Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Coverage-Aware Partial Transport for Region-Level Attribution of Reused News Images

Taoqian Lin

lintaoqian@126.com

Qingdao Agricultural University

Meijing He

Qingdao Agricultural University

Research Article

Keywords: news image reuse, region attribution, image provenance, partial optimal transport, local feature matching

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11766225/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Coverage-Aware Partial Transport for Region-Level Attribution of Reused News Images

Taoqian Lin[1*] and Meijing He[1]

[1*]School of Animation and Media, Qingdao Agricultural University, 700 Changcheng Road, Qingdao, 266109, Shandong, China.

*Corresponding author(s). E-mail(s): lintaoqian@126.com;
Contributing authors: 734716461@qq.com;

Abstract

News-image reuse is often local: a query can crop one historical photograph, combine regions from two photographs, or contain areas with no counterpart in an available gallery. Whole-image retrieval and independent pairwise matching do not directly resolve which source explains each query region because several candidates can claim the same pixels. We formulate this task as closed-gallery region-to-source attribution and introduce a joint partial-transport method. Frozen patch descriptors define appearance costs, affine source-to-query transforms supply geometric residuals and projected-area capacities, and a shared transport budget makes candidate sources compete with an explicit null destination. Candidate subsets are selected with an entropy-free evidence score so regularization smooths each transport plan without favoring subsets with more columns. On a capture-cluster-separated benchmark derived from 1,663 downloadable images in fact-checking records, the method reaches 0.772 mask mean intersection over union (77.23%) across 1,170 controlled queries. It improves this metric by 0.087 over LightGlue, 0.106 over RootSIFT, and 0.100 over an independent partial-transport control; **95%** paired cluster-bootstrap intervals exclude zero. The control produces overlapping source claims on **37.5%** of query patches before conflict resolution. The joint system has lower exact-source-set accuracy than LightGlue and RootSIFT, showing localization and source-set recovery are distinct objectives. Null mass ranks missing-source regions well under one-parent omission, with 0.920 area under the receiver operating characteristic curve, but remains imperfect when all parents are omitted. These results establish joint transport as an effective mechanism for controlled, closed-gallery localization of reused news-image regions.

Keywords: news image reuse, region attribution, image provenance, partial optimal transport, local feature matching

1 Introduction

News images are frequently reused through cropping, recompression, layout overlays, and composition. For a fact checker, recovering a visually related image is only the beginning: the useful question is which part of the query is explained by which earlier image. A whole-image neighbor cannot express two visual parents, while a collection of independent local matches can assign several candidates to the same region. Region-level attribution therefore requires sources to compete for finite query area and must leave unsupported regions unexplained.

This spatial question is increasingly relevant because evidence in a news image is rarely presented in the form in which it was first captured. Publishers crop photographs to fit a layout, place text or logos over salient regions, and combine visual material from different sources. Social posts can further recompress or screenshot those layouts. A retrieval system may still return the correct photograph, but an analyst cannot tell from an image-level score whether the match explains the full query, only a crop, or one side of a composite. The same ambiguity appears when several near-duplicate gallery items are individually plausible. A useful attribution system must therefore expose both positive support and unassigned area rather than force a single global identity.

Multimodal fact-checking surveys distinguish evidence retrieval and cross-modal consistency from final factuality assessment [1, 2]. We adopt that separation at a finer spatial scale. The proposed output is an auditable visual intermediate: a source identity or null label for each query patch. A reviewer can inspect the selected images, the estimated warps, and the resulting masks before using any contextual or chronological evidence. This design is especially important for news use, where visual relatedness alone cannot establish whether a caption is accurate or which outlet first published the photograph.

Research on out-of-context imagery has concentrated on image-caption consistency, binary misuse detection, and recovery of an image's original context [3–7]. These tasks provide evidence about whether a news use is consistent with an image, but their principal output is not a spatial map from query regions to external visual references. Image-provenance systems and local feature matchers address a related problem by retrieving transformed copies and estimating pairwise correspondences [8–10]. Pairwise evidence is strong when a single source dominates. It becomes ambiguous when the number of contributing sources is unknown or when the gallery covers only part of the query.

We cast closed-gallery region attribution as a joint allocation problem. Each query patch owns a fixed amount of mass. Shortlisted gallery images propose source patches, whose costs combine frozen visual similarity with source-to-query geometric consistency. Real source columns have upper capacities derived from their projected support, and every query patch can instead send mass to a null column. We solve a regularized partial-transport problem for each admissible active subset and compare subsets with an entropy-free evidence energy. This separation is important: entropy stabilizes a fixed solve, but it must not reward a subset merely because that subset contains more transport columns. A source-count penalty and a minimum mass requirement for two-source subsets control the inferred cardinality.

Figure 1 summarizes the inference path. Global and local evidence first produce a small candidate set; affine verification then supplies geometry for both costs and capacities. The transport layer is the only stage that sees all active sources together. Consequently, one patch cannot be fully claimed by two candidates, and a poor match can be routed to null without invalidating the remaining source. This differs from independently scoring candidates and resolving conflicts after the fact, because competition occurs while the source plans are formed.

The method is evaluated on controlled derivatives of downloadable images from 5Pils fact-checking records [6]. The protocol first clusters likely same-capture gallery items, splits complete clusters, and only then creates single-source crops, crops with news-style overlays, and two-source composites. This design provides exact parent masks while preventing derivatives of an accepted same-capture pair from crossing protocol splits. The resulting test contains 1,170 queries against a fixed gallery of 1,663 images.

The controlled design is intentional. The released 5Pils evidence URLs are retrieval candidates rather than pixel-level parent annotations, so treating them as region ground truth would confound provenance with search-engine output. By generating derivatives only after a capture-cluster-aware split, we know the visual parents and masks exactly while retaining real photographs drawn from fact-checking records. The resulting claim is correspondingly narrow: the experiments test controlled, closed-gallery localization, not natural open-Web provenance reconstruction.

Our contributions are:

- We define closed-gallery region-to-source attribution as an unknown-cardinality patch assignment problem with a null label, separating spatial source localization from whole-image context prediction and source-set retrieval.
- We introduce a joint partial-transport formulation that combines frozen appearance evidence, source-to-query geometry, projected-support capacities, capture-cluster exclusivity, and entropy-free subset comparison.
- We provide a capture-cluster-aware controlled evaluation against global retrieval, patch similarity, MNN-RANSAC, a same-cost independent transport control, RootSIFT, and LightGlue, together with paired uncertainty, mechanism ablations, omission stress tests, and explicit trade-offs.

Joint transport obtains 0.772 mask mIoU. Its paired improvement is 0.100 over independent transport, 0.087 over LightGlue, and 0.106 over RootSIFT under bootstrap resampling of 228 primary parent capture clusters. RootSIFT and LightGlue remain better at exact-set recovery, and the omission experiments distinguish useful continuous null ranking from weaker hard abstention. The evidence thus identifies the setting in which the joint formulation helps: spatial attribution within a fixed candidate gallery.

