ANONYMOUS IMAGE PROCESSING-CHAIN FORENSICS BY EXAMPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering an image's processing history is a central problem in digital image forensics. Existing methods recover named operation sequences or compare processing traces. When operation names and configurations are unavailable, a few before-and-after examples provide another source of evidence. We introduce anonymous image processing-chain inference, a new task for ranking candidate histories between two image versions using these examples. Our central finding is that a history decision can be determined before its operations are fully identified. We derive sufficient conditions under which all example-compatible models in a stated parameter neighborhood prefer the same history, accounting for repeated operations and propagated estimation errors. We propose ChainPrint, which estimates reusable operation functions from paired examples within a declared model library and composes them to address diverse and compound history-inference problems without task-specific retraining. Shuffling example pairings degrades recovery; replacing sequential composition with additive operation changes lowers eight-step AI Top-1 accuracy by 41.4 percentage points. Across three image sources, eight-step Top-1 reaches 100.0% for conventional histories and 77.8% for AI-containing histories, with eight examples per operation and sixteen candidates. A fully enumerated two-dimensional control shows that ranking stability depends on which parameter directions the examples constrain, even when example fit and overall operation disagreement are matched. These results focus history recovery on the operation knowledge needed to distinguish competing histories, making example-based inference possible without requiring complete operation identification.

1 INTRODUCTION

An image may be edited by a generative model (Brooks et al., 2023), resized for publication, and repeatedly compressed between platforms. Its *processing history* is the ordered sequence of operations connecting earlier and later versions. Investigators may need to distinguish routine redistribution from an intervening content edit. The Coalition for Content Provenance and Authenticity (C2PA) standard binds signed creation and editing records to content, but these records may omit processing steps (Coalition for Content Provenance and Authenticity). Image-based history recovery complements them by testing explanations of undocumented changes against available versions.

Prior forensic work recovers named operation sequences (Li et al., 2023), estimates editing transformations from image pairs (Hu et al., 2013), and compares processing traces, including those unseen during training (Mayer & Stamm, 2020). Learned editing representations (Wu et al., 2026) and transformations specified through examples (Hertzmann et al., 2001) also characterize image changes. Yet examples may not determine operation behavior on intermediate images of a long history. Earlier operations change later inputs, and later effects may obscure earlier traces (Chen et al., 2022; Verde et al., 2023). Models that agree on isolated examples can therefore predict different outcomes when composed. We ask: **what must operation examples determine to distinguish candidate histories?**

We introduce *anonymous image processing-chain inference*, a new task that ranks candidate histories from two image versions and operation examples. Each operation has a few before-and-after pairs generated with the same settings and implementation; its name and configuration are withheld. Local identifiers, such as A and B, link examples to histories. Comparing $A \to B \to A$ with $A \to A \to B$

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.

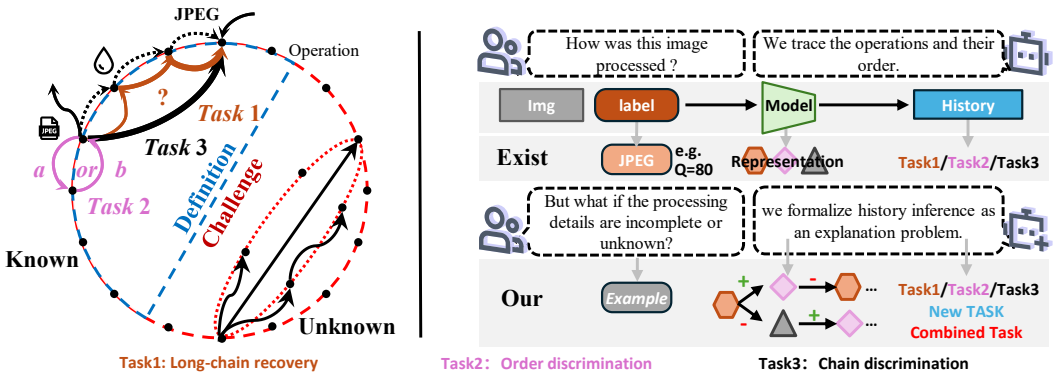

Figure 1: Anonymous image processing-chain inference. Examples specify reusable operations for ranking candidate histories. Long-chain recovery, order discrimination, and chain discrimination are representative queries; new and combined queries are specified through candidate histories.

tests the order of the same demonstrated operations. Candidates can mix conventional processing with local AI editing. Complete-sequence demonstrations are not required.

Our central finding is that **a history decision can be determined before its operations are fully identified**. Several operation estimates may explain the same examples and predict different endpoints, yet favor the same history. The deciding quantity is the difference between candidates' prediction errors. We identify shared-composition conditions under which changes in the operation estimates shift every candidate's squared prediction error equally, preserving all such differences. Beyond this exact invariance, we characterize stability within a stated neighborhood of the fitted operation estimates. The preferred history remains unchanged when the uncertainty allowed by the examples is sufficiently constrained in directions that affect candidate comparisons. We further establish when that stable preference identifies the history that generated the observation. Thus history recovery can require less operation knowledge than full operation identification.

