SAME PHOTO, MORE EVIDENCE? MULTI-IMAGE VLMS OVERCOUNT REPEATED IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models increasingly answer questions about a set of images rather than one: the photographs on a product listing, the images attached to an incident report, the results of a web search. In such sets the same photograph routinely re-enters re-compressed, cropped or screenshotted. Existing multi-image benchmarks vary how many images a model sees and what they depict, but never whether an added image is a copy of one already present. The standard remedy, de-duplication by embedding similarity, does not settle this either: a re-upload of a photograph and a genuinely new photograph of the same object look alike to it, so any threshold either keeps the copies or throws away real evidence. We show that current multi-image VLMs count images rather than new observations. On a controlled spectrum that fixes image and token count, renderings of one photograph raise a model's belief almost as much as genuinely new viewpoints; on real product listings, three re-uploads of one mis-attributed photo change up to a third of decisions, on open 3B–72B models and on API-served ones alike. Copies of a correct photograph inflate confidence as well, and stating in the prompt that copies count once does not help. Attention analysis places the effect not in how much the answer attends to the repeated images, which is flat across conditions, but in reinforcement among those images' own tokens in the middle layers. The remedy is to remove the copies before the model sees the set. **Provenance-Aware Evidence Aggregation** (PAEN) scores each pair of images for provenance, asking whether they are renderings of one observation, by geometric verification: SIFT correspondences under a RANSAC homography, backed by a perceptual hash. This separates re-uploads from nearby viewpoints where semantic similarity cannot, even with a descriptor trained for copy detection. PAEN is training-free, returns four open architectures to within two points of their base accuracy with near-zero decision change while leaving genuinely new photographs untouched, and, applied before the images are sent, does the same for three models served only through an API (DeepSeek, GPT-5.6, Kimi). The signal multi-image aggregation needs is provenance, not similarity.

1 INTRODUCTION

An AI shopping assistant built on a vision–language model (VLM) is shown the photographs on a product listing and asked what is being sold. Imagine that one photograph was uploaded to the wrong listing, and that it appears three more times: re-compressed by a messenger app, cropped square, and screenshotted (Fig. 1b). The three copies add no new information about the product, yet four of the six photographs now show the wrong item. Multi-image VLMs (Jiang et al., 2024; Li et al., 2025b; Laurençon et al., 2024) are deployed on exactly such collections, including product listings, incident reports, news coverage and the images an agent gathers from the web, and in all of them the same photograph often comes back in a different form. A person who notices that the three extra photos are re-uploads of one picture treats them as a single observation, not four; we ask whether multi-image VLMs do the same.

To our knowledge, no prior work has tested this. Multi-image benchmarks vary how many images a model sees and what they show (Wang et al., 2025a; Jiang et al., 2024; Li et al., 2025c; Guo

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

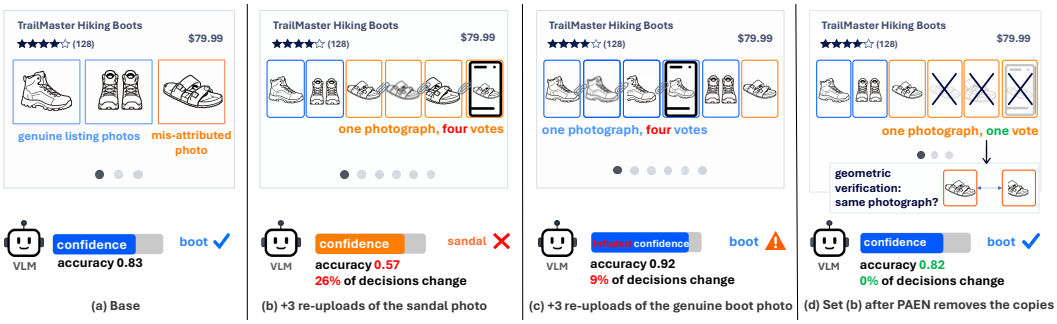

Figure 1: **One photograph, counted several times (a–c), and counted once with PAEN (d).** A product listing whose photo set contains one mis-attributed photo (orange). **(a)** The model answers correctly. **(b)** Adding three simulated re-upload copies of that same photo (a messenger re-encoding, a square crop and a phone screenshot) changes the decision on 26% of listings and costs 0.83→0.57 in accuracy, although no new observation was added. **(c)** The same copies made from a genuine photo change 9% of decisions and raise confidence by +1.1 nats, although they add no new observation. **(d)** PAEN asks whether two images are renderings of the same observation and removes the copies it flags, restoring base accuracy with 0% decision change. Numbers: Qwen2.5-VL-7B on 312 product listings, with the mis-attributed photo placed first, in the middle and last (Sec. 2.2).

et al., 2026; Das et al., 2026), and report that redundant views hurt, but never whether an added image is a new observation, so a rendering and a viewpoint are never separated. On the text side, retrieval-augmented models are known to follow the majority of retrieved passages and to be swayed by repetition (Wang et al., 2025b; Ge et al., 2025; Jin et al., 2024; Schuster et al., 2026); detection there is easy, because a re-posted passage is near-verbatim. A re-uploaded photograph is not, and the natural detector, embedding similarity (Radford et al., 2021), is the wrong instrument: a re-upload of one photograph and a photograph taken a step to the side are equally similar to it, so any threshold either admits the copies or discards real evidence. What is needed is not a measure of resemblance but an answer to another question: were the two images produced by the same act of observation?

So do multi-image VLMs treat copies of one photograph as a single observation? We find that they largely do not. To test this we construct a *visual novelty spectrum*: image sets that hold the number of images, the number of visual tokens and the semantic conclusion fixed, and vary only how much new observation each additional image contributes, from exact duplicates and label-preserving renderings of one view, through partially overlapping crops, to adjacent and far viewpoints and independent instances. Across two datasets and four architectures, renderings of a single photograph are credited almost exactly like new viewpoints. Only exact, pixel-identical copies receive discount, and only on some models; others, including the 72B model, credit even those in full (Sec. 2.1).

The failure is not an artifact of log-odds. On real product listings, asked the plain question a shopper would ask, three simulated re-upload copies of one mis-attributed photo change 27–34% of the decisions of models served through an API, including GPT-5.6, DeepSeek-V4.1 and Kimi K2.6, and 12–26% on open models, whereas three genuinely new photographs of the product raise accuracy. Two controls show what the model is doing: copies of a genuine photo also inflate its confidence, so what is being counted is images rather than observations; and writing the counting rule into the question leaves the failure intact, so the model is not misreading the question but cannot see that the copies are one photograph (Sec. 2.2). MIHBench (Li et al., 2025c) reduces multi-image hallucination by balancing attention across images, but this makes the failure worse (Sec. 2.3).

Why do copies raise the model's belief? In a decoder-only VLM, information from the images can reach the answer in two ways: the answer position reads the image tokens directly, and the tokens of each image also read the tokens of earlier images. We find that only the second path carries the effect (Sec. 3). Attention from the answer position to the repeated images is essentially identical across all spectrum conditions, so the answer does not "look more" at duplicated content, and re-weighting text→image attention leaves most of the amplification in place. What removes it is masking the attention between the repeated images' own tokens, mostly in the middle layers, while a shuffled mask of the same strength does not and genuinely new viewpoints are untouched.

We summarize our contributions as follows:

- We show, on a controlled novelty spectrum and a realistic product-listing task, that multi-image VLMs, both open and API-served, credit re-uploaded copies of a photograph as new evidence. Prompting, similarity-based de-duplication and attention balancing do not fix this.
- We locate where the extra belief that copies produce comes from. It is not that the answer attends more to the repeated images. Instead, the copies' own tokens reinforce one another through cross-image attention, mostly in the middle layers, and masking those links alone removes the gain while leaving genuinely new views unaffected.
- We propose PAEN, a training-free step that detects copies by geometric verification and removes them before the model sees the set; it restores base accuracy on open models from 3B to 72B and on three API-served models.

2 Counting Images versus Counting Observations

We first measure the effect on a controlled spectrum that varies only how much new observation each added image contributes (Sec. 2.1), then on a decision task built from real product listings (Sec. 2.2). Sec. 2.3 shows that three simple fixes do not work. Duplicates that occur naturally inside listings are treated in App. A.4.

Renderings versus observations. Let x be a photograph and T a label-preserving transformation, such as re-sizing, JPEG re-compression, a brightness or contrast change, blur, grayscale, cropping, panning, a mild rotation or perspective change. We call $T(x)$ a rendering of the same observation, because whatever the scene has been observed to be, $T(x)$ adds nothing to it. A photograph taken from a different viewpoint, or of a different instance, is a new observation. New viewpoints of one object are not statistically independent, yet they do add observations, which is why the axis we vary is novelty rather than correlation. A horizontal flip also counts as a rendering for the questions we ask, which do not depend on left and right; App. B.7 reports it separately.

Protocol and metrics. Every experiment starts from a base set of images that supports one answer, adds k images of a single kind (copies of an image already in the set, or genuinely new photographs), and compares the model's answer before and after the addition. Every question in the main text offers two candidate answers, A and B, and the model replies with one word. A four-candidate and an open-ended version of the central comparison give the same result (App. B.1). We take its belief to be $\log P(A)/P(B)$ at the first generated token, measured in nats, and its decision to be whichever answer is the more likely of the two. We report three kinds of metrics: accuracy against the gold answer; decision change, the fraction of decisions that differ from the decision on the same set before the images are added; and calibration, measured by wrong-answer confidence, Brier score (Brier, 1950) and negative log-likelihood (NLL). Intervals are bootstrapped (Efron, 1992).