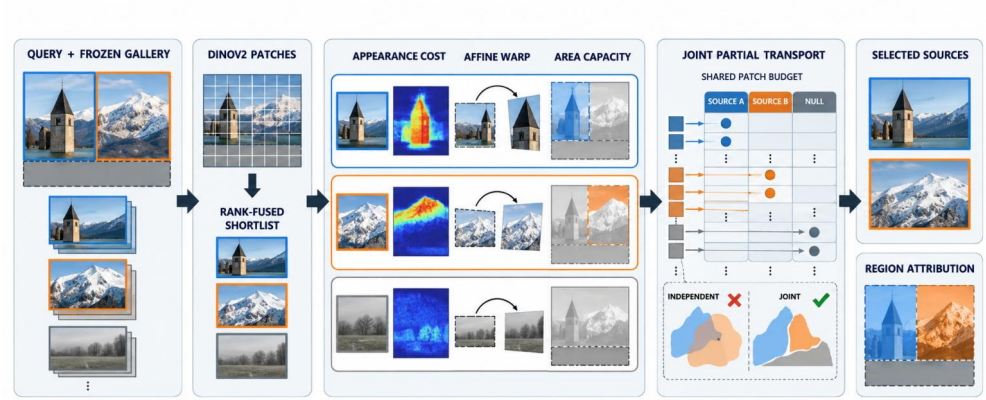


Fig. 1 Overview of the closed-gallery region attribution pipeline. Frozen query and gallery descriptors produce a rank-fused shortlist. Appearance costs, source-to-query affine warps, and projected-area capacities then enter one joint partial-transport problem with a shared query-patch budget and an explicit null destination. The comparison inset highlights the central distinction: independent candidate solves can overlap, whereas joint allocation produces mutually exclusive source and null regions.

2 Related Work

2.1 News-image context and multimodal fact checking

Multimodal fact checking covers several tasks that are easy to conflate: determining whether an image-text pair is mutually consistent, retrieving external evidence, contextualizing an image, detecting digital manipulation, and deciding a claim's veracity. Recent surveys organize these tasks by the role of the visual evidence and document persistent problems with data access, annotation provenance, and leakage [1, 2]. Our task occupies a narrower position in this taxonomy. It asks which available visual source explains each query region. It does not infer whether the associated news claim is true, which publication came first, or whether the pixels are camera-authentic.

Early image-centered work made this separation visible. Fauxtography evaluates the relationship between an image and a textual claim, including cases in which genuine photographs accompany false descriptions [11]. Cross-modal entity consistency measures whether people, places, and events named in news text agree with the accompanying photograph [12]. NewsCLIPpings constructs hard image-caption mismatches while preserving plausible visual semantics [3], and COSMOS uses self-supervised visual representations and object-aware reasoning to detect out-of-context reuse [5]. These methods address contextual agreement; none defines a multi-source spatial assignment over a gallery.

Evidence-oriented systems extend the problem beyond the posted pair. Open-domain multimodal fact checking retrieves online visual and textual evidence [4], SNIFFER generates evidence-aware explanations with a multimodal language model [7], and 5Pils asks a system to reconstruct the original meta-context of fact-checked

images [6]. M4FC broadens real-world coverage across languages, cultures, and several intermediate verification tasks [13]. This progression makes retrieval and evidence traceability increasingly important, but a retrieved item is still an image-level object. A region map is needed when a query contains a crop from one source, a second visual parent, and graphics or borders unsupported by either.

The distinction between visual relation and truth assessment is also methodological. Similarity over Factuality shows that a benchmark can reward image similarity even when its stated objective is factual verification [14]. We therefore report a deliberately visual claim: under a fixed closed gallery and a controlled generator, the method localizes which source explains each patch. This task can support a fact checker by exposing reusable evidence, but its output is not a fact-checking verdict.

2.2 Image provenance, copy detection, and retrieval

Image provenance systems retrieve related images and organize pairwise transformations into a graph [8]. Image repurposing detection similarly learns representations that remain informative under contextual or adversarial changes [15]. Both lines rely on the ability to recognize an instance despite cropping, resizing, recompression, and partial occlusion. Our work shares that visual foundation but changes the output unit. A provenance edge states that two images are related; region attribution states which query patches are explained by each relation when several edges are simultaneously plausible.

Large-scale retrieval usually separates inexpensive global ranking from more expensive local verification. DELF learns attentive local features for retrieval and geometric verification [16], while DOLG fuses local and global information into one compact representation [17]. Recent list-wise re-ranking processes interactions among the query and an entire shortlist at the local-descriptor level [18]. These systems improve candidate ordering. Our rank-fused shortlist serves the same computational purpose, but shortlist ranking is not the final decision: retained candidates must still compete for finite query area.

Classical SIFT correspondence remains a strong copy-detection primitive because it tolerates scale and viewpoint changes [19]; RootSIFT improves its descriptor normalization for retrieval [9], and RANSAC converts noisy matches into a geometric hypothesis [20]. Learned local features and matchers progressively replace stages of this pipeline. D2-Net jointly detects and describes local features [21]; SuperGlue performs context-aware assignment between detected features [22]; LoFTR directly predicts detector-free dense correspondences [23]; and LightGlue adaptively reduces matching computation [10]. Efficient LoFTR improves the speed-accuracy balance for semi-dense matching [24]. RoMa combines frozen foundation features with specialized fine features for robust dense correspondence [25], while MATCHA targets geometric, semantic, and temporal correspondence within a unified representation [26]. These advances strengthen pairwise evidence, but they do not by themselves resolve conflicts between multiple candidate sources.

Our evaluation therefore retains pairwise geometry and includes RootSIFT and LightGlue as strong baselines. The methodological difference begins after matching.

Running a matcher independently for every candidate can leave two candidates claiming the same region, and selecting the pair with the highest image-level scores does not specify how pixels should be divided. We make candidate images share one query-patch budget and evaluate an independent partial-transport control that uses the same appearance, geometry, capacity, and null constructions before post-hoc conflict resolution.

2.3 Frozen visual features and partial transport

DINOv2 supplies transferable global and dense visual descriptors without task-specific supervision [27]. Freezing this encoder serves two purposes in our setting. It prevents the controlled query generator from becoming a training shortcut, and it makes the comparison focus on inference structure rather than representation learning. The resulting patches are coarser than keypoint-level correspondences, so affine consistency and source-capacity constraints remain necessary to separate locally similar candidates.

Entropic optimal transport provides an efficient framework for soft assignment through Sinkhorn scaling [28]. Iterative Bregman projections extend this view to intersections of convex constraints [29], and partial optimal transport formalizes matching when not all mass should be paired [30]. SuperGlue's learned dustbin is a related mechanism for unmatched features [22]. The important design choice here is the scope of the transport problem. Pairwise partial matching optimizes one query-source relation at a time, so separately solved candidates never share query mass. Our real columns instead combine patches from a jointly selected source subset, while one null column absorbs unsupported area. Entropy smooths a fixed feasible plan; an entropy-free evidence energy compares source subsets so that adding columns does not receive an artificial regularization advantage.

This formulation connects retrieval, correspondence, and partial assignment without claiming to replace them. Retrieval defines which candidates enter the decision, geometry determines whether their local evidence is spatially coherent, and joint transport converts those candidate-wise signals into a mutually competitive region map. The experiments separately evaluate each responsibility through shortlist recall, geometric ablation, capacity ablation, an independent-transport control, and missing-source stress tests.

