | .895 | .839 | **.967** | **.962** | **.961** | **.958** | **.850** | **.894** |
| UnivFD [9] | .751 | .781 | .752 | .852 | .805 | .876 | .861 | .852 | .845 | .676 | .805 |
| **2CAP** | **.895** | **.899** | **.892** | **.923** | **.911** | .929 | .905 | .910 | .893 | .772 | .893 |

### D.7 Matched-source Four-class Diagnostic

We additionally test whether four-way predictions remain reliable when all four labels are derived from the same source scene. We select 100 original-retouched pairs from the existing Generative Photography test-source pool, with 25 pairs per camera setting (bokeh, color temperature, focal length, and shutter speed). The camera and retouched images are the existing paired files. For each camera original, Big-LaMa removes a masked object to form a tampered candidate, with pixels outside the mask copied from the original. Separately, FLUX.1-dev generates a fully synthetic image from the original's caption alone, without image conditioning. This produces 100 source groups and 400 images (100 per class).

We evaluate frozen Stage A and Stage B checkpoints without retraining or tuning on these images. Camera and SigLIP2 patch-token embeddings are extracted once and supplied to the full model and its camera-only and semantic-only ablations. Each model receives one image at a time, without access to the group identity. Table 12 reports four-way argmax accuracy, macro-F1, macro one-vs-rest AUC, and class recalls. The full model attains 46.25% accuracy, compared with 44.50% for semantic-only. Its 1.75-point accuracy difference has a paired 95% bootstrap interval of $[-0.75, 4.50]$ points when resampling the 100 original-image groups, and therefore does not establish a reliable gain over the semantic channel alone. In particular, the full model identifies only 4% of camera images and 8% of tampered candidates.

The full model's recalls are 4%, 99%, 8%, and 74% for camera, retouched, tampered, and fully synthetic images, respectively; the corresponding per-class one-vs-rest AUCs are 0.450, 0.848, 0.582, and 0.999. These ranking scores should not obscure the poor camera and tampered argmax recalls. The cohort reuses the paper's existing test-source

**Table 12: Matched-source four-class diagnostic.** All values are percentages.

| Frozen model | Acc. | Macro-F1 | Macro-AUC |
|---|---|---|---|
| Full 2CAP | 46.25 | 39.14 | 71.99 |
| Camera only | 28.75 | 23.88 | 64.33 |
| Semantic only | 44.50 | 36.51 | 66.42 |

pool, rather than providing a new independent model-selection holdout. Exact source-path and pixel-change checks passed, but visual review of all Big-LaMa removals and scene consistency is incomplete; some spot-checked removals are small or visibly blurred. We therefore treat this as a failure-oriented diagnostic, not evidence that matched-source confounding has been resolved. The group manifest, model-specific generation records, predictions, and inference logs are retained with the dataset for auditing.

### D.8 Validation of VLM Evaluation with Annotated Bounding Box

In Sec. 4 of the main paper, the VLM evidence evaluation uses location text as a proxy ground truth for localization quality (IoU) as not all the datasets have bounding-box annotation. To validate this proxy, we manually annotate bounding boxes for a subset of tampered image pairs from SAGID (those VLM actually produces a bounding box), using the ground-truth masks provided by the dataset. We then compare the VLM-generated bounding boxes against both the human-annotated boxes and the saliency-derived proxy boxes.

**Table 13: Proxy GT validation.** Agreement between VLM-derived bounding boxes and human-annotated bounding boxes on 50 tampered pairs.

| | Baseline | | Ours | |
|---|---|---|---|---|
| Metric | Mean | Std | Mean | Std |
| IoU | .500 | .241 | .597 | .228 |
| Precision | .722 | .244 | .823 | .203 |
| Recall | .638 | .297 | .690 | .254 |

Table 13 reports higher agreement with mask-derived human boxes for OGR on the selected 50-pair subset (IoU .500 to .597). This supports localization on that subset, but does not directly validate the text-derived proxy: doing so requires comparing proxy boxes themselves with mask-derived boxes. Selecting pairs on which a VLM outputs a box can also exclude localization failures. The subset selection and missing-box handling therefore limit the conclusion.