Controls. Two controls keep the comparison honest. First, multi-image VLMs have strong position effects (Tian et al., 2025), so every condition is evaluated with the conflicting image first, in the middle, and last, and averaged. Second, we report belief changes relative to a NEUTRAL control that inserts gray or noise images at the same positions, so that what is measured is what a rendering adds over and above one view of it. We restrict the main results to $k \leq 3$ additions: neutral insertions have negligible average effect here, whereas longer sequences are unstable (App. A.6).

Data and models. CO3D object sequences (Reizenstein et al., 2021) provide real multi-viewpoint imagery (10 frames per object, 12 categories spanning textured, texture-poor, thin and large objects); COCO (Lin et al., 2014) provides independent instances; Amazon Berkeley Objects (ABO) (Collins et al., 2022) provides real product listings with several catalog photographs per product and product-type/color metadata. Models: Qwen2.5-VL-7B (Bai et al., 2025), Qwen2-VL-7B (Wang et al., 2024), InternVL3-8B (Zhu et al., 2025) and LLaVA-OneVision-7B (Li et al., 2025a) are used throughout. On the realistic product-listing task we also evaluate Qwen2.5-VL at 3B, 32B and 72B, and three models served only through an API: GPT-5.6 Luna, DeepSeek-V4.1-Flash and Kimi K2.6. These give decision-level results only, since their attention is not accessible; App. A.3 describes how we query them.

Table 1: **The novelty spectrum.** Log-odds gain over a single supporting image when four more are added (nats; CO3D, 72 trials, conflict position balanced; trial-level bootstrap 95% intervals in App. A.5). Column headings give the new observation each added image contributes.

Model	DUP	DERIVED	PARTIAL	ADJ	FAR	INDEP
new observation	none	none	partial	yes	yes	other object
Qwen2.5-VL-7B	+0.10	+3.18	+1.98	+3.59	+3.61	+0.69
InternVL3-8B	+4.07	+5.05	+2.60	+5.76	+5.79	+6.38
Qwen2-VL-7B	+2.27	+3.36	+1.07	+3.24	+3.23	+2.28
LLaVA-OneVision-7B	+1.87	+1.94	+0.26	+2.02	+1.97	+1.48
Qwen2.5-VL-72B	+3.04	+3.15	+2.07	+3.37	+3.54	+0.25

2.1 RENDERINGS ARE CREDITED LIKE NEW VIEWS

Given a supporting observation x_A and one conflicting observation x_B, we build image sets of fixed size in which the images supporting A are one of: DUP (exact copies of x_A), DERIVED (renderings $T_i(x_A)$), PARTIAL (overlapping crops that never show the whole of x_A), ADJ and FAR (nearby and widely separated viewpoints of the same object), and INDEP (different instances of the category). A model that tracks novelty should credit DUP$\approx$DERIVED$<$PARTIAL$<$ADJ$\leq$FAR.

Tab. 1 shows the credit the models actually assign. On every one DERIVED is close to ADJ/FAR: four renderings of one photograph are worth about as much as four genuinely new viewpoints. This is not because a rendering is more persuasive on its own. When we add only one image instead of four, a rendering, an adjacent view, a far view and another instance each raise the log-odds by about the same amount (1.8–2.1 nats). The difference therefore arises when the model combines several images, not from the images themselves. The same pattern holds on COCO, on a twelve-category CO3D spectrum, and the neutral-centred k=3 protocol used in the rest of the paper (App. A.5).

2.2 THE FAILURE FLIPS DECISIONS ON A REALISTIC TASK

A spectrum is not an application. We therefore build a decision task around real product listings.

Task. Each product listing in Amazon Berkeley Objects has at least five real catalog photos of the product. We build a base set of three images: two of these photos and one catalog photo of a different, easily confused product, e.g. a sandal on a boot listing (Fig. 1a). This simulates a common error on real platforms, where a seller uploads a wrong image to a listing. The question is "These are the photos on one product listing. What is the product being sold: a boot or a sandal?"; the gold answer is the listing's product type from its metadata. A second task asks the product's color.

Conditions. We then add three images of one kind to the base set and compare the model's answer with its original answer. The added images are re-upload copies of the mis-attributed photo (DERIVED$_3$, Fig. 1b), which add no new observation; three exact copies of the mis-attributed photo (DUP$_3$); re-upload copies of a genuine photo (TRUEDUP$_3$, Fig. 1c), which add none either but sit on the correct side; three further genuine catalog photos of the product (NEW$_3$); or three gray images (NEUTRAL$_3$). The re-upload copies are produced by six fixed pipelines, each imitating what one real platform does to a forwarded photo: messenger forwarding, a square crop, a phone screenshot, a CDN thumbnail, double re-encoding, and a link-preview crop (App. A.1); each set draws three of the six. We use 312 listings for product type and 300 for color.

Tab. 2 gives the result. DERIVED$_3$ changes 12% to 26% of decisions, depending on the model, and every changed decision moves towards the wrong product. DUP$_3$ shows that even exact copies are discounted only partly. TRUEDUP$_3$ shows that the failure is not about the wrong photo in particular: copies of a correct photo add no observation either, yet they raise confidence by about half of what three new photographs add. NEW$_3$ moves decisions the other way, as genuine evidence should. Stating the counting rule in the question ("some photos may be re-uploads of the same photo; count each distinct photo once") does not help: the model does not misread the question, it cannot see that the copies are one photograph. Duplicates that occur naturally inside listings are credited the same way (App. A.4). The color question and the rule-stated prompt give the same picture (App. B.2).

Table 2: **The failure on real product listings.** Product-type question. For the base set and after each addition we report accuracy (acc) and the share of decisions that change (Δdec.). For TRUEDUP$_3$ we also give the change in log-odds (ΔLO), since its copies push towards the correct answer and accuracy alone hides the effect. Color question and the rule-stated prompt: App. B.2.

		DERIVED$_3$		DUP$_3$		TRUEDUP$_3$			NEW$_3$	
Model	Base	acc	Δdec.	acc	Δdec.	acc	Δdec.	ΔLO	acc	Δdec.
Qwen2.5-VL-7B	0.83	0.57	26%	0.68	15%	0.92	9%	+1.1	0.96	13%
InternVL3-8B	0.76	0.57	21%	0.60	18%	0.84	10%	+0.1	0.93	17%
Qwen2-VL-7B	0.80	0.61	20%	0.64	17%	0.89	9%	+1.1	0.92	12%
LLaVA-OneVision-7B	0.71	0.60	12%	0.60	12%	0.79	9%	+0.7	0.84	13%
Qwen2.5-VL-72B	0.95	0.71	24%	0.79	16%	0.97	3%	+0.7	0.98	4%

Controlled counterpart. We repeat the same protocol on CO3D object sequences, where the base set is two views of one object and one real photo of a confusable category, and the copies are synthetic renderings (179 sets, App. B). All four architectures show the same pattern, with new views intact. Because CO3D fixes the viewpoint of every frame, it is also the setting we use for the mechanism analysis, where control over viewpoint is needed.

2.3 WHY CHEAP FIXES FAIL

We tried three obvious fixes. However, none of them works.

Prompting. Stating the counting rule in the question (Tab. 9, App. B.2) or instructing the model that some photos may be near-duplicate copies that should count once does not reduce the failure on any architecture. Nor does the model raise the matter itself: in free-form explanations it never remarks that any of the images are copies. The instruction has nothing to act on: as Sec. 3 shows, the copies reinforce one another among the image tokens before the answer is formed, and the question text does not reach that stage.

Similarity de-duplication. CLIP similarity does not separate renderings of one view from new viewpoints of the same object (0.918 vs. 0.915 on CO3D), so any threshold either leaks renderings or merges genuinely new views. On the CO3D task, hard CLIP de-duplication merges the two true views and lowers base accuracy to 0.79 while still leaking renderings (DERIVED$_3$: 0.62), which is worse than no intervention; DINOv2 (Oquab et al., 2024) de-duplication behaves the same way (Tab. 11). A descriptor trained for copy detection, SSCD (Pizzi et al., 2022), draws the line better than CLIP but still not where it has to be: at the threshold that separates renderings from adjacent viewpoints it misses six in ten phone screenshots, and a threshold that catches them merges a third of the adjacent viewpoints (App. A.2). Perceptual hashing (pHash) is strong on photometric renderings but leaks geometric ones (crops, zoom, pan) and flips (Tab. 13). Both pHash and SSCD are reported as baselines throughout.

Attention balancing. Equalizing the attention each image receives (the method of MIHBench (Li et al., 2025c), re-implemented on Qwen2.5-VL) gives each copy a full share. The gain from renderings does not decrease, and the small discount the model applies to exact copies is removed as well (exact-copy gain +0.7→+2.5 nats).

The pattern is consistent: a remedy must know whether two images are the same observation, and semantic similarity does not carry that information. Copies add no observation, yet they raise the model's belief; we call this extra gain the amplification. Before building a remedy, we ask where in the model it arises.

3 MECHANISM: COPIES REINFORCE ONE ANOTHER

A remedy has to act somewhere. Two sites are available in a decoder-only VLM: the answer's attention to each image, and the images' attention to one another. We test both on the novelty spectrum, on all four architectures.