3 Joint Region Attribution

3.1 Problem formulation and shortlist

Let a query image be represented by N unit-normalized patch descriptors $Q = \{q_i\}_{i=1}^{N}$ and let gallery image j contain K_j descriptors $G_j = \{g_{jk}\}_{k=1}^{K_j}$. The task returns a label $y_i \in S \cup \{\emptyset\}$ for each query patch, where S is an inferred subset of a fixed gallery and $\emptyset$ denotes area unsupported by an active source. We allow at most $L = 2$ source capture clusters, matching the single- and two-source scope of the controlled protocol.

All images use a frozen DINOv2-small encoder [27]. We resize inputs to 224×224, average-pool the native 16×16 patch grid to 8×8, and set $N = K_j = 64$. The gallery

shortlist is the rank-fused union of 15 global-descriptor neighbors and five local-vote neighbors, capped at 20 unique items. Shortlisting is fixed before regional inference and reaches 98.12% all-parent recall on the test queries.

The two retrieval channels cover complementary failure modes. Global descriptors efficiently identify images with similar overall content, but a small reused crop may contribute too little to dominate a whole-image embedding. Local votes can promote a candidate supported by only part of the query, although repetitive structures may also create spurious votes. Rank fusion avoids calibrating the raw scales of these two scores. The downstream method receives only the fused identities and their cached patch descriptors; it does not fine-tune the encoder or revise the shortlist.

The closed-gallery assumption means that every predicted real source must be one of these candidates. It does not mean that every query patch has a source. Borders, captions, overlays, and a missing visual parent are represented by $\emptyset$. This distinction motivates a row-conserving formulation: each query patch must allocate all of its mass, but the allocation may terminate at a real source patch or at null.

3.2 Appearance and source-to-query geometry

The appearance dissimilarity between query patch i and source patch (j, k) is

$$d^{\text{app}}_{i,j,k} = 1 - q_i^\top g_{jk}. \tag{1}$$

Mutual-nearest dense matches estimate an affine transform H_j with RANSAC [20]. The transform maps normalized source coordinates z_{jk} into query coordinates x_i. For a valid estimate, the truncated geometric residual is

$$d^{\text{geo}}_{i,j,k} = \min\left(\frac{\|x_i - H_j z_{jk}\|_2^2}{\sigma_g^2}, c_g\right), \tag{2}$$

and an invalid estimate receives $d^{\text{geo}}_{i,j,k} = c_g$. The combined cost is

$$C_{i,j,k} = d^{\text{app}}_{i,j,k} + \lambda_g d^{\text{geo}}_{i,j,k}. \tag{3}$$

Geometry also defines an upper capacity. Let A_{jk} be source grid cell (j, k) and Ω_q the unit query domain. We compute

$$b_{jk} = \frac{\text{area}(H_j(A_{jk}) \cap \Omega_q)}{\text{area}(\Omega_q)} + \delta_{jk}, \tag{4}$$

where δ_{jk} is a small numerical slack for visible cells. Polygon clipping evaluates the intersection. Invalid transforms use the bounded fallback $b_{jk} = \kappa/N$. This capacity prevents a source cell from explaining more query mass than its projected support; Section 5 tests its empirical trade-off against uniform capacity.

The transform direction is source to query. This direction matters for cropped reuse: a cell that lies outside the visible transformed source receives no geometric support in the query, whereas a visible cell receives capacity in proportion to its

projected area. The same transform also lets the cost reward a query patch that lies near the warped source coordinate. Thus geometry affects where mass is inexpensive and how much mass a source cell may accept. The two effects are ablated separately through the no-geometry and uniform-capacity variants.

Affine estimation can be unreliable for weakly textured photographs, repeated patterns, or very small retained regions. Rather than discard such a candidate and risk turning a retrieval failure into a hard false negative, the implementation assigns a truncated geometric cost and a bounded fallback capacity. Appearance evidence can still support the candidate, but it receives no benefit from a precise warp. This behavior keeps the candidate in the same optimization problem and makes geometry a graded source of evidence.

3.3 Partial transport for a fixed subset

Each query patch has mass $a_i = 1/N$. For an active subset S, a plan $P_{i,j,k} \geq 0$ sends query mass to real source patches, and $p_{i,0} \geq 0$ sends mass to null. The feasible set is

$$\mathcal{F}_S = \{(P, p_0) : \sum_{j \in S} \sum_k P_{i,j,k} + p_{i,0} = a_i, \quad \sum_i P_{i,j,k} \leq b_{jk}, \quad P, p_0 \geq 0\}. \tag{5}$$

The null column makes every subset feasible, including $S = \emptyset$.

For nonempty S, a reference measure reserves ρ_0 mass for null and divides the remaining mass equally among sources and their columns: $r_{jk} = (1 - \rho_0)/(\mathrm{card}(S)K_j)$ and $r_0 = \rho_0$. With $R_{i,j,k} = a_i r_{jk}$ and $R_{i,0} = a_i r_0$, the regularized plan is

$$(P_S^\epsilon, p_{S0}^\epsilon) = \arg \min_{(P,p_0) \in \mathcal{F}_S} \{\langle P, C \rangle + \gamma_0 \|p_0\|_1 + \epsilon \mathrm{KL}([P, p_0] \| R_S)\}. \tag{6}$$

Positive reference mass and the strictly convex KL term give a unique plan on the feasible support.

The shared row equality in Eq. (5) is the central coupling. Suppose two candidates both match the same building facade. Independent solvers may each send the full row mass to their own source patches; a later tie-breaking rule can choose one label, but neither plan was formed in the presence of the other. In the joint problem, the candidates and null collectively receive exactly a_i. Increasing the mass assigned to one explanation immediately reduces the mass available to the others. The plan therefore represents competition rather than a collection of independently normalized confidences.

The null column has no finite upper capacity. Its linear cost γ_0 sets the price of leaving a patch unexplained, while its reference share ρ_0 controls only the regularized smoothing target. A low null cost would reject weak but valid sources; a high cost would force unsupported overlays into a real source. We freeze both quantities on validation data and evaluate continuous null mass separately from the hard source-or-null label.

We solve Eq. (6) with generalized Sinkhorn dual scaling [28, 29]. For the Gibbs kernel obtained by multiplying the reference measure by the elementwise exponential

of the negative augmented cost, the updates are

$$u \leftarrow a \otimes (K_S v), \qquad v_r \leftarrow \min\{1, b \otimes (K_S^\top u)_r\}, \qquad v_0 = 1, \tag{7}$$

where r indexes the real source columns and b is the capacity vector over these columns. The implementation requires row residual, capacity violation, and relative objective change to remain below their tolerances for ten iterations. A nonconverged subset is retried with a larger iteration budget and otherwise raises an error.

These checks are important because subset energies are comparable only when the corresponding inner problems are solved to similar accuracy. The final test produced no nonconverged subset. For a fixed subset, the dominant storage is the cost kernel between N query patches and the concatenated source patches plus null. With $L = 2$ and 64 patches per image, this kernel remains small; most runtime comes from evaluating several admissible singleton and pair subsets rather than from one transport projection.