### D.9 Design-Space Comparison with VLM-Based Forensic Methods

Table 14 provides a systematic design-space comparison with VLM-based forensic methods, clarifying fundamental architectural and formulation differences.

**Scope of a direct comparison.** Forensic VLMs differ in reference access, training data, and output format. These differences require explicit reporting; they do not make practical comparison infeasible. Our current experiments measure improvements to the same frozen VLM under pairwise inputs, rather than competitiveness against fine-tuned forensic systems. A complementary comparison could evaluate a released forensic VLM on a shared camera/tampered subset with mask-derived boxes, showing its native single-image results separately from any pair-input adaptation and reporting the reference advantage of OGR.

**Key advantage.** The OGR interface uses frozen VLMs without forensic instruction tuning, while the camera encoder and fusion classifier are trained. The experiments demonstrate improvements across the tested proprietary and open-weight configurations. Extension to other backends is plausible but requires evaluation, and the current results do not establish a performance ranking against fine-tuned forensic VLMs.

### D.10 Discussion of Reference-Based Baselines

We consider the following reference-based similarity alternatives for change assessment (Sec. 4.3):

**Table 14: Design-space comparison with VLM-based forensic methods.** 2CAP is the only method that (i) uses reference-based verification, (ii) decomposes camera and semantic channels, (iii) distinguishes retouching from tampering, and (iv) uses the tested VLMs without forensic instruction tuning.

| Property | FakeVLM | FakeShield | ForgeryGPT | SIDA | 2CAP |
|---|---|---|---|---|---|
| Input type | Single | Single | Single | Single | **Pair** |
| VLM training | SFT | SFT | Staged tuning | SFT | **Frozen** |
| Channel decomp. | × | × | × | × | ✓ |
| Retouch vs. tamp. | × | × | × | × | ✓ |
| Forgery localization | × | Mask | Mask | Mask | **Saliency** |
| Output type | Binary+text | Binary+mask | Desc.+mask | Label+mask+desc. | **Scores+sal.+OGR** |
| VLM portability | Fixed | Fixed | Fixed | Fixed | **Tested VLMs** |