Table 3: **Where the amplification lives.** Log-odds gain over a single supporting image when four images are added (nats; CO3D, 72 trials). Which images are copies is known from the construction of the data and given to the intervention directly; the estimator of Sec. 4 is not used here. Columns: no intervention (none); the answer's attention to each copy scaled down by the number of copies (text-side); text-side plus masked attention between copies in all layers (+cross-image) or only in the early, middle or late third of the layers; and text-side plus a mask of the same total strength on randomly chosen image pairs (shuffled). Upper block: four renderings of one photograph, where the ideal gain is 0. Lower block: four genuinely new viewpoints, where the gain should stay unchanged.

	all layers			cross-image mask, one third of layers			
Model	none	text-side	+cross-image	early	middle	late	shuffled
DERIVED$_4$*: four renderings of one photograph (target: 0)*							
Qwen2.5-VL-7B	+3.2	+2.6	-0.1	+2.0	+1.6	+3.1	+1.9
InternVL3-8B	+5.1	+3.3	+1.3	+4.8	+2.5	+4.7	+3.3
Qwen2-VL-7B	+3.4	+2.5	+0.3	+3.5	+1.4	+2.9	+2.4
LLaVA-OneVision-7B	+1.9	+0.5	+0.6	+2.1	+0.8	+1.1	+2.5
ADJ$_4$*: four genuinely new viewpoints (target: unchanged)*							
Qwen2.5-VL-7B	+3.6	+3.5	+3.5	+3.6	+3.5	+3.6	+3.6
InternVL3-8B	+5.8	+5.4	+5.1	+5.5	+5.4	+5.7	+5.3
Qwen2-VL-7B	+3.2	+3.2	+3.0	+3.3	+3.0	+3.2	+3.2
LLaVA-OneVision-7B	+2.0	+1.8	+1.8	+2.0	+1.9	+1.9	+2.0

The first site is not where the amplification lives. On Qwen2.5-VL the attention mass from the answer position onto the supporting images is nearly constant across DUP, DERIVED, ADJ, FAR and INDEP (0.042–0.046) while the log-odds these conditions produce differ by four nats; InternVL3 (0.043–0.044) and Qwen2-VL (0.076–0.079) repeat this, so the answer does not "look more" at repeated content. Consistently, re-weighting text→image attention by effective multiplicity leaves most of the amplification in place (Tab. 3, "text-side").

The second site is. Tab. 3 shows that masking image→image attention between dependent images brings four renderings down to the worth of about one image (DERIVED$_4$ from +3.2 to -0.1 on Qwen2.5-VL), while genuinely new viewpoints keep their full value (ADJ$_4$ from +3.6 to +3.5). Three controls make the reading specific. Restricting the mask to the middle third of the decoder recovers between half and two thirds of the effect, whereas the early and late thirds recover little, so the amplification is built in a particular band of depth rather than accumulated everywhere. A shuffled-dependence mask that applies the same total bias to the wrong pairs is far weaker, so the effect comes from masking the right links rather than from masking as such. And under every variant ADJ$_4$ moves by at most 0.7 nats, against drops of three to five nats on DERIVED$_4$, so the mask does not simply lower every gain: it removes the gain from copies and leaves the gain from genuine views. LLaVA-OneVision-7B behaves differently, and we discuss it separately in App. B.5.

Two consequences shape the method: the amplification is carried by the copies' own tokens, so removing the copies takes it out at its source, and what decides whether an image is a copy must be pairwise provenance. This also explains why prompting and attention balancing (Sec. 2.3) fail: they act on the text side.

4 METHOD: PROVENANCE-AWARE EVIDENCE AGGREGATION

In this section we propose Provenance-Aware Evidence Aggregation (PAEN), a training-free step that de-duplicates an image set by provenance before the set reaches the model. For every pair of images it estimates whether the two are renderings of the same observation, and it keeps one image per observation, so the model receives the de-duplicated set in place of the raw one. It has two parts: a pairwise provenance estimate, and a rule that removes the copies the estimate flags.

Provenance estimation. For each pair of images (i, j) we estimate a dependence $d_{ij} \in [0, 1]$: how strongly the evidence says that the two are renderings of the same observation. The cue is geometric. A crop, a screenshot or a re-encoding of one photograph maps onto the original by a single planar transformation, a homography (Hartley et al., 2003), whereas two viewpoints of one object do not,

because parts of the object shift relative to one another. We therefore use classical geometric verification: local keypoints are detected and matched between the two images with the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), and Random Sample Consensus (RANSAC) (Fischler & Bolles, 1981) searches for the one homography that explains the most matches. The fraction of matches that this homography explains, the *inlier ratio* r_{ij}, is high for renderings and low for viewpoints (mean 0.52 vs. 0.04), where CLIP similarity cannot tell the two apart (Sec. 2.3). Photometric renderings that destroy keypoints, such as heavy re-compression or blur, are caught by a perceptual-hash match (Zauner, 2010):

$$d_{ij}^{\mathrm{geo}} = \min\bigl(1,\, r_{ij}/0.5\bigr), \qquad d_{ij} = \max\Bigl(d_{ij}^{\mathrm{geo}},\, \mathbf{1}\bigl[\mathrm{ham}(h_i, h_j) \le \tau\bigr]\Bigr). \tag{1}$$

where h_i is a 64-bit difference hash and τ=10 bits.

Aggregation. Images are visited in order, and an image is dropped if $d_{ij} \ge 0.5$ for an image j already kept (Fig. 1d), so that only the first version of each photograph remains. The threshold follows from the two means above: at an inlier ratio of 0.52, a rendering reaches $d^{\mathrm{geo}} = 1$, whereas an adjacent viewpoint at 0.04 stays below 0.1. Because the copies are removed before the model sees the set, the amplification of Sec. 3 cannot arise. An in-model alternative that keeps every image and masks the attention between dependent images by $1 - d_{ij}$ is reported as an ablation in App. B.6; it does not improve on removal.

Cost. PAEN is training-free. The overhead is the pairwise estimate of Eq. 1 on the CPU, about a second for ten images and under half that with per-image features cached, and the forward pass itself becomes shorter.

5 Experiments

All conditions follow the protocol of Sec. 2. We ask three questions: does PAEN repair the failure, is it the signal or the rule that matters (Sec. 5.1), and what does it cost (Sec. 5.2).

5.1 Main results on real product listings

Baselines. We compare PAEN against no intervention and against two de-duplication baselines that use the removal rule of Sec. 4 with a different signal: perceptual hashing (pHash), the standard tool for removing re-encoded copies, which drops an image whose difference hash lies within 10 bits of one already kept (the hash that backs up Eq. 1, used on its own); and SSCD (Pizzi et al., 2022), a descriptor trained for image copy detection, with its cosine threshold set on the spectrum (App. A.2). Similarity-based de-duplication and attention balancing fail for the structural reasons given in Sec. 2.3.

Tab. 4 reports the repair on the sets of Tab. 2. PAEN returns every architecture to within two points of its base accuracy with at most 1% decision change, removes the confidence inflation from copies of a genuine photo, and leaves genuinely new photographs untouched. pHash leaks screenshots and crops; SSCD recovers more but stops short of PAEN on every architecture, since the threshold that spares adjacent viewpoints misses most screenshots. The three models served only through an API, GPT-5.6 Luna, DeepSeek-V4.1-Flash and Kimi K2.6, fail in the same way, and PAEN applied on the client before the images are sent brings their decision change down to the floor of their own run-to-run noise. The controlled CO3D setting and the earlier ABO construction give the same ordering (App. B.6), and so does model scale: from 3B to 72B in the Qwen2.5-VL family the failure does not shrink, and PAEN repairs every size (App. B.4).

The signal matters, not the aggregation rule. Holding the rule fixed and changing only the signal, similarity-based de-duplication with CLIP or DINOv2 merges genuinely different views and lowers base accuracy, while pHash and SSCD leak screenshots and crops; the provenance signal does neither. Holding the signal fixed and changing the rule, masking the attention between flagged images instead of removing them, or a hybrid of the two, also restores accuracy (App. B.6). Neither improves on removal: masking over-corrects on exact copies and recovers less on real re-upload chains. Together with Tab. 3, this is the evidence for the central claim: the failure is repaired by knowing provenance, not by any particular way of acting on it.

Table 4: **Failure and repair on real product listings** (sets and protocol of Tab. 2; columns as in Tab. 2, with TRUEDUP$_3$ reporting Δdec. only). pHash and SSCD use PAEN's removal rule with a similarity signal in place of provenance. Above the double rule are models reached only through an API, with PAEN applied on the client before the images are sent. Please note that these APIs return no log-odds and are not deterministic: identical inputs re-sent change 5% (DeepSeek), 9% (GPT) and 4% (Kimi) of decisions, the noise floor for their Δdec. columns.