3.4 Entropy-free subset selection and labels

Regularization should smooth a fixed plan without determining how many sources are selected. We therefore compare subsets with the entropy-free evidence energy

$$E(S) = \langle P_S^\epsilon, C \rangle + \gamma_0 \|p_{S0}^\epsilon\|_1 + \lambda_s \mathrm{card}(S). \tag{8}$$

The admissible family contains the empty set, all singletons, and cross-capture-cluster pairs formed from a validation-frozen beam of the six strongest singletons. Two items from the same capture cluster cannot be active together. Each member of a two-source subset must receive at least τ transported mass. We select the subset that minimizes $E(S)$.

Capture-cluster exclusivity prevents duplicate gallery items from doubling the representational capacity of one photograph. We retain duplicate items as separate singleton hypotheses because the exact gallery item can still matter for attribution, but two members of the same accepted capture cluster cannot be activated together. The beam reduces the quadratic pair search after singleton screening. The minimum-mass gate removes nominal two-source solutions in which one candidate receives only a negligible residual assignment.

The empty set is evaluated with all mass at null and provides a common reference for source selection. For nonempty subsets, Eq. (8) deliberately omits the KL term used in Eq. (6). Including entropy in subset comparison can favor larger subsets simply because they expose more reference columns, even when their appearance and null evidence are unchanged. We use entropy for numerical smoothing within a subset and use transport cost, null cost, and an explicit source-count penalty for evidence comparison across subsets.

For each selected source j, source mass at query patch i is the source mass $m(i, j)$, obtained by summing the selected transport columns over k. The patch receives the

source with largest m_{ij} when this mass exceeds p_{i0}; otherwise it receives null. The continuous null score is Np_{i0} clipped to $[0, 1]$. This score is evaluated separately from the hard label because ranking and fixed-threshold abstention answer different questions.

The final source set contains only labels that survive patch decoding. This prevents a numerically active source with no winning query region from appearing in the reported set. Region masks are obtained by reshaping the 8×8 labels to the query grid; evaluation occurs at this same patch resolution. We do not post-process the masks with a learned segmenter, so differences among methods reflect correspondence and allocation rather than boundary refinement.

4 Experimental Protocol

4.1 Data and controlled queries

We downloaded 1,663 valid images from the 1,676 released 5Pils records [6]; 13 URLs returned HTTP 404. The dataset field named `SOURCE` denotes a textual record source rather than the identity of a visual parent, and released reverse-search URLs are candidate evidence rather than region-level ground truth. We therefore use the decoded images as known parents for a deterministic controlled protocol.

This choice avoids converting retrieval metadata into labels that it was not designed to provide. An evidence URL may contain a related screenshot, a near duplicate, a repost, or an image that helps recover context without being the exact pixel source. Conversely, absence from the released evidence list does not establish visual irrelevance. The controlled protocol uses only the downloaded photograph itself as a parent and records every crop window, resize, overlay, seam, compression parameter, and mask label in a query manifest. Parentage is therefore defined by the generator rather than inferred from filename, Web domain, or semantic similarity.

The protocol first assigns source images to train, validation, and test and then generates derivatives within each split. Every source produces four single-source queries with retained areas 0.25, 0.50, 0.75, and 1.00. Half are crop-and-resize variants; half add a border and a news-style overlay whose pixels are labeled null. Each source also produces one hard-seam composite with a distinct within-split source, balanced between horizontal and vertical seams. JPEG quality is sampled from $[45, 65, 85, 95]$. The resulting 8,315 queries comprise 5,835 training, 1,310 validation, and 1,170 test examples. The test set contains 936 single-source and 234 two-source queries.

The retained-area schedule creates a controlled range from small partial reuse to an uncropped parent. Crop-and-resize queries test whether a source can be recovered when the visible region changes scale. Overlay queries add genuinely unsupported pixels without changing the identity of the underlying photograph. Two-source composites test competition directly because the ground-truth parents occupy disjoint areas in the same query. Hard seams are simpler than natural collage boundaries, but they make the source masks exact and prevent an image-generation model from introducing ambiguous blended pixels.

All random choices are deterministic under the released protocol seed. The gallery is fixed independently of each query and contains the original downloaded images, not query derivatives. Training derivatives are retained for possible future representation

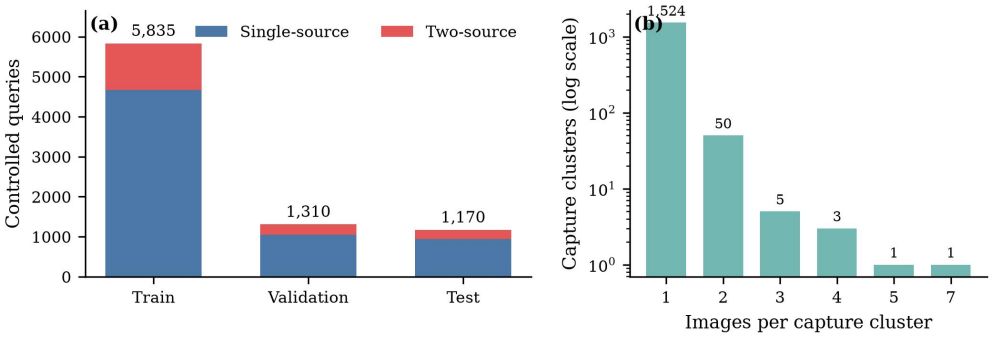


Fig. 2 Composition of the controlled benchmark. (a) Single-source and two-source query counts after capture-cluster-separated assignment; totals above the bars are 5,835 training, 1,310 validation, and 1,170 test queries. (b) Distribution of the 1,663 downloaded images over 1,584 capture clusters. The logarithmic axis exposes the small number of multi-image clusters without hiding the 1,524 singleton clusters.

learning, but the present method keeps the encoder frozen and uses only validation queries for hyperparameter selection. Test queries are processed once with the frozen configuration.

Figure 2 summarizes the evaluation population rather than method performance. The split is dominated by single-source derivatives because each parent contributes four such queries and one composite. Most capture clusters are singletons, while the multi-image tail motivates assigning complete connected components, rather than individual files, to protocol splits.

4.2 Capture-cluster separation

A preliminary perceptual-hash split was rejected after a frozen DINOv2 screen revealed crop and screenshot variants across splits. All measurements from that split were quarantined. For the corrected protocol, the 30 nearest DINOv2 neighbors above cosine 0.70 proposed 427 possible same-capture pairs. RootSIFT correspondence and affine RANSAC accepted an edge with at least eight inliers and an inlier ratio of at least 0.25. The 83 accepted edges formed 1,584 connected components among 1,663 images; 60 components contain multiple images and the largest contains seven. Complete components were assigned to one split before query generation. The automatic audit detected zero cross-split capture clusters and zero missing-file, hash, parent-map, or mask-label errors. Independent human confirmation of the 83 accepted edges remains a pre-submission integrity gate.

Splitting connected components rather than individual files addresses two forms of leakage. First, visually equivalent crops or screenshots must not appear as nominally different sources across train and test. Second, bootstrap uncertainty should not treat derivatives of closely related parent files as independent evidence. We therefore preserve the component identifier throughout generation and evaluation. The clustering procedure is conservative rather than exhaustive: it prevents activation of accepted

duplicate candidates and defines the resampling unit, but relations missed by the proposal or acceptance thresholds may remain.