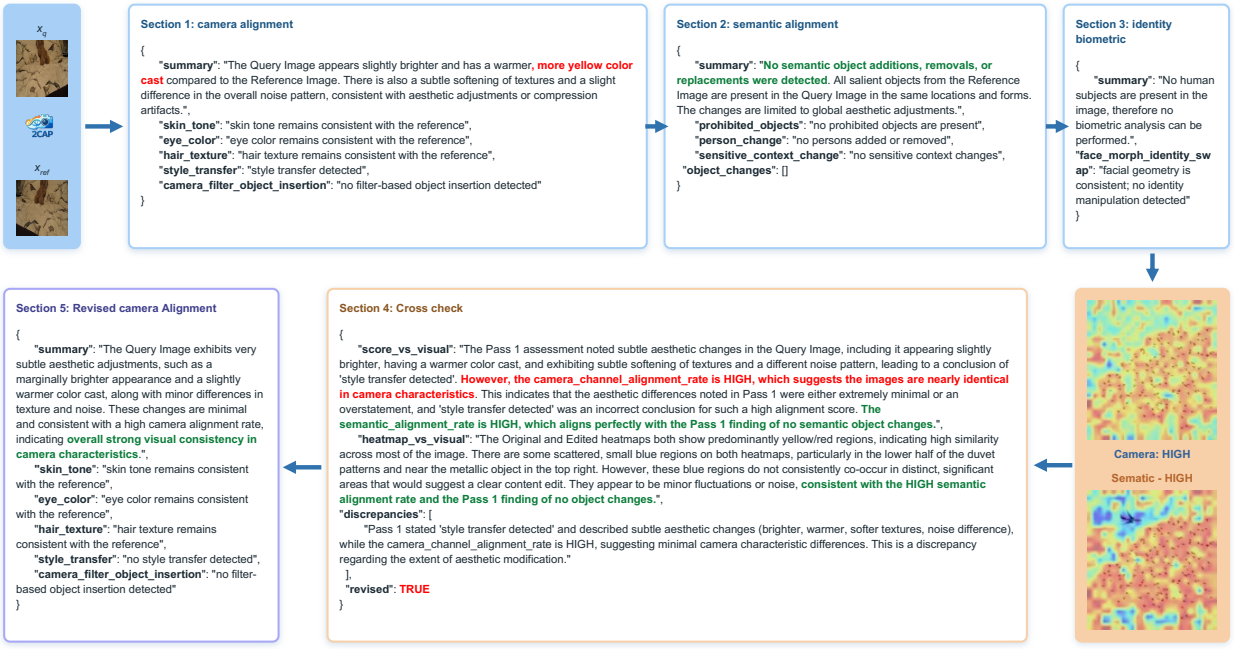


**Figure 10: Stage 3 pipeline example: tampered image.** The VLM identifies localized semantic edits guided by channel scores and saliency heatmaps, producing object-level change descriptions with bounding boxes.

- **C2PA/Content Credentials:** Cryptographic provenance, not learned forensic analysis. Requires adoption at capture time and does not work for legacy content.
- **Perceptual hashing (pHash) [60]:** Detects near-duplicates but cannot characterize *what* changed or distinguish retouching from tampering.
- **Image-pair similarity (LPIPS [61], DISTS [62]):** Measures perceptual distance but provides no forensic decomposition into change type, affected objects, or spatial localization.

A quantitative comparison with these reference-based alternatives requires the same evaluation pairs, a specified multi-class decision rule, and thresholds selected on a separate validation split. We leave this matched comparison to further evaluation.

# E Additional Qualitative Results and User Study

## E.1 Extended Qualitative Examples

Figures 10-14 present additional qualitative results beyond the main paper, covering tampered, retouched, and fully synthetic categories. Each example shows the query-reference pair, the two-pass OGR pipeline output (Pass 1 visual-only assessment, cross-check, and Pass 2 revision), and the final forensic conclusion. These examples demonstrate how

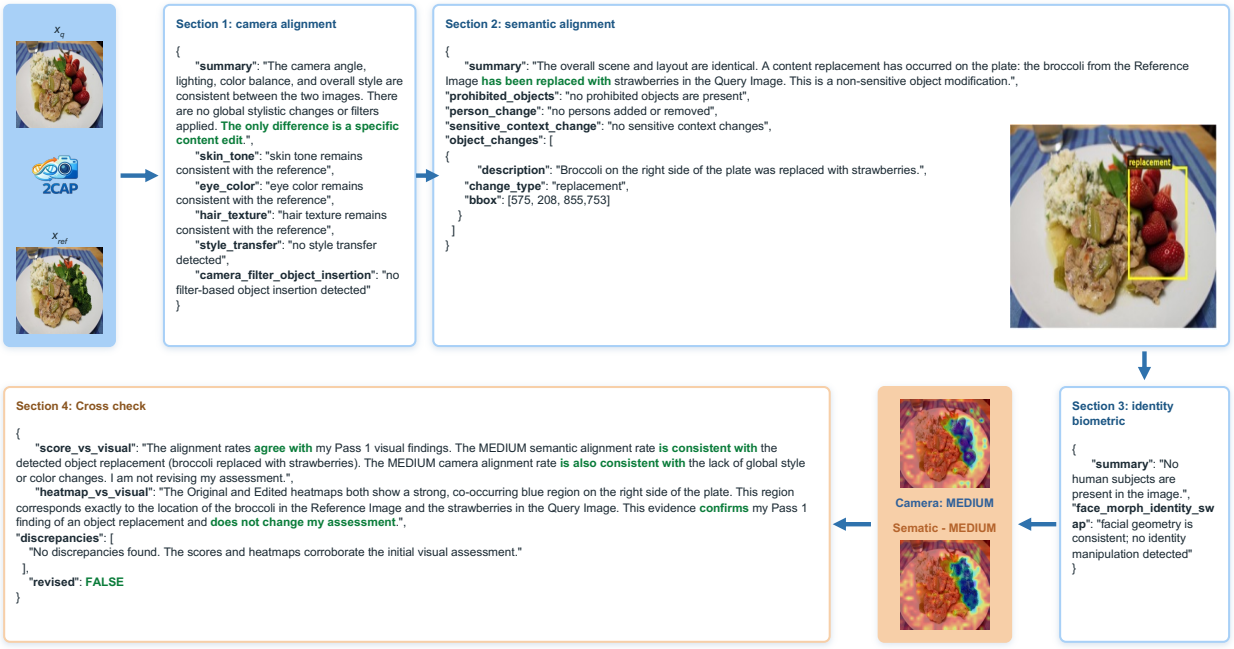


**Figure 11: Stage 3 pipeline example: retouched image.** Camera-channel misalignment triggers the VLM to focus on global processing artifacts rather than semantic edits, correctly distinguishing retouching from tampering.