	Model	Method	Base	DERIVED$_3$ acc	DERIVED$_3$ Δdec.	TRUEDUP$_3$ Δdec.	NEW$_3$ acc
Commercial models (API)	DeepSeek-V4.1-Flash	none	0.85	0.57	28%	8%	0.97
		pHash de-dup	0.81	0.70	13%	7%	0.97
		SSCD de-dup	0.81	0.76	6%	6%	0.97
		PAEN	0.80	**0.80**	**3%**	2%	0.97
	GPT-5.6 Luna	none	0.81	0.58	27%	9%	0.96
		pHash de-dup	0.79	0.67	17%	8%	0.96
		SSCD de-dup	0.79	0.74	12%	8%	0.95
		PAEN	0.78	**0.77**	**8%**	8%	0.96
	Kimi K2.6	none	0.89	0.55	34%	7%	0.98
		pHash de-dup	0.87	0.67	21%	8%	0.98
		SSCD de-dup	0.88	0.79	11%	6%	0.98
		PAEN	0.87	**0.86**	**5%**	4%	0.97
Open-weight models	Qwen2.5-VL-7B	none	0.83	0.57	26%	9%	0.96
		pHash de-dup	0.83	0.71	12%	6%	0.95
		SSCD de-dup	0.82	0.77	5%	4%	0.95
		PAEN	0.82	**0.82**	**0%**	1%	0.95
	InternVL3-8B	none	0.76	0.57	21%	10%	0.93
		pHash de-dup	0.76	0.67	10%	6%	0.92
		SSCD de-dup	0.76	0.73	4%	3%	0.92
		PAEN	0.76	**0.76**	**0%**	0%	0.92
	Qwen2-VL-7B	none	0.80	0.61	20%	9%	0.92
		pHash de-dup	0.80	0.70	10%	4%	0.91
		SSCD de-dup	0.80	0.76	4%	2%	0.90
		PAEN	0.80	**0.80**	**1%**	0%	0.91
	LLaVA-OneVision-7B	none	0.71	0.60	12%	9%	0.84
		pHash de-dup	0.71	0.67	4%	3%	0.83
		SSCD de-dup	0.70	0.68	2%	2%	0.83
		PAEN	0.70	**0.70**	**1%**	0%	0.83
	Qwen2.5-VL-72B	none	0.95	0.71	24%	3%	0.98
		pHash de-dup	0.95	0.83	12%	2%	0.98
		SSCD de-dup	0.94	0.90	4%	2%	0.98
		PAEN	0.95	**0.95**	**0%**	0%	0.98

5.2 WHAT THE REPAIR COSTS: CORROBORATION, CALIBRATION, ROBUSTNESS

Preserving legitimate corroboration. A provenance filter is only useful if it does not discard real evidence. Under three genuinely new views accuracy moves by at most two points on CO3D, and on the spectrum PAEN zeroes the DUP/DERIVED gains while leaving FAR and INDEP untouched and removing on average 0.25 of four adjacent views. Adjacent or orbit-closing frames of one capture are the boundary: there the estimator sometimes merges two true views, and on a boundary-regime set (Tab. 20) PAEN costs 3–4 base-accuracy points (Qwen2.5-VL 0.94→0.90) while still restoring accuracy under three renderings from 0.70 to 0.89. The main CO3D task uses views with frame gap 3–7, the regime our claims cover; a learned provenance estimator is the natural way to extend them.

Calibration. Renderings inflate wrong-answer confidence (0.18→0.37, CO3D) and Brier score (0.120→0.333 on the realistic product-type sets); PAEN returns Brier to 0.126. Every intervention improves calibration over no intervention in all ten model×setting cells in which all were run, and PAEN has the lowest Brier in nine of them (the exception is LLaVA-OneVision on the realistic product-type task, where the masking ablation is 0.01 better).

Robustness. The findings do not depend on the evaluation surface: generated answers agree with the first-token decision in 88–100% of cases and a second question framing reproduces the effect (App. B.3).

6 Related Work

Multi-image VLM evaluation. Benchmarks such as MuirBench and Mantis evaluate multi-image understanding broadly (Wang et al., 2025a; Jiang et al., 2024; Fu et al., 2024; Meng et al., 2025); MI-HBench studies multi-image hallucination (existence, counting, identity consistency) as a function of image count and same-object ratio and proposes dynamic attention balancing (Li et al., 2025c); SSMNBench shows that redundant views cause distraction degradation and that models tend to average single-image semantics rather than synthesize across views (Guo et al., 2026); controlled analyses catalogue aggregation and multi-concept tracking failures (Das et al., 2026). None of these separates renderings of one observation from new observations, which is the variable we control, and none evaluates what happens when the same photograph re-enters a set in a different rendering. Position effects in long contexts are well documented (Liu et al., 2024; Tian et al., 2025); we balance and zero-center for them rather than study them. Studies of single-image VLMs cover hallucination (Li et al., 2023; Leng et al., 2024; Huang et al., 2024) and how attention passes visual information to the text across layers (Chen et al., 2024; Kaduri et al., 2025; Zhang et al., 2025).

Evidence aggregation and repetition in LLMs. In retrieval-augmented text settings, LLMs handle conflicting evidence poorly (Wang et al., 2025b; Xie et al., 2024); in fact-checking they follow the majority of retrieved sources and prioritize repeated low-credibility content (Ge et al., 2025), and their source preferences can be reversed by repetition from less credible sources (Schuster et al., 2026). We show the visual analogue, that it survives explicit instruction, and that it has a specific attention-level mechanism. Repetition also raises perceived truth in people (the illusory-truth effect (Hasher et al., 1977; Dechêne et al., 2010)); the standard we apply is normative, not a comparison against readers. The visual case differs in one respect that shapes the remedy: a re-posted passage is near-verbatim and trivially detectable, whereas a re-uploaded photograph is not, and the natural detector (embedding similarity) cannot separate it from a new viewpoint.

Near-duplicate detection and de-duplication. Detecting re-encoded, resized or cropped copies of an image is classical (local features with geometric verification (Lowe, 2004; Fischler & Bolles, 1981; Philbin et al., 2007); perceptual hashing (Zauner, 2010); learned descriptors trained for copy detection (Pizzi et al., 2022)), and embedding similarity (Radford et al., 2021) is the common modern shortcut. Learned matchers such as SuperGlue and LoFTR improve local correspondence (Sarlin et al., 2020; Sun et al., 2021), and large training sets are de-duplicated with hashes or embeddings (Lee et al., 2022; Abbas et al., 2023). Our contribution is not a new detector but the diagnosis that multi-image aggregation in VLMs needs a provenance signal rather than a similarity signal, which cannot separate renderings from nearby viewpoints, and the finding that once provenance is known the aggregation side is simple: remove what is flagged.

7 Limitations and Conclusion

Limitations. The classical estimator sets the boundary of the method. On held-out re-upload chains it detects 94% of phone screenshots and misses the rest, it merges adjacent frames of one capture into a single observation, and it merges recoloured variants of a product photo that share the photograph's geometry but not its evidence (App. A.2). Each of these is a property of the signal, not of the removal rule, and a learned provenance estimator is the natural replacement.

Conclusion. Across open and API-served multi-image VLMs, copies of one photograph, whether synthetic renderings, simulated re-upload chains or duplicates that occur naturally in product listings, are credited nearly as much as new photographs. When a mis-attributed photo is in the set, this flips decisions on the plain application question, whether or not the question says that copies count once. The failure survives prompting and similarity-based de-duplication, and it has a specific mechanism: cross-image reinforcement among the renderings' tokens in the middle layers. Asking "is this a new observation?" rather than "do these look alike?", and removing what is not, returns every evaluated model to within two points of its base accuracy. The signal that multi-image aggregation needs is provenance, not similarity.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting dataset. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, supporting qualitative and thematic data analysis. Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for language editing of the manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All models we evaluate are publicly available, the open-weight models through their released checkpoints and the API models through their public interfaces (App. A.3). All datasets are public. The task construction (Sec. 2.2), the re-upload pipelines (App. A.1) and the method (Sec. 4) are specified in full, and we provide sample code in the supplementary material.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. ABO: Dataset and benchmarks for real-world 3D object understanding. In *CVPR*, 2022.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of VLM failure modes. In *Findings of ACL*, 2026.

Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, 14(2): 238–257, 2010.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *ECCV*, 2024.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. In *IJCAI*, 2025.

Tianchen Guo, Chen Liu, Ling Chen, and Xin Yu. SSMNBench: Diagnosing image-based cross-view human-object understanding via single-view sufficiency and multi-view necessity. In *ECCV*, 2026. arXiv:2606.25634.

Richard Hartley, Andrew Zisserman, et al. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

Lynn Hasher, David Goldstein, and Thomas Toppino. Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1):107–112, 1977.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. MANTIS: Interleaved multi-image instruction tuning. *TMLR*, 2024.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *LREC-COLING*, 2024.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *CVPR*, 2025.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, 2024.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *ACL*, 2022.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *TMLR*, 2025a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3D in large multimodal models. In *ICLR*, 2025b.

Jiale Li, Mingrui Wu, Zixiang Jin, Hao Chen, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. MIHBench: Benchmarking and mitigating multi-image hallucinations in multimodal large language models. In *ACM Multimedia*, 2025c.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *TACL*, 12:157–173, 2024.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *ICLR*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *CVPR*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In *ICCV*, 2021.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In *ACL*, 2026.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *CVPR*, 2025.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. MuirBench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025a.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *COLM*, 2025b.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *ICLR*, 2024.

Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus, 2010.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *CVPR*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Protocol details

A.1 Simulated re-upload chains

The realistic setting produces copies with six fixed pipelines: messenger—the source is treated as a $2.5\times$ larger photo, scaled to a 1280 px longest side and JPEG-encoded at quality 70 with chroma subsampling; social—center square crop keeping 90% of the short side, JPEG q80; screenshot—the photo letterboxed in a 1080×1920 phone frame with status and navigation bars, JPEG q85, padded to square; thumbnail—downscale to 256 px, JPEG q60, upscaled by the client; rescreen—JPEG q90 $\rightarrow$ WebP q75 $\rightarrow$ JPEG q85 with a 0.9 resize; sidecrop—a 16:9 link-preview band, JPEG q80, padded to square without aspect distortion.