Hyperparameters were chosen on the first 100 validation queries: $\lambda_g = 0.08$, $\sigma_g = 0.12$, $c_g = 4$, $\varepsilon = 0.04$, $\gamma_0 = 0.65$, $\rho_0 = 0.2$, $\lambda_s = 0.09$, and $\tau = 0.12$. The solver uses 1,000 iterations initially and retries up to 20,000. The pair beam contains six singleton candidates.

The validation slice was fixed before confirmatory testing. We selected a single operating point rather than optimizing separate settings for single-source, composite, or omission conditions. This is a demanding choice for the null threshold and source-count penalty, because the settings must balance source parsimony against a legitimate second parent. Sensitivity runs are treated as diagnostics and are not used to select the reported test result.

4.3 Baselines and metrics

The global baseline assigns the whole query to its top DINOv2 neighbor. Geometry-aware MaxSim labels each query patch with its strongest candidate-patch similarity after adding the same geometric cost as Eq. (3); its validation-selected threshold is 0.48. MNN-RANSAC greedily covers query inliers from at most two candidates and requires 0.18 new-query coverage. The independent partial-transport control uses the same shortlist, costs, capacities, null reference, source penalty, and singleton solver as our method. It solves each candidate separately, retains up to two candidates whose penalized singleton energy is below the empty-set energy, and resolves overlapping positive claims by maximum source-over-null mass advantage. Thus it compares coupled multi-source optimization with independent singleton plans followed by heuristic selection and conflict resolution. RootSIFT combines ratio-tested local features with affine verification [9, 19, 20]. LightGlue is used with its released matcher and weights [10]. Every method uses the same gallery and reports at most two sources.

The baselines span four levels of spatial reasoning. Global retrieval has no region model. MaxSim makes patch decisions without enforcing a shared source budget. MNN-RANSAC converts sparse geometric inliers into greedy source coverage. RootSIFT and LightGlue provide strong pairwise correspondences with different handcrafted and learned feature families. Independent partial transport is the closest mechanism control, although its candidate retention and post-hoc conflict rule necessarily differ from joint subset optimization. We therefore interpret it as a test of the complete independent-plan alternative rather than as an isolated intervention on one constraint.

The primary metric is source-mask mean intersection over union (mIoU), averaged over ground-truth source labels. Patch accuracy includes the null label. Exact-set accuracy requires equality between predicted and true source sets. Omission experiments additionally report null AUROC, average precision (AUPR), and hard null IoU. Confidence intervals use 2,000 percentile bootstrap replicates with seed 20261002. The resampling unit is the primary parent capture cluster; all queries assigned to a sampled cluster are retained. This yields 228 test clusters and preserves dependence among derivatives and multi-file clusters.

Table 1 Closed-gallery results on 1,170 controlled test queries. Values are percentages. Exact-set and patch accuracy are computed over all queries; the last three columns report all-query and stratified mask mIoU. Exact set is correct only when the complete predicted source set matches the target. Bold marks the best point estimate in each column.

Method	Exact set ↑	Patch acc. ↑	All mIoU ↑	Single mIoU ↑	Composite mIoU ↑
Global retrieval	72.65	66.78	62.21	72.06	22.84
Geometry-aware MaxSim	21.28	72.77	71.56	69.07	81.51
MNN-RANSAC	80.00	50.08	41.31	33.12	74.07
Independent partial OT	29.91	65.38	67.28	63.74	81.44
RootSIFT	**96.41**	70.96	66.67	67.99	61.37
LightGlue	90.60	72.15	68.53	68.57	68.35
Joint transport	72.56	**76.19**	**77.23**	**75.50**	**84.15**

No single metric summarizes all intended behavior. Mask mIoU measures whether each true source occupies the correct query area and is therefore primary. Patch accuracy additionally rewards correct null assignment but can be influenced by the amount of overlay area. Exact-set accuracy ignores localization once the source identities are correct; it is included to reveal the trade-off between regional allocation and source retrieval. In omission tests, AUROC and AUPR assess the ordering induced by continuous null mass, whereas hard null IoU evaluates the frozen operating point. Reporting both prevents a well-ranked but poorly calibrated null score from being described as reliable abstention.

All feature tensors, shortlists, controlled-query manifests, capture-cluster assignments, solver configurations, and per-query predictions are stored separately from aggregate tables. Figure and table scripts read the frozen JSON result files directly. This organization supports a reproducibility audit in which every plotted value can be traced to a machine-readable result rather than transcribed from manuscript text.

5 Results

5.1 Localization and source-set recovery

Table 1 reports both evaluation objectives. Joint transport reaches 77.23% overall mask mIoU (95% CI 75.59–78.75) and 76.19% patch accuracy. Its paired mIoU improvement is 9.95 points over independent partial OT (95% CI 8.74–11.23), 8.70 over LightGlue (7.31–10.08), 10.56 over RootSIFT (8.68–12.50), and 5.67 over geometry-aware MaxSim (4.42–6.85). The intervals resample capture clusters rather than individual derived queries.

The direct control tests coupled multi-source inference against independent singleton transports with post-hoc selection and conflict resolution. Independent partial OT reaches 67.28% overall mIoU and its selected sources make simultaneous positive claims on 37.48% of query patches before conflict resolution (95% CI 35.37–39.72). On composites, joint transport improves over the control by 2.72 mIoU points (0.63–5.01), while pre-resolution overlap remains 27.02% (24.78–29.38). The result supports

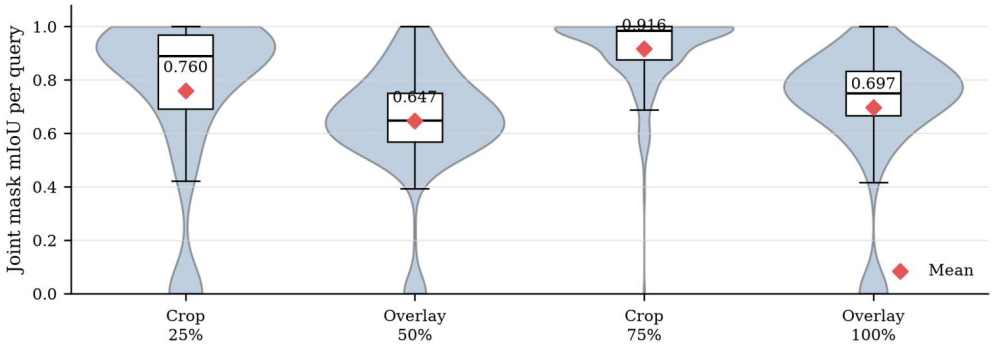


Fig. 3 Per-query joint mask mIoU for the four deterministic single-source generator settings ($n = 234$ per setting). Violins show the complete distributions, boxes show the interquartile ranges and medians, and diamonds mark means. Retained area and edit type are jointly defined by the generator, so the plot describes condition-specific difficulty rather than a causal area-only trend.

joint competition over independent plans plus post-hoc resolution; it does not isolate the row budget from every associated selection difference.

The larger overall gain than composite-only gain is explained by the complete independent pipeline rather than by overlap frequency alone. On single-source queries, independently retained candidates can introduce an unnecessary second explanation before the conflict rule is applied. Joint subset scoring can instead prefer a singleton or the empty set while still solving source and null assignments together. Because candidate retention, minimum-mass gating, and conflict handling are coupled to the chosen inference strategy, the comparison establishes the value of the joint system as implemented but does not assign the full difference to one mathematical constraint.