**Figure 12: Stage 3 pipeline example: fully synthetic image.** Both channels show low alignment, and the VLM identifies the absence of camera traces alongside semantic inconsistencies.

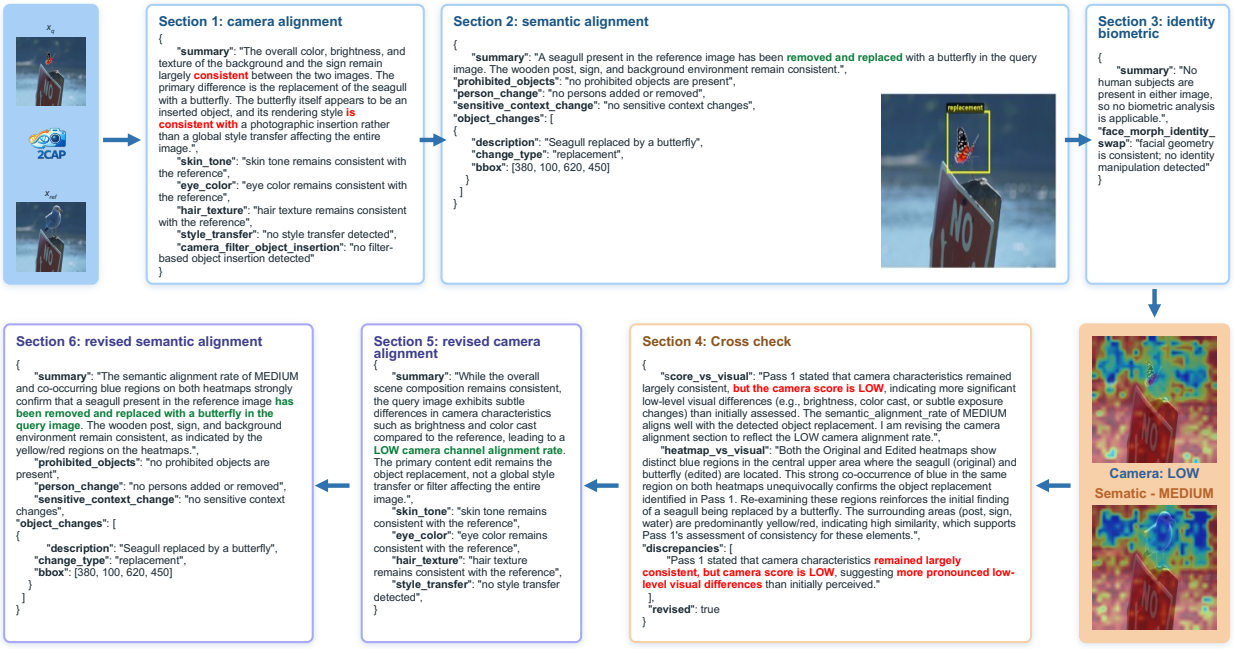


**Figure 13: Stage 3 pipeline example: fully synthetic image.** Both channels show low alignment, and the VLM identifies the absence of camera traces alongside semantic inconsistencies.


**Figure 14: Stage 3 pipeline example: fully synthetic image.** Both channels show low alignment, and the VLM identifies the absence of camera traces alongside semantic inconsistencies.

the dual-channel scores and saliency heatmaps guide the VLM through the Observe-Generate-Refine loop, producing structured evidence that decomposes *what* changed, *where*, and *how*. We include both clear-cut cases where the dual-channel signatures are distinctive and borderline cases where scores are ambiguous, illustrating the range of 2CAP's behavior across diverse editing scenarios.