A.2 HELD-OUT EVALUATION OF THE PROVENANCE ESTIMATOR

We evaluate the estimator on held-out data at the PAEN threshold, $d \geq 0.5$. On 300 random listing photos passed through each of the six re-upload chains, it flags 98.7% of the copies: every messenger re-encoding, thumbnail and double re-encoding, 99.3% of square crops, 98.3% of preview crops, and 94.3% of phone screenshots, whose interface chrome and reduced photo size make keypoint matching hardest. At 0.9, the removal threshold of the hybrid ablation in App. B.6, screenshot detection falls to 72.3%. On 1,000 random pairs of photos from different listings of the same product type it flags 2.7%. Of ten flagged pairs we inspected, half are the same photograph reused across listings, such as a packaging shot or a ring-on-hand template, and are correct merges; the other half are recoloured variants of one photograph or different products photographed in an identical studio template, which share a photograph's geometry but not its evidence. A recoloured variant is genuine evidence for a colour question. The single further pair merged at the lower threshold is two different shoes on white backgrounds. Among the 3,120 within-listing pairs of genuine photos in the realistic setting it flags 5.2%. This is an upper bound on false merges, since some of these pairs are real duplicates: a manual audit of 50 random within-listing pairs from a separate ABO sample found one identical pair and three possible close-up crops of the same photograph. The merges cost little in practice: base accuracy under PAEN stays within one point of no intervention on both questions. When the two true views are adjacent or orbit-closing frames, geometric verification flags them as one observation although they are distinct captures—a conservative false merge at the estimator's boundary; we therefore require a viewpoint gap of 3–7 frames in the main task and report the boundary regime separately.

A learned copy-detection descriptor as the signal. Tab. 5 compares the geometric estimator with SSCD (Pizzi et al., 2022), a self-supervised descriptor trained for image copy detection, used in the same removal rule with its cosine threshold set on the spectrum (the cut that best separates DERIVED from ADJ pairs, τ^*=0.67; AUROC 0.97 against 0.99 for the inlier ratio). SSCD is invariant to the edits it was trained on, messenger re-encodings and thumbnails, but a phone screenshot or a preview crop moves its embedding as much as a change of viewpoint does, so no threshold separates the two: at τ^* it detects 41% of screenshots, and at the threshold that matches the geometric estimator's recall it merges 31% of adjacent-viewpoint pairs. The same shows on the natural duplicates of App. A.4: at τ^* SSCD flags 55% of listings against 10.5% for the geometric estimator, and the pairs it adds are overwhelmingly the same product photographed from another angle, whereas the pairs only the geometric estimator finds are the same photograph with dimension annotations drawn on it or re-cropped. As a signal for PAEN it recovers part of the loss on real re-upload chains but not all of it (Tab. 4).

Table 5: **Copy-detection descriptor versus geometric verification as the provenance signal.** Spectrum columns: fraction of rendering pairs and of adjacent-viewpoint pairs flagged as dependent (CO3D, 144 pairs each). Held-out columns: re-upload copies detected over the six chains, the screenshot chain alone, and pairs flagged among photos of different listings of the same product type (1,000) and among genuine photos within a listing (3,120). †geometry alone; the hash component of Eq. 1 flags the remaining exact and photometric renderings.

	spectrum (CO3D)		held-out ABO photos			
Signal	renderings flagged	adjacent views flagged	re-uploads detected	screenshots	cross-listing flagged	within-listing flagged
SSCD, τ=0.55	99%	31%	98.5%	93.0%	2.4%	8.5%
SSCD, τ^*=0.67	94%	12%	87.9%	40.7%	1.3%	4.9%
SSCD, τ=0.80	69%	4%	74.8%	0.0%	1.2%	2.5%
Geometric verification, $d \geq 0.5$ (PAEN)	79%†	1.4%	98.7%	94.3%	2.7%	5.2%

A.3 MODELS SERVED THROUGH AN API

We ran the realistic product-listing task on DeepSeek-V4.1-Flash, GPT-5.6 Luna and Kimi K2.6 through their public chat APIs, using the same listings, images, question and conditions as for the open models. Each image was scaled to 768 pixels on its longer side and sent inline. Thinking modes were switched off, the output was capped at 48 tokens, and the temperature was set to 0 where the API allows it (GPT-5.6 and Kimi fix it). Since these APIs return no usable log-odds, the decision is the first of the two candidate words that appears in the reply. Replies that name neither

word (0.2% for DeepSeek, up to 2.7% for GPT-5.6, none for Kimi) are left out of the decision-change rate. The APIs are not deterministic: sending the identical base set a second time changes 5%, 9% and 4% of decisions for DeepSeek, GPT-5.6 and Kimi. We measured this floor once per model and report it next to their change rates in Tab. 4. PAEN, pHash and SSCD were applied on our side before the images were sent, so the API only ever saw the reduced set. All runs were made on 23–24 September 2026; the three models together cost about 25 US dollars.

A.4 In the wild: naturally occurring duplicates

The re-uploads above are produced by us, even if by realistic chains. Real listings also contain duplicates that nobody injected: of 22,608 listings of the product types above, 10.5% contain a within-listing near-duplicate pair, typically an exact re-upload, the same photo with dimension annotations overlaid, or a duplicated line drawing. We ask whether such a natural duplicate is credited like a distinct view.

Design. Without an injected conflict the model answers with near-certainty, and removing or reversing any image then moves belief by less than the order-noise floor; we therefore keep one injected confusable photo in the set, to hold belief off the ceiling, and inject nothing else. For 600 listings with a natural duplicate pair (a, b) and two further distinct photos (c, d), we compare, at matched image count, replacing the duplicate b by a gray image against replacing the distinct photo c by a gray image. A reader who knows that b is a copy of a loses nothing in the first case. Because these duplicates were found with the same signal PAEN uses, the test measures the weight a model puts on a duplicate, not the estimator's generalization, which is evaluated separately (App. A.2).

Table 6: **Weight of a naturally occurring duplicate.** Change in log-odds (nats) when the content of the natural duplicate, or of a genuinely distinct photo, is replaced by a gray image at fixed image count; paired difference with 95% CI. A rational reader loses nothing in the first column.

Model	Method	duplicate → gray	distinct → gray	difference
Qwen2.5-VL-7B	none	-0.40	-0.71	+0.31 [+0.24, +0.38]
InternVL3-8B	none	-0.72	-0.76	+0.04 [-0.02, +0.11]
Qwen2.5-VL-7B	PAEN	-0.00	-0.69	+0.68 [+0.62, +0.75]
InternVL3-8B	PAEN	-0.33	-0.88	+0.55 [+0.47, +0.64]

Tab. 6 shows that neither architecture treats the natural duplicate as the copy it is. InternVL3 loses as much belief when the duplicate's content is removed as when a distinct photo's is; Qwen2.5-VL discounts the duplicate to roughly half the weight of a genuine view, but not to zero. With PAEN, which removes the duplicate before the forward pass, the picture becomes the rational one: on Qwen2.5-VL removing the duplicate's content costs nothing while removing a genuine view's still costs what it did, and on InternVL3 the small residual in the duplicate column is the cost of the inserted gray image itself, common to both columns.

A.5 The spectrum on other data and protocols

Tab. 7 repeats Tab. 1 with trial-level bootstrap intervals. On COCO cats, four renderings gain +3.6 against +3.0 for four different cats and +1.5 for exact copies. On a twelve-category CO3D spectrum (144 sequences) renderings gain $+2.0$ against $+2.3$ and $+2.4$ for adjacent and far views and $+0.7$ for copies. Under the neutral-centred k=3 protocol on CO3D, renderings gain +4.1, adjacent and far views +4.5 and +4.7, exact copies +0.9, and overlapping crops +2.8. Three prompt framings (object, witness, pet) give the same ordering.

A.6 Neutral insertions and sequence length

Inserting $k \leq 3$ gray/noise images changes the average log-odds by less than 0.4 nats but by up to ± 1.3 depending on position, which is why all results average over positions; with $k = 7$ neutral insertions Qwen2.5-VL flips 54% of decisions, a sequence-length instability separate from the effect we study.

Table 7: **Novelty spectrum with confidence intervals.** Log-odds gain over a single supporting image when four images are added (nats; CO3D, 72 trials; trial-level bootstrap 95% intervals, 2,000 resamples).

Model	DUP	DERIVED	PARTIAL	ADJ	FAR	INDEP
Qwen2.5-VL-7B	+0.10 [-0.37, +0.59]	+3.18 [+2.50, +3.87]	+1.98 [+1.47, +2.53]	+3.59 [+2.88, +4.30]	+3.61 [+2.93, +4.29]	+0.69 [+0.39, +0.99]
InternVL3-8B	+4.07 [+3.72, +4.44]	+5.05 [+4.59, +5.53]	+2.59 [+2.23, +2.97]	+5.76 [+5.31, +6.23]	+5.79 [+5.40, +6.20]	+6.38 [+6.01, +6.80]
Qwen2-VL-7B	+2.27 [+2.06, +2.47]	+3.36 [+3.13, +3.61]	+1.07 [+0.83, +1.32]	+3.24 [+2.93, +3.55]	+3.23 [+2.92, +3.55]	+2.28 [+2.01, +2.55]
LLaVA-OneVision-7B	+1.87 [+1.74, +2.00]	+1.94 [+1.81, +2.07]	+0.26 [+0.12, +0.40]	+2.02 [+1.85, +2.19]	+1.97 [+1.80, +2.14]	+1.48 [+1.31, +1.67]

B ADDITIONAL RESULTS

Tabs. 14–20 report every model/method/dataset combination with accuracy, decision change, Brier, NLL and neutral-centered Δlog-odds.