The result changes with the objective. RootSIFT achieves 96.41% exact-set accuracy and LightGlue 90.60%, whereas joint transport achieves 72.56%. Joint inference therefore improves where the task asks for spatial allocation, not where success requires only the correct set of source identities. The distinction is also visible by query type. On single-source queries, joint mIoU is 75.50%. On two-source composites, it rises to 84.15%, compared with 81.51% for MaxSim, 81.44% for independent partial OT, 68.35% for LightGlue, and 61.37% for RootSIFT. The paired composite gain over MaxSim is 2.65 points (95% CI 0.44–4.90).

Composites are not necessarily harder under the primary metric. They contain two large, spatially coherent source regions and no overlay text by construction, whereas single-source overlay queries contain null pixels and aggressive retained-area settings. The higher composite mIoU therefore reflects the controlled generator rather than a general claim that multi-source attribution is easier. Exact-set accuracy and the independent-overlap diagnostic remain necessary to expose source-cardinality errors that region averages alone may obscure.

Figure 3 exposes variation hidden by the single-source average in Table 1. The two crop settings and two overlay settings produce distinct distributions, with a pronounced low-score tail when unsupported border or headline pixels are present.

A. Composite ablations

Table 2 Mechanism and missing-source analyses. Panel A uses 234 composites (Full exact-set on composites: 81.62%); the no-penalty overall exact-set over all queries is 35.56% versus 72.56% for the full method. Panel B reports one-parent ranking and hard null IoU and all-parent hard null IoU; only methods with an explicit null score are shown. Values except latency are percentages.

Config.	Exact	Patch	mIoU	Med. s
Full	81.62	89.24	84.15	4.04
No geometry	70.51	83.81	76.51	0.84
Uniform cap.	82.91	89.72	84.82	5.51
No penalty	**88.46**	**91.32**	**87.80**	4.06

B. Omitted-source stress tests

Method	1P ROC	1P PR	1P IoU	All IoU
Geo-MaxSim	85.38	86.84	64.73	69.39
MNN-RANSAC	87.22	80.59	**80.51**	**95.38**
Joint transport	**92.00**	**89.83**	63.05	59.46

Because transform type changes with the retained-area schedule, these differences should not be attributed to retained area alone.

The shortlist contains every available parent for 98.12% of test queries. Conditional joint mIoU is 78.17%, compared with the end-to-end 77.23%. For the evaluated joint pipeline, the 0.94-point difference indicates that candidate retrieval is a smaller limitation than the downstream attribution stage.

The conditional result also bounds what regional inference can recover: if a parent never enters the shortlist, no closed-gallery allocator can label its region correctly. The small end-to-end gap shows that this ceiling is reached infrequently in the controlled test, but it should not be extrapolated to an open-Web gallery. Larger and more heterogeneous collections may require a different retrieval index or query expansion before the same allocation layer is applied.

5.2 Mechanism ablations

Panel A of Table 2 isolates the main design choices on the 234 composites. Removing the geometric residual lowers mIoU from 84.15% to 76.51%. The paired difference is 7.64 points (95% CI 5.23–10.20), and exact-set accuracy falls by 11.11 points (6.90–15.55). Geometry therefore contributes information beyond the frozen appearance descriptors.

Projected-area capacity is not an accuracy improvement in this benchmark. Uniform capacity reaches 84.82% mIoU, 0.67 points above the full system; full minus uniform has a 95% CI of $[-1.48, -0.03]$. The area-based formulation is faster in the

recorded composite runs, with 4.04 s median latency versus 5.51 s, and retains a direct geometric interpretation, but the observed accuracy favors uniform capacity slightly.

The source-count penalty controls a different trade-off. Removing it raises composite mIoU to 87.80% but reduces overall exact-set accuracy from 72.56% to 35.56% and single-source exact-set accuracy from 70.30% to 22.33%. Across all queries, the full-minus-no-penalty exact-set difference is 37.01 points (95% CI 34.49–39.66). In this operating regime, penalizing additional sources supports mixed-cardinality parsimony while sacrificing some two-source recall.

5.3 Missing-source stress tests

Panel B removes one or both true parents from the gallery while preserving query labels for null-region evaluation. When one parent of a composite is absent, joint null mass yields 92.00% AUROC (95% CI 90.23–93.67) and 89.83% AUPR (86.53–93.45), exceeding the continuous rankings of MaxSim and MNN-RANSAC. At the frozen hard decision, however, joint null IoU is 63.05% (59.19–66.82), below MNN-RANSAC at 80.51% (78.92–82.17).

The gap widens when both parents are absent. Joint hard null IoU is 59.46% (95% CI 56.30–62.63), compared with 69.39% for MaxSim and 95.38% for MNN-RANSAC. Null mass is therefore effective as a ranked missing-source signal under partial omission, whereas the frozen operating point is not a reliable all-parent-omission decision rule.

5.4 Solver behavior and qualitative cases

All 1,170 final joint solves converged. The maximum row residual was 9.99×10^{-7} and the maximum capacity violation was 5.55×10^{-17}. Median end-to-end attribution latency was 0.465 s per query, with a 5.60 s 95th percentile. Composite latency had a 4.04 s median, reflecting the additional pair-subset evaluations required by the implemented inference procedure.

The difference between median and tail latency is consistent with variable subset search. Singletons that clearly lose to the empty set terminate with fewer competitive pair hypotheses, whereas composites and ambiguous near duplicates require more pair evaluations. The timing is an implementation measurement on the stated hardware, not a hardware-normalized efficiency claim. Cached gallery descriptors remove encoder cost for the fixed gallery, but each query still requires feature extraction, shortlisting, geometric estimation, and transport solves.

Figure 4 shows that the latency tail is concentrated among composite queries and high-iteration solves. The trend is not perfectly one-dimensional because candidate verification and subset enumeration also contribute to wall-clock time, but it confirms that the reported median alone does not describe the computational profile.

Figure 5 illustrates a crop, a successful composite, and a composite failure. The successful cases show that the joint prediction separates source regions while preserving null areas. The failure is consistent with the complementary risk that subset selection may suppress a visually valid source when its evidence gain does not offset the source-count penalty.

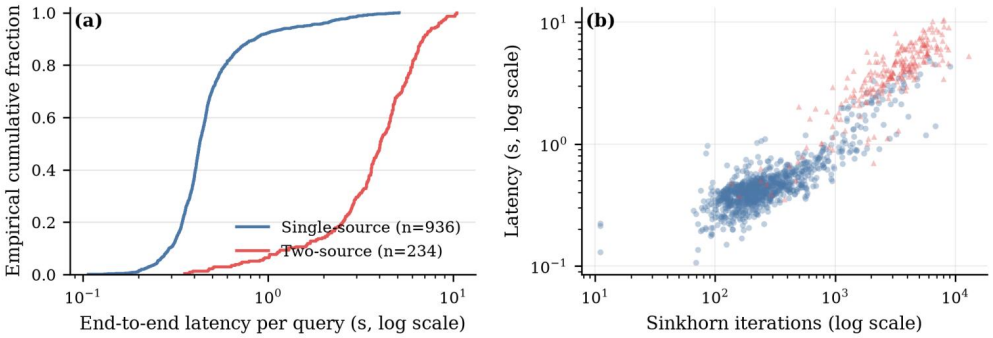


Fig. 4 Solver diagnostics over all 1,170 frozen test queries. (a) Empirical latency distributions separate 936 single-source queries from 234 two-source composites. (b) Per-query latency against generalized Sinkhorn iterations; both axes are logarithmic. The plot reports implementation behavior rather than a hardware-normalized efficiency comparison.