### E.2 Failure Case Analysis

Understanding failure modes is critical for assessing deployment readiness. We present three representative failure cases in Figure 15, each selected because it represents a genuinely difficult sample that challenges not only 2CAP but also human annotators. These edge cases reveal the boundaries of the current dual-channel framework and inform directions for future improvement.

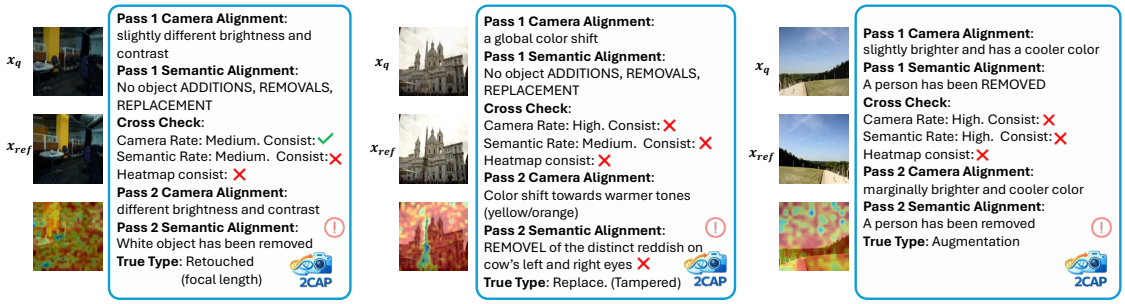


**Figure 15: Failure cases of 2CAP.** Three challenging examples where the two-pass pipeline produces incorrect conclusions. Left: retouching misclassified due to VLM hallucination under ambiguous saliency. Middle: subtle tampering missed in Pass 1, with Pass 2 producing an imprecise description. Right: augmentation misclassified as tampering due to VLM hallucinating a person removal.

**Case 1: Retouching with hallucinated object change (left).** The query image is a focal-length retouching of the reference: a global camera-parameter edit that preserves all semantic content. This is an inherently difficult sample because the focal-length change introduces perspective cropping that creates apparent object disappearances at image borders. Pass 1 correctly identifies brightness and contrast differences with no object changes. However, the cross-check flags heatmap inconsistency, and Pass 2 hallucinates that a "white object has been removed," likely triggered by border-region saliency artifacts from the perspective shift. The result is a misclassification from retouched to tampered.

**Case 2: Subtle tampering with imprecise description (middle).** The query contains a localized replacement edit (ground truth: tampered), but the modification is small and blends seamlessly with the surrounding context: a cow's eye region with subtle color changes. Pass 1 detects only a global color shift and reports no object changes, missing the tampering entirely. The cross-check correctly flags inconsistencies across all three signals, triggering Pass 2 revision. While Pass 2 identifies a change in the cow's eye region, the description ("removal of reddish on cow's eyes") is imprecise–the actual edit is a replacement, not a removal. This case illustrates the difficulty of detecting and accurately describing small, well-blended tampered regions where the semantic mismatch is below the SigLIP2 patch resolution ($27 \times 27$).

**Case 3: Augmentation with hallucinated person removal (right).** The query is an authentic augmentation (camera-parameter restyling) with no semantic changes, yet Pass 1 confidently reports "a person has been REMOVED." The cross-check flags inconsistencies, but Pass 2 reinforces the hallucination rather than correcting it. This failure arises because the VLM's instruction prompt implicitly biases toward finding edits; when saliency maps show diffuse activity (common in augmentations with global brightness/color shifts), the VLM may anchor on spurious visual differences and fabricate a plausible edit narrative.

**Discussion and mitigations.** These failure cases share a common theme: they are *adversarially difficult* samples that lie near the decision boundaries between modification categories. Two key observations emerge:

1. **VLM hallucination under weak saliency** (Cases 1 and 3): When both channels produce ambiguous scores, saliency maps are diffuse, providing the VLM with insufficient spatial grounding. The +OGR cross-check partially

mitigates this–it correctly detects inconsistencies in all three cases–but the current Pass 2 revision cannot fully suppress hallucination when the VLM has high confidence in its fabricated description. One potential mitigation is training a lightweight hallucination detector that cross-references VLM claims against the quantitative channel scores. We leave these extensions to future work.