B.1 FOUR-CANDIDATE AND OPEN-ENDED READOUTS

Binary questions give a clean scalar readout, but they also make the wrong answer the only alternative. We therefore repeat the realistic task in two non-binary forms on the same listings. In the four-candidate form the prompt names the true product type, the confusable type and two unrelated types in random order, and the decision is the highest-scoring of the four. In the open form no candidates are named: the model answers in its own words and the answer is classified as naming the true type, the mis-attributed type, or neither.

Table 8: **Non-binary readouts** (product-type task, no intervention). Four-candidate: accuracy, decision change, and how often the mis-attributed type is the chosen answer. Open: how often the model's free-form answer names the mis-attributed type. PAEN on the four-candidate form: Tab. 21.

	four candidates					open answer: names the wrong type		
	Base	DERIVED$_3$		wrong type chosen		Base	DERIVED$_3$	TRUEDUP$_3$
Model	acc	acc	Δdec.	base	DERIVED$_3$			
Qwen2.5-VL-7B	0.80	0.45	36%	0.19	0.55	0.16	0.40	0.09
InternVL3-8B	0.71	0.51	21%	0.27	0.47	0.23	0.32	0.17

Both forms reproduce the effect (Tab. 8), and the four-candidate form makes its direction explicit: with four answers available, three re-upload copies of the mis-attributed photo turn that type from a minority choice into the plurality one on Qwen2.5-VL, with a larger decision change than in the binary task. The model does not retreat to a vaguer answer when copies arrive; it names the wrong product. The open form agrees: the rate at which the mis-attributed type is named rises on both models, and copies of a genuine photo again push the other way. PAEN returns the four-candidate task to base with about 1% decision change.

B.2 COLOUR QUESTION AND RULE-STATED PROMPT

Tab. 9 gives the same protocol on the ABO colour question (300 listings) and the variant whose prompt states that re-uploads count once.

B.3 ROBUSTNESS OF THE READOUT

The findings do not depend on the evaluation surface: greedy-decoded answers (100-set subsamples per model, substring-parsed) agree with the first-token decision in 88–100% of cases and show the same accuracy drop under renderings (e.g. Qwen2.5-VL, conflict-last position: 0.95$\rightarrow$0.40). A second question framing ("witnesses") reproduces the effect (0.61 under renderings vs. base 0.85). Full per-model tables with confidence intervals and the per-image voting baselines are Tabs. 14–20.

Table 9: **The failure on real product listings** (no intervention; application question; simulated re-upload chains). Columns as in Tab. 2: accuracy of the base set; DERIVED$_3$ (accuracy, decision change); DUP$_3$ (accuracy); TRUEDUP$_3$ and NEW$_3$ (accuracy, Δlog-odds vs. the base set). Last block: the same question with "re-uploads count once" stated.

			DERIVED$_3$		DUP$_3$	TRUEDUP$_3$		NEW$_3$	
Task	Model	Base	acc	Δdec.	acc	acc	ΔLO	acc	ΔLO
product type	Qwen2.5-VL-7B	0.83	0.57	26%	0.68	0.92	+1.1	0.96	+2.1
	InternVL3-8B	0.76	0.57	21%	0.60	0.84	+0.1	0.93	+1.3
	Qwen2-VL-7B	0.80	0.61	20%	0.64	0.89	+1.1	0.92	+1.4
	LLaVA-OneVision-7B	0.71	0.60	12%	0.60	0.79	+0.7	0.84	+1.0
color	Qwen2.5-VL-7B	0.79	0.37	42%	0.51	0.88	+1.4	0.92	+2.1
	InternVL3-8B	0.83	0.65	19%	0.67	0.87	+0.2	0.92	+0.9
product type, question states "re-uploads count once"	Qwen2.5-VL-7B	0.82	0.58	24%	–	0.90	–	–	–
	InternVL3-8B	0.71	0.50	23%	–	0.78	–	–	–
	Qwen2-VL-7B	0.82	0.62	20%	–	0.89	–	–	–
	LLaVA-OneVision-7B	0.72	0.61	12%	–	0.77	–	–	–

B.4 MODEL SCALE

Tab. 10 repeats the realistic protocol of Tab. 2 across the Qwen2.5-VL family from 3B to 72B (bf16, the 72B model sharded over four GPUs), with the same 312 listings, conditions and $\geq$5-image filter. The failure does not shrink with scale: three re-upload copies of the mis-attributed photo change 29%, 26%, 43% and 24% of decisions at 3B, 7B, 32B and 72B, and copies of a genuine photo still raise its log-odds at every size. PAEN returns every size to its base accuracy with at most 1% decision change.

Table 10: **The realistic task across model sizes** (Qwen2.5-VL family; 312 listings $\times$ three conflict positions; columns as in Tab. 4, plus the log-odds shift under copies of a genuine photo).

			DERIVED$_3$		TRUEDUP$_3$		NEW$_3$
Model	Method	Base	acc	Δdec.	Δdec.	ΔLO	acc
Qwen2.5-VL-3B	none	0.81	0.53	29%	9%	+0.6	0.94
	pHash de-dup	0.81	0.63	17%	5%	+0.3	0.94
	PAEN	0.81	**0.80**	**0%**	1%	+0.0	0.93
Qwen2.5-VL-7B	none	0.83	0.57	26%	9%	+1.1	0.96
	pHash de-dup	0.83	0.71	12%	6%	+0.5	0.95
	PAEN	0.82	**0.82**	**0%**	1%	+0.0	0.95
Qwen2.5-VL-32B	none	0.85	0.42	43%	9%	+1.8	0.97
	pHash de-dup	0.85	0.67	18%	4%	+0.9	0.97
	PAEN	0.84	**0.84**	**0%**	0%	+0.1	0.97
Qwen2.5-VL-72B	none	0.95	0.71	24%	3%	+0.7	0.98
	pHash de-dup	0.95	0.83	12%	2%	+0.8	0.98
	PAEN	0.95	**0.95**	**0%**	0%	+0.0	0.98

B.5 LLAVA-ONEVISION

LLaVA-OneVision does not discriminate on the spectrum: exact copies, renderings and new viewpoints all gain about the same (Tab. 1), so it has no implicit copy discount to begin with. On it the text side is the effective lever and cross-image masking adds nothing beyond it (Tab. 3). Its answer-position attention probe is omitted: the model emits 1,485 visual tokens per image regardless of input size, so materialising its full attention for five images exceeds memory. We therefore describe the mechanism of Sec. 3 as a mutually reinforcing visual evidence cluster formed in the middle layers of the architectures that discriminate between copies and viewpoints, and note that a model which never separates the two does not form such a cluster in the first place. The failure and its repair are unaffected: on the realistic task LLaVA-OneVision loses 0.71$\rightarrow$0.60 under three

re-upload copies and PAEN returns it to 0.70 (Tab. 4), since removal does not depend on where the amplification is built.

B.6 MASKING INSTEAD OF REMOVAL

The mechanism of Sec. 3 suggests an in-model alternative to removal: keep every image and scale the attention between dependent images by the estimate. For queries that are tokens of image i, keys of image j receive a logit bias

$$b_{i\to j} = \log(1 - d_{ij}), \tag{2}$$

which scales the attention weight between dependent images by $(1 - d_{ij})$, with full masking at $d_{ij} = 1$; for text queries, keys of image j receive

$$b_{\cdot\to j} = \log w_j, \qquad w_j = \frac{1}{1 + \sum_{k\neq j} d_{jk}}, \tag{3}$$

so that a cluster of m mutually dependent renderings carries the text-side weight of one image. We call this rule *masking*; a *hybrid* removes images with $d_{ij} \geq 0.9$ and masks the remaining pairs. Both are training-free and add only an additive attention bias.

Table 11: **Signal versus aggregation rule, controlled setting** (synthetic label-preserving renderings; CO3D: 179 sets; ABO: 258 sets for product type and 273 for colour; renderings sampled from the full label-preserving family). Accuracy after adding three renderings of the mis-attributed photo; base accuracy in parentheses where a method changes it. Similarity-based de-duplication (DINOv2) merges genuinely different views and lowers base accuracy; the provenance signal of Eq. 1 restores accuracy whether it is used to remove (PAEN), to mask, or both (hybrid).