The examples also clarify the resolution of the reported masks. Boundaries follow the 8×8 patch grid rather than object contours, and a patch containing both an overlay edge and photograph content receives one label. The qualitative figure is therefore best read as an explanation of allocation behavior. It is not evidence for fine-boundary segmentation, and the aggregate results remain the basis for method comparison.

6 Discussion and Limitations

The results separate two questions that are often merged in provenance pipelines. Exact-set recovery asks whether the correct identities appear in the output; region attribution asks whether each identity explains the correct area. RootSIFT and LightGlue answer the first question more accurately in this controlled gallery, while joint transport answers the second more accurately. The independent-plan pipeline frequently produces overlapping claims and is less accurate after its selection and conflict-resolution stages; the comparison does not attribute the entire gap to conflict resolution alone. The source-count ablation reveals a cardinality trade-off: the penalty improves single-source exact-set recovery but can suppress a legitimate second parent. A future system should couple localization with a separately calibrated cardinality model instead of using one scalar penalty for both regimes.

This distinction suggests a practical two-stage interface. A retrieval-oriented matcher can present high-confidence source identities, while joint transport can explain where those identities are supported and which regions remain unmatched. Disagreement between the two outputs is informative rather than necessarily erroneous. A source may have strong sparse correspondences and therefore enter the retrieved set, yet explain too little area to survive the transport gate. Conversely, a source may receive a coherent region mask while exact-set recovery includes or omits an additional weak candidate. Presenting both views would let an analyst inspect the evidence instead of collapsing all behavior into one score.

The capacity ablation also narrows the mechanism claim. Projected area gives each source patch a geometrically interpretable upper bound and reduced median composite latency in the recorded run. Uniform capacity was nevertheless 0.67 mIoU points better. Capacity calibration, rather than the existence of a cap alone, remains unresolved. Learned or uncertainty-aware capacities could preserve the competition mechanism without treating one affine estimate as exact.

The result is consistent with two opposing effects. A projected-area cap can prevent a warped cell from absorbing implausibly large mass, but the cap inherits error from affine estimation and coarse grid clipping. Uniform capacity removes that error at the cost of weaker geometric meaning. Because the benchmark uses hard-seam composites and moderate affine edits, appearance cost plus geometric residual may already provide most of the useful spatial restriction. More severe projective transformations could change this balance, but the present data do not establish such an advantage.

The null column behaves differently as a score and as a decision. Under one-parent omission, its continuous ranking is strong, but the fixed hard label is weaker than MNN-RANSAC, especially when all parents are absent. Deployment would require a validation protocol that includes complete omissions, explicit risk-coverage selection, and shift-aware calibration.

Null behavior also depends on the candidate gallery. In the one-parent test, the remaining true parent anchors part of the query and makes the missing region comparatively distinct. When both parents disappear, visually related distractors can absorb mass and reduce the hard null area. A useful open-set extension should model gallery adequacy at both image and region levels. Candidate-level confidence can detect a globally poor shortlist, while patch-level null mass can localize unsupported content after a plausible source has been selected.

The evaluation has four material scope limits. First, all test queries are deterministic crops, overlays, and hard-seam composites derived from a single 5Pils image pool; they do not sample the natural distribution of online reuse. Second, the gallery is closed and contains at most two visual parents, so the method does not recover earliest publication or propagation direction. Third, affine geometry does not model severe projective or nonrigid edits. Fourth, the capture-cluster audit is defined by a frozen DINOv2 proposal stage and RootSIFT-RANSAC acceptance thresholds. It found no cross-split clusters under that protocol, but undetected same-capture relations remain possible.

Two additional design limits follow from the patch representation. The 8×8 grid cannot resolve thin overlays or precise boundaries, and frozen descriptors may favor semantic resemblance over exact instance identity in texture-poor regions. Increasing resolution would improve boundary granularity but enlarge every candidate cost matrix and may amplify noisy local similarities. A hierarchical solver could first allocate coarse mass and refine only uncertain boundaries, preserving the joint budget without paying the full high-resolution cost for every candidate.

The controlled benchmark remains valuable despite these limits because it tests a causal question that natural Web data cannot answer without annotation: whether sources should compete during regional inference. Exact parent masks, frozen query generation, and capture-cluster resampling make that comparison auditable. The next

empirical step is not to relabel released evidence URLs as truth, but to construct a natural-reuse tier with independently reviewed source relations, region masks, and a query-independent frozen gallery. The present hyperparameters should then be transferred without retuning before any natural-domain claim is made.

Region-to-source maps should be treated as retrieval evidence for human inspection, not as a verdict about news truth. Same-capture correspondence does not establish which publisher appeared first or whether the accompanying claim is correct. Natural reuse evaluation will require independently annotated source relations and a query-independent frozen gallery.

There are also data-governance considerations. News photographs may remain copyrighted even when their URLs are public, and the availability of remote images changes over time. Releasing manifests, hashes, transforms, and masks separately from the underlying photographs allows other researchers to reconstruct the benchmark under the original dataset terms without silently redistributing third-party media. Human review is still required for the accepted same-capture edges and for any future natural-reuse annotations.

7 Conclusion

We formulated reused-news-image analysis as closed-gallery region-to-source attribution and solved it with joint partial transport. Frozen patch descriptors and source-to-query geometry define candidate evidence, while a shared query-patch budget forces active sources and an explicit null destination to compete. An entropy-free subset energy separates numerical smoothing from source-cardinality selection.

On 1,170 capture-cluster-separated controlled queries, the complete system achieved 0.772 mask mIoU and improved over independent partial transport, LightGlue, and RootSIFT under paired cluster bootstrap. The comparison supports joint multi-source inference for regional localization. It does not imply superiority for source-set retrieval: RootSIFT and LightGlue recovered exact source sets more accurately. Ablations further show that geometry is useful, uniform capacity is slightly more accurate than projected-area capacity in this benchmark, and the source-count penalty trades composite recall for mixed-cardinality parsimony.

The omission tests identify a second boundary. Continuous null mass ranks missing-source regions well when one parent remains available, but the frozen hard decision is weak when the gallery omits all parents. Region localization, exact-set recovery, and open-set abstention therefore require distinct operating objectives. The next step is an independently annotated natural-reuse evaluation with a frozen query-independent gallery, followed by higher-resolution allocation and separately calibrated source cardinality and abstention.

Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Competing Interests. The authors declare that they have no known competing financial or non-financial interests that could have influenced the work reported in this paper.
Data availability. The study uses images released with the 5Pils dataset. The controlled-query manifests, split hashes, and evaluation protocol will be released subject to the original dataset terms and URL availability. Third-party news images are not redistributed beyond those terms.
Code availability. Code for query generation, integrity auditing, feature extraction, transport inference, baselines, and statistical analysis will be released with the final version.
Author contributions. T.L. conceived the study, designed the method, conducted the experiments, analyzed the data, and wrote the manuscript. M.H. contributed to data curation and manuscript revision. All authors reviewed the manuscript.
Generative AI statement. Generative AI tools assisted with code drafting and language editing. The authors are responsible for verifying the methods, experiments, numerical results, citations, and final text.