2. **Sub-patch-resolution edits** (Case 2): Tampering that falls below the SigLIP2 patch granularity ($27 \times 27$ grid, $\approx 16 \times 16$ pixel patches) produces minimal semantic-channel signal, forcing reliance on the camera channel alone. A promising future direction is to incorporate multi-scale patch features or higher-resolution saliency maps to capture fine-grained edits.

These examples characterize possible failure mechanisms. Aggregate improvements on the 400-pair evaluation do not establish their prevalence in deployment or the frequency of hallucination; that requires explicit failure annotation and measurement.

### E.3 User Study: Forensic Explanation Quality

This section provides full details of the user study summarized in Sec. 4.3. We assess whether score-contextualized +OGR explanations are preferred over baseline (score-agnostic) explanations by human evaluators.

**Setup.** We randomly sample 10% image pairs spanning multiple modification categories and present each pair to 5 evaluators alongside two anonymized forensic explanations (A/B): one from the baseline VLM (no alignment scores or heatmaps) and one from the +OGR pipeline (with dual-channel scores and saliency maps). The A/B assignment is randomized per pair. Evaluators are asked: *"Which explanation is more helpful for determining what changed and whether it matters?"* (A / B / Tie).

**Results.** The reported study found a majority preference for +OGR on every sampled pair, with at least 60% of evaluators choosing it on each pair. This is pair-level majority agreement, not unanimous agreement among evaluators. Evaluators highlighted several qualitative advantages of the +OGR explanations:

- **Greater specificity:** +OGR explanations provided more granular descriptions, identifying specific entities (e.g., biometric details, object boundaries) rather than generic summaries.
- **Saliency grounding:** The inclusion of saliency analysis helped evaluators verify *where* changes occurred, a capability absent in baseline explanations.
- **Fewer misidentifications:** Baseline explanations exhibited reliability issues such as confusing object types (e.g., mistaking a fence for a barrier), while +OGR descriptions were more consistent with the visual evidence.

Evaluators also noted minor limitations of +OGR: saliency maps were occasionally sparse for subtle edits, and background environmental features were sometimes overlooked. These observations are consistent with the failure modes identified in Sec. E.2 and suggest that higher-resolution saliency and improved VLM grounding remain valuable directions for future work.

## F Broad Impact and Limitations

**Policy-dependent decisions.** 2CAP provides channel similarity scores, localized evidence, and natural-language explanations to support policy-dependent human review. Different stakeholders, such as newsroom editors, scientific reviewers, legal professionals, platform moderators, may apply different thresholds and rubrics. 2CAP is designed to produce structured evidence without assuming a universal definition of disallowed manipulation; this policy-agnosticism is a deliberate design choice.

**Adversarial robustness.** Sophisticated adversaries may craft modifications that fool both channels simultaneously. While we observe robustness to standard perturbations and common post-processing pipelines, dedicated adversarial attacks that preserve apparent camera traces while introducing semantic modifications remain an open challenge for all forensic systems.

**Score context influence.** While the VLM can in principle disagree with the channel scores, we observe that score context exerts a strong anchoring effect: when scores are misleading (*e.g.*, a sophisticated edit that preserves camera traces), the VLM may under-weight its own visual observations. Presenting scores as context rather than hard labels mitigates but does not fully eliminate this anchoring bias. Developing score-aware calibration strategies that balance VLM autonomy with quantitative guidance is an important direction.

**Semantic encoder scope.** Our evidence-generation experiments use frozen SigLIP2 for both the semantic score and patch-level saliency map. The paired F−N contrast tests both channel scores and the heatmap jointly. It does not isolate semantic evidence or measure sensitivity to the choice of semantic encoder, including cases where highly

realistic synthetic or edited images confuse the encoder. Comparing multiple frozen semantic encoders and adversarially stress-testing their semantic matches are important next steps.

**VLM reliability.** Generated explanations may occasionally be inaccurate or hallucinate forensic details not present in the image. The structured evidence trail (dual-channel scores, saliency maps) provides independent verification beyond the text, and human-in-the-loop verification remains essential for high-stakes decisions.