Domain	Model	none	pHash	DINO de-dup	masking	**PAEN**	hybrid
CO3D, category	Qwen2.5-VL-7B	0.67	0.78	0.77 (0.85)	0.84	**0.88**	0.88
	InternVL3-8B	0.63	0.79	0.83 (0.90)	0.87	**0.92**	0.95
ABO, product type	Qwen2.5-VL-7B	0.55	0.71	0.79	0.75	**0.83**	0.83
	InternVL3-8B	0.55	0.71	0.77	**0.80**	0.79	0.79
ABO, color	Qwen2.5-VL-7B	0.51	0.69	–	0.78	**0.84**	–
	InternVL3-8B	0.71	0.82	–	**0.90**	0.88	–

Table 12: **Removal versus masking on the novelty spectrum** (Qwen2.5-VL-7B, three additions, 24 trials per cell; gain over a single supporting image in nats, estimated dependence). Masking overcorrects on copies and renderings and costs more on adjacent viewpoints; on overlapping crops the two rules agree.

condition	none	masking	**PAEN**
DUP$_3$	+0.31	−0.42	+0.00
DERIVED$_3$	+3.56	−0.58	+0.00
PARTIAL$_2$	+2.22	+0.75	+0.62
ADJ$_3$	+3.92	+3.49	+3.82
FAR$_3$	+4.12	+4.06	+4.12

Tab. 11 separates the two ingredients of PAEN. Holding the aggregation rule fixed and changing the signal, similarity-based de-duplication (DINOv2, like CLIP in Sec. 2.3) merges genuinely different views (base accuracy drops to 0.85/0.90) while still leaking renderings, and pHash leaks geometric renderings; the provenance signal does neither. Holding the signal fixed and changing the rule, all three rules restore accuracy in the controlled setting. Masking matches removal within two points on InternVL3's ABO sets and trails it by four to eight points on the others, and on the simulated re-upload chains of Sec. 5.1 it recovers only part of the loss (Tab. 14). Together with Tab. 3 (text-side re-weighting inert, image→image masking decisive), this is the evidence that the failure is repaired by knowing provenance, not by any particular way of acting on it.

Partial overlap. Overlapping crops are the one condition in which removal discards evidence: on PARTIAL$_2$ (two crops of one photograph, each showing 60% of it) PAEN drops the second crop in 11 of 24 trials and with it the strip only that crop shows. An oracle input that keeps only that strip (the overlap painted gray through the RANSAC homography) gains +1.77 nats over a single image; no intervention scores +2.21, removal +0.62 and whole-image masking +0.75, so both rules over-correct here by about one nat. We therefore keep removal and record partial overlap as its known cost.

Tab. 12 shows why masking has no regime of its own. On exact copies and renderings it over-corrects, pushing the gain below the single-image value, whereas removal returns it to exactly zero. On overlapping crops, the case a soft rule is designed for, the two are within 0.15 nats of each other, removal keeping on average 2.5 of the three crops. On adjacent viewpoints masking costs more. Screenshots and preview crops in the re-upload chains give partial geometric matches, and a partial mask leaves part of the reinforcement in place; when the estimate is clear-cut, removal is the right regime.

B.7 PER-TRANSFORMATION BREAKDOWN

Table 13: Single renderings ($k = 1$; 72 decisions per cell): flip rate / Δ log-odds. Original model vs. masking vs. pHash de-duplication (Qwen2.5-VL-7B, CO3D). PAEN is exact on every family that the estimator flags.

Rendering family	original	masking	pHash
photometric (resize, JPEG, brightness, contrast, blur, gray)	0.21 / +2.6	0.02 / +0.6	0.00 / 0.0
geometric (crop, zoom, pan, corner crop, rot. 5°, persp.)	0.17 / +2.5	0.03 / +0.7	0.12 / +1.8
flip	0.24 / +2.8	0.01 / +0.6	0.24 / +2.8

Tab. 13 isolates single renderings ($k = 1$). Every rendering family, photometric, geometric or flip, moves the original model by +2.5 to +2.8 nats and flips 17–24% of decisions. Masking brings all of them to $\leq$+0.8 and PAEN to zero on every family the estimator flags; perceptual hashing is perfect on photometric renderings but leaks geometric ones and flips. The repair is therefore not specific to any rendering family.

Table 14: ABO listings, product type: realistic setting (312 sets; application prompt, real re-upload chains): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.83	0.57 [0.52,0.61]	0.26	0.333	1.12	-1.63	0.96
Qwen2.5-VL-7B	pHash de-dup	0.83	0.71 [0.67,0.75]	0.12	0.219	0.72	-0.53	0.95
Qwen2.5-VL-7B	masking (ablation)	0.83	0.62 [0.58,0.66]	0.21	0.285	0.94	-1.23	0.96
Qwen2.5-VL-7B	PAEN	0.82	0.82 [0.79,0.85]	0.00	0.126	0.42	+0.46	0.95
InternVL3-8B	none	0.76	0.57 [0.52,0.61]	0.21	0.338	1.26	-1.35	0.93
InternVL3-8B	pHash de-dup	0.76	0.67 [0.63,0.72]	0.10	0.251	0.93	-0.45	0.92
InternVL3-8B	masking (ablation)	0.76	0.67 [0.63,0.71]	0.12	0.248	0.89	-0.51	0.93
InternVL3-8B	PAEN	0.76	0.76 [0.72,0.80]	0.00	0.181	0.68	+0.55	0.92
Qwen2-VL-7B	none	0.80	0.61 [0.57,0.64]	0.20	0.313	1.14	-1.37	0.92
Qwen2-VL-7B	pHash de-dup	0.80	0.70 [0.67,0.73]	0.10	0.228	0.77	-0.57	0.91
Qwen2-VL-7B	masking (ablation)	0.80	0.67 [0.64,0.70]	0.13	0.256	0.88	-0.73	0.90
Qwen2-VL-7B	PAEN	0.80	0.80 [0.77,0.83]	0.01	0.146	0.48	+0.09	0.91
LLaVA-OneVision-7B	none	0.71	0.60 [0.56,0.64]	0.12	0.301	1.01	-1.01	0.84
LLaVA-OneVision-7B	pHash de-dup	0.71	0.67 [0.63,0.71]	0.04	0.255	0.85	-0.36	0.83
LLaVA-OneVision-7B	masking (ablation)	0.71	0.70 [0.66,0.74]	0.10	0.213	0.70	-0.34	0.83
LLaVA-OneVision-7B	PAEN	0.70	0.70 [0.66,0.74]	0.01	0.225	0.76	-0.05	0.83

Table 15: ABO listings, color: realistic setting (300 sets): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.79	0.37 [0.32,0.43]	0.42	0.454	1.43	-2.18	0.92
Qwen2.5-VL-7B	pHash de-dup	0.79	0.63 [0.58,0.67]	0.16	0.262	0.84	-0.87	0.92
Qwen2.5-VL-7B	masking (ablation)	0.79	0.47 [0.42,0.52]	0.33	0.376	1.19	-1.71	0.92
Qwen2.5-VL-7B	PAEN	0.78	0.78 [0.74,0.81]	0.01	0.162	0.54	+0.14	0.92
InternVL3-8B	none	0.83	0.65 [0.61,0.69]	0.19	0.265	0.98	-1.09	0.92
InternVL3-8B	pHash de-dup	0.83	0.77 [0.74,0.81]	0.06	0.180	0.75	-0.04	0.91
InternVL3-8B	masking (ablation)	0.83	0.74 [0.70,0.78]	0.11	0.202	0.74	-0.43	0.91
InternVL3-8B	PAEN	0.82	0.82 [0.79,0.86]	0.00	0.138	0.63	+0.90	0.91

Table 16: CO3D multi-view, model-independent construction (179 sets; main CO3D task): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.90	0.67 [0.62,0.71]	0.29	0.256	0.91	-1.83	0.91
Qwen2.5-VL-7B	pHash de-dup	0.90	0.78 [0.73,0.82]	0.13	0.176	0.58	-0.55	0.91
Qwen2.5-VL-7B	CLIP de-dup (τ=.9)	0.79	0.62 [0.56,0.68]	0.17	0.290	1.03	-0.72	0.85
Qwen2.5-VL-7B	DINOv2 de-dup	0.85	0.77 [0.72,0.81]	0.08	0.192	0.65	-0.29	0.91
Qwen2.5-VL-7B	masking (ablation)	0.88	0.84 [0.81,0.88]	0.09	0.127	0.42	-0.41	0.89
Qwen2.5-VL-7B	PAEN	0.89	0.88 [0.85,0.92]	0.00	0.108	0.39	+0.42	0.91
InternVL3-8B	none	0.94	0.63 [0.58,0.69]	0.31	0.254	0.89	-2.03	0.96
InternVL3-8B	pHash de-dup	0.94	0.79 [0.74,0.84]	0.15	0.152	0.57	-0.66	0.96
InternVL3-8B	CLIP de-dup (τ=.9)	0.82	0.70 [0.64,0.76]	0.12	0.229	0.83	-0.77	0.90
InternVL3-8B	DINOv2 de-dup	0.90	0.83 [0.79,0.88]	0.07	0.137	0.52	-0.27	0.94
InternVL3-8B	masking (ablation)	0.93	0.87 [0.83,0.91]	0.07	0.102	0.38	-0.12	0.94
InternVL3-8B	PAEN	0.93	0.92 [0.88,0.95]	0.00	0.075	0.35	+0.64	0.95
Qwen2-VL-7B	none	0.89	0.69 [0.66,0.72]	0.19	0.254	0.97	-1.93	0.96
Qwen2-VL-7B	pHash de-dup	0.89	0.78 [0.75,0.81]	0.11	0.164	0.57	-0.87	0.96
Qwen2-VL-7B	masking (ablation)	0.88	0.76 [0.74,0.79]	0.11	0.171	0.55	-0.41	0.93
Qwen2-VL-7B	PAEN	0.88	0.87 [0.85,0.90]	0.00	0.095	0.31	-0.02	0.95
LLaVA-OneVision-7B	none	0.79	0.67 [0.63,0.70]	0.12	0.254	0.89	-1.43	0.90
LLaVA-OneVision-7B	pHash de-dup	0.79	0.73 [0.70,0.77]	0.05	0.202	0.71	-0.55	0.90
LLaVA-OneVision-7B	masking (ablation)	0.78	0.77 [0.74,0.81]	0.04	0.173	0.60	-0.43	0.86
LLaVA-OneVision-7B	PAEN	0.79	0.78 [0.76,0.81]	0.00	0.165	0.59	+0.04	0.88