References

[1] Akhtar M, Schlichtkrull M, Guo Z, Cocarascu O, Simperl E, Vlachos A. Multimodal automated fact-checking: a survey. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics; 2023. pp. 5430–48.

[2] Gallegos Carvajal IM, Portelli B, Zini L, Baraldi L, Serra G. An in-depth survey on multimodal automatic fact-checking datasets. Multimedia Tools and Applications. 2026;85(588).

[3] Luo G, Darrell T, Rohrbach A. NewsCLIPpings: automatic generation of out-of-context multimodal media. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online; Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. pp. 6801–17.

[4] Abdelnabi S, Hasan R, Fritz M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. pp. 14940–9.

[5] Aneja S, Bregler C, Niessner M. COSMOS: catching out-of-context image misuse using self-supervised learning. Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37(12):14084–92.

[6] Tonglet J, Moens M-F, Gurevych I. "Image, tell me your story!" Predicting the original meta-context of visual misinformation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics; 2024. pp. 7845–64.

[7] Qi P, Yan Z, Hsu W, Lee ML. SNIFFER: multimodal large language model for explainable out-of-context misinformation detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. pp. 13052–62.

[8] Moreira D, Bharati A, Brogan J, Pinto A, Parowski M, Bowyer KW, et al. Image provenance analysis at scale. IEEE Transactions on Image Processing. 2018;27(12):6109–23.

[9] Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012. pp. 2911–8.

[10] Lindenberger P, Sarlin P-E, Pollefeys M. LightGlue: local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. pp. 17627–38.

[11] Zlatkova D, Nakov P, Koychev I. Fact-checking meets fauxtography: verifying claims about images. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Association for Computational Linguistics; 2019. pp. 2099–108.

[12] Müller-Budack E, Theiner J, Diering S, Idahl M, Ewerth R. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. Association for Computing Machinery; 2020. pp. 16–25.

[13] Geng J, Tonglet J, Gurevych I. M4FC: a multimodal, multilingual, multicultural, multitask real-world fact-checking dataset [Internet]. 2025. Available from: http://arxiv.org/abs/2510.23508

[14] Papadopoulos S-I, Koutlis C, Papadopoulos S, Petrantonakis PC. Similarity over factuality: are we making progress on multimodal out-of-context misinformation detection? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2025. pp. 5041–50.

[15] Jaiswal A, Wu Y, AbdAlmageed W, Masi I, Natarajan P. AIRD: adversarial learning framework for image repurposing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 11330–9.

[16] Noh H, Araujo A, Sim J, Weyand T, Han B. Large-scale image retrieval with attentive deep local features. In: Proceedings of the IEEE International Conference on Computer Vision. 2017. pp. 3456–65.

[17] Yang M, He D, Fan M, Shi B, Xue X, Li F, et al. DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 11772–81.

[18] Xiao Z, Suma P, Sachdeva A, Wang H-J, Kordopatis-Zilos G, Tolias G, et al. LOCORE: image re-ranking with long-context sequence modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. pp. 9580–90.

[19] Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision. 2004;60(2):91–110.

[20] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications in image analysis and automated cartography. Communications of the ACM. 1981;24(6):381–95.

[21] Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, et al. D2-Net: a trainable CNN for joint description and detection of local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. pp. 8092–101.

[22] Sarlin P-E, DeTone D, Malisiewicz T, Rabinovich A. SuperGlue: learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 4937–46.

[23] Sun J, Shen Z, Wang Y, Bao H, Zhou X. LoFTR: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. pp. 8922–31.

[24] Wang Y, He X, Peng S, Tan D, Zhou X. Efficient LoFTR: semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. pp. 21666–75.

[25] Edstedt J, Sun Q, Bökman G, Wadenbäck M, Felsberg M. RoMa: robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. pp. 19790–800.

[26] Xue F, Elflein S, Leal-Taixé L, Zhou Q. MATCHA: towards matching anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025. pp. 27081–91.

[27] Oquab M, Darcet T, Moutakanni T, Vo HV, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision. Transactions on Machine Learning Research [Internet]. 2024; Available from: https://openreview.net/forum?id=a68SUt6zFt

[28] Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems. 2013. pp. 2292–300.

[29] Benamou J-D, Carlier G, Cuturi M, Nenna L, Peyré G. Iterative Bregman projections for regularized transportation problems. SIAM Journal on Scientific Computing. 2015;37(2):A1111–38.

[30] Chapel L, Alaya MZ, Gasso G. Partial optimal transport with applications on positive-unlabeled learning. In: Advances in Neural Information Processing Systems. 2020. pp. 2903–13.

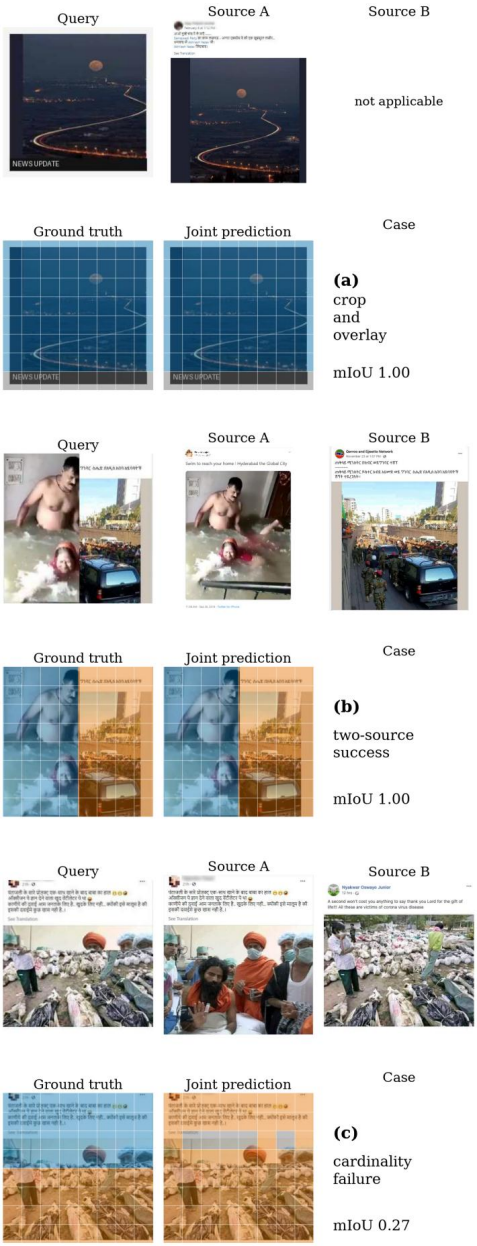


Fig. 5 Qualitative closed-gallery attributions for a single-source crop, a successful two-source composite, and a failure case. For each case, the upper row shows the query and available sources; the lower row shows the ground-truth patch mask, the joint prediction, and row-level source-mask mIoU. Blue and orange denote source identities and gray denotes null. The examples illustrate allocation behavior and do not replace aggregate evaluation.