Table 17: ABO listings, category (106 sets after requiring $\geq$5 catalog images per listing): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.87	0.60 [0.53,0.67]	0.28	0.301	1.04	-1.84	0.98
Qwen2.5-VL-7B	pHash de-dup	0.87	0.77 [0.71,0.83]	0.10	0.173	0.61	-0.40	0.97
Qwen2.5-VL-7B	masking (ablation)	0.87	0.80 [0.75,0.86]	0.07	0.138	0.43	-0.02	0.97
Qwen2.5-VL-7B	PAEN	0.87	0.87 [0.82,0.92]	0.00	0.098	0.34	+0.90	0.96
Qwen2.5-VL-7B	per-image soft vote (mean log-odds)	0.92	0.45 [0.36,0.55]	0.47	0.429	1.53	-2.49	0.99
Qwen2.5-VL-7B	per-image soft vote + pHash	0.92	0.74 [0.65,0.81]	0.18	0.189	0.63	-0.66	0.99
Qwen2.5-VL-7B	per-image majority vote	0.92	0.19	–	–	–	–	0.97
Qwen2.5-VL-7B	per-image majority vote + pHash	0.90	0.39	–	–	–	–	0.98
InternVL3-8B	none	0.79	0.63 [0.54,0.71]	0.17	0.305	1.13	-1.62	0.94
InternVL3-8B	pHash de-dup	0.79	0.72 [0.65,0.80]	0.07	0.227	0.84	-0.43	0.94
InternVL3-8B	masking (ablation)	0.79	0.79 [0.71,0.85]	0.06	0.152	0.49	+0.03	0.94
InternVL3-8B	PAEN	0.79	0.79 [0.72,0.85]	0.00	0.169	0.65	+0.51	0.93
Qwen2-VL-7B	none	0.81	0.69 [0.63,0.75]	0.12	0.252	1.03	-1.21	0.93
Qwen2-VL-7B	pHash de-dup	0.81	0.74 [0.69,0.80]	0.07	0.193	0.73	-0.56	0.92
Qwen2-VL-7B	masking (ablation)	0.82	0.75 [0.70,0.80]	0.09	0.181	0.62	-0.28	0.91
Qwen2-VL-7B	PAEN	0.81	0.81 [0.76,0.86]	0.00	0.128	0.43	+0.06	0.91
LLaVA-OneVision-7B	none	0.81	0.72 [0.65,0.78]	0.10	0.219	0.74	-1.25	0.89
LLaVA-OneVision-7B	pHash de-dup	0.81	0.77 [0.71,0.82]	0.04	0.179	0.59	-0.63	0.89
LLaVA-OneVision-7B	masking (ablation)	0.82	0.81 [0.75,0.86]	0.04	0.152	0.50	-0.65	0.91
LLaVA-OneVision-7B	PAEN	0.81	0.81 [0.75,0.86]	0.00	0.147	0.48	-0.12	0.88

Table 18: ABO listings, category, scaled (258 sets after the ≥5-image protocol filter): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.84	0.55 [0.50,0.59]	0.30	0.347	1.24	-1.63	0.96
Qwen2.5-VL-7B	pHash de-dup	0.84	0.71 [0.67,0.75]	0.13	0.220	0.81	-0.34	0.95
Qwen2.5-VL-7B	CLIP de-dup (τ=.9)	0.81	0.78 [0.74,0.81]	0.03	0.174	0.64	+0.47	0.91
Qwen2.5-VL-7B	DINOv2 de-dup	0.84	0.79 [0.75,0.83]	0.05	0.163	0.61	+0.43	0.94
Qwen2.5-VL-7B	masking (ablation)	0.84	0.75 [0.72,0.79]	0.10	0.185	0.64	-0.05	0.94
Qwen2.5-VL-7B	PAEN	0.84	0.83 [0.80,0.87]	0.00	0.135	0.53	+0.76	0.94
Qwen2.5-VL-7B	hybrid (ablation)	0.84	0.83 [0.80,0.86]	0.01	0.139	0.54	+0.72	0.94
InternVL3-8B	none	0.80	0.55 [0.50,0.61]	0.25	0.339	1.26	-1.67	0.92
InternVL3-8B	pHash de-dup	0.79	0.71 [0.66,0.75]	0.09	0.230	0.89	-0.52	0.92
InternVL3-8B	CLIP de-dup (τ=.9)	0.76	0.73 [0.68,0.78]	0.03	0.205	0.81	+0.17	0.88
InternVL3-8B	DINOv2 de-dup	0.79	0.77 [0.72,0.81]	0.03	0.186	0.74	+0.22	0.91
InternVL3-8B	masking (ablation)	0.79	0.80 [0.76,0.84]	0.07	0.162	0.61	+0.08	0.91
InternVL3-8B	PAEN	0.79	0.79 [0.75,0.83]	0.00	0.161	0.65	+0.52	0.91
InternVL3-8B	hybrid (ablation)	0.79	0.79 [0.75,0.83]	0.00	0.161	0.65	+0.46	0.91

Table 19: ABO listings, color (273 sets after the ≥5-image protocol filter): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.86	0.51 [0.46,0.57]	0.35	0.346	1.11	-1.52	0.96
Qwen2.5-VL-7B	pHash de-dup	0.86	0.69 [0.65,0.74]	0.16	0.208	0.68	-0.39	0.95
Qwen2.5-VL-7B	masking (ablation)	0.85	0.78 [0.74,0.83]	0.08	0.156	0.50	-0.02	0.95
Qwen2.5-VL-7B	PAEN	0.85	0.84 [0.80,0.88]	0.01	0.117	0.41	+0.82	0.95
InternVL3-8B	none	0.89	0.71 [0.67,0.76]	0.18	0.205	0.72	-0.94	0.94
InternVL3-8B	pHash de-dup	0.88	0.82 [0.78,0.86]	0.07	0.141	0.56	-0.12	0.93
InternVL3-8B	masking (ablation)	0.89	0.90 [0.87,0.93]	0.05	0.084	0.32	+0.32	0.94
InternVL3-8B	PAEN	0.88	0.88 [0.85,0.91]	0.00	0.096	0.43	+0.90	0.93

Table 20: CO3D, near-duplicate true views (boundary regime, 116 sets; the two true views are adjacent or orbit-closing frames of one capture): base accuracy; accuracy under three renderings of the mis-attributed photo (with trial-level bootstrap 95% CI), decision change, Brier, NLL, Δlog-odds (neutral-centered); accuracy under three new views.

Model	Method	Base	DERIVED$_3$ acc [CI]	Δdec.	Brier	NLL	ΔLO	NEW$_3$
Qwen2.5-VL-7B	none	0.94	0.70 [0.64,0.75]	0.26	0.223	0.77	-1.79	0.92
Qwen2.5-VL-7B	pHash de-dup	0.93	0.80 [0.75,0.85]	0.13	0.156	0.51	-0.56	0.92
Qwen2.5-VL-7B	masking (ablation)	0.91	0.85 [0.81,0.89]	0.07	0.114	0.37	-0.37	0.92
Qwen2.5-VL-7B	masking v2 (ablation)	0.90	0.85 [0.80,0.89]	0.07	0.117	0.37	-0.35	0.91
Qwen2.5-VL-7B	PAEN	0.90	0.89 [0.85,0.93]	0.00	0.105	0.36	+0.43	0.92
Qwen2.5-VL-7B	hybrid (ablation)	0.90	0.89 [0.84,0.93]	0.01	0.104	0.35	+0.37	0.92
InternVL3-8B	none	0.97	0.70 [0.64,0.76]	0.27	0.219	0.73	-2.13	0.97
InternVL3-8B	pHash de-dup	0.96	0.88 [0.82,0.92]	0.08	0.107	0.38	-0.57	0.97
InternVL3-8B	masking (ablation)	0.95	0.89 [0.84,0.93]	0.07	0.075	0.26	-0.14	0.97
InternVL3-8B	masking v2 (ablation)	0.95	0.89 [0.84,0.94]	0.07	0.071	0.25	-0.12	0.96
Qwen2-VL-7B	none	0.89	0.71 [0.68,0.74]	0.18	0.252	0.95	-1.99	0.96
Qwen2-VL-7B	pHash de-dup	0.89	0.80 [0.77,0.83]	0.09	0.152	0.49	-0.88	0.96
Qwen2-VL-7B	masking (ablation)	0.87	0.76 [0.73,0.78]	0.12	0.173	0.54	-0.39	0.93
Qwen2-VL-7B	masking v2 (ablation)	0.86	0.76 [0.73,0.78]	0.11	0.171	0.53	-0.37	0.92

Table 21: Four-candidate readout on the realistic ABO task (312 listings × 3 conflict positions): the prompt names the true product type, the confusable type and two unrelated types in random order.

Model	Method	Base acc	DERIVED$_3$ acc	Δdec.	wrong-type (base)	wrong-type (DERIVED$_3$)	TRUEDUP$_3$ acc	NEW$_3$ acc
Qwen2.5-VL-7B	no intervention	0.80	0.45	0.36	0.19	0.55	0.89	0.95
Qwen2.5-VL-7B	PAEN	0.78	0.77	0.01	0.21	0.22	0.78	0.94
InternVL3-8B	no intervention	0.71	0.51	0.21	0.27	0.47	0.79	0.89
InternVL3-8B	PAEN	0.73	0.73	0.01	0.27	0.27	0.73	0.90