We propose **ChainPrint**, which makes this operation knowledge reusable by estimating an *image-to-image function* for each operation within a declared model library. Held-out prediction errors guide selection among fitted estimates, refinements, and shared corrections. Each function acts on the image predicted by preceding functions and is reused across candidates and repeated occurrences of its operation. Sharing couples estimation changes across candidate predictions, supplying the structure studied in our analysis. With the functions fixed, new candidate histories define new and compound forensic questions without further fitting.

We evaluate ChainPrint across three image sources on conventional and AI-containing histories, testing longer compositions, reordered and repeated operations, configuration changes, and transfer. With eight examples per operation and sixteen candidates, it reaches 100.0% Top-1 on conventional eight-step histories and 77.8% on histories containing a local AI edit.

Our contributions are threefold.

(1) **A new task: anonymous image processing-chain inference.** We formulate history recovery from two image versions and a few paired examples per operation, with names and configurations withheld at inference. Candidates can combine conventional processing with local AI editing and reuse operations.

(2) **ChainPrint: composing reusable operation functions.** The same fitted functions support diverse and compound history questions. Replacing sequential composition with additive operation changes lowers eight-step AI Top-1 by 41.4 percentage points. Against joint fitting with the same prior, ChainPrint reduces setup cost by $5.30\times$ on six fixed timing sources.

(3) **History decisions before full operation identification.** We derive shared-composition conditions that preserve candidate loss differences, and conditions for stable history decisions and correct recovery. A fully enumerated directional control isolates which uncertainty matters. Four example-compatible configurations of one real AI operation provide an image-based instance: they produce differing effects while preserving correct decisions in the evaluated eight-step comparisons.

2 Related Work

Processing histories and provenance. Processing-history recovery includes detectors for JPEG–resampling–JPEG (Chen et al., 2017) and sequence models that do not treat each complete chain as a separate class (Li et al., 2023). Inverse editing infers semantic histories from original and edited image pairs (Hu et al., 2013). Provenance analysis retrieves related images and reconstructs ancestry (Moreira et al., 2018); pairwise analysis combines image comparisons and forensic tools to infer manipulation histories (Zhang et al., 2020).

Forensic representations and example-defined transformations. Camera fingerprints characterize acquisition traces (Cozzolino & Verdoliva, 2020); forensic similarity also compares unseen traces and parameter differences (Mayer & Stamm, 2020). Editprint learns editing fingerprints from simulated histories for several forensic tasks (Wu et al., 2026). Image Analogies learns transformations from paired examples (Hertzmann et al., 2001); visual prompting and Painter specify prediction tasks through examples (Bar et al., 2022; Wang et al., 2023), while ReEdit transfers exemplified edits (Srivastava et al., 2025). ChainPrint composes functions estimated from examples and compares their predicted endpoints. We study whether uncertainty in these transformations can change history comparisons.

3 Anonymous Image Processing-Chain Inference

We rank candidate operation sequences connecting two observed image versions. Each of G operation instances has a local identifier $g \in \{1, \dots, G\}$ and K before-and-after examples:

$$S_g = \{(x_{gk}^-, x_{gk}^+)\}_{k=1}^K, \qquad x_{gk}^+ = T_g(x_{gk}^-), \qquad \mathcal{S} = \{S_g\}_{g=1}^G. \tag{1}$$

T_g generates these pairs with fixed settings and implementation; its name and true configuration are withheld. The identifier links examples to history occurrences without a semantic class label. Conventional processing and AI editing share this interface; editing configurations fix prompts, masks, and randomness. Inference receives S_g and may use declared model priors, which can include the generating configuration without revealing its correspondence to g.

A query contains an earlier image x, a later image y, and a finite candidate set $\mathcal{H}$. Each history $h = (h_1, \dots, h_{L_h})$ lists L_h local identifiers in execution order, defining true composition T_h. A scoring rule s ranks histories by agreement with the observed change:

$$T_h = T_{h_{L_h}} \circ \cdots \circ T_{h_1}, \qquad \widehat{h} = \arg\max_{h \in \mathcal{H}} s(x, y, h \mid \mathcal{S}). \tag{2}$$

Each query is ranked independently. In (A, B, A), the same operation A is applied before and after B. With y fixed, choosing x as its immediate predecessor, an intermediate version, or the original image asks for one-step, partial-chain, or full-history recovery. The same functions can serve these spans while operation configurations remain fixed.

Candidate histories define the question. Changing order tests ordering; changing operations at selected positions tests missing-step hypotheses. Comparing (A, B, C), (A, B, A, C), and (A, C, A, B) jointly tests length, repetition, and order through the same ranking interface. Investigators may obtain candidates from workflow records or hypotheses and use operation examples when the original tools are unavailable.

When $\mathcal{H}$ contains the recorded history, we measure its rank; otherwise, we compare recommended and recorded histories' effects on independent images. Behavioral similarity measures usefulness, not historical identity. Example and query content are disjoint in the benchmark. Appendices A and H give the information boundary and metrics.

4 ChainPrint

ChainPrint estimates one executable function $\widetilde{f}_g$ from each operation's examples S_g, then fixes these functions for history inference (Figure 2). Each candidate composes them on the earlier image x and compares its endpoint with y. Sharing functions couples changes across candidate predictions, which need not change the preferred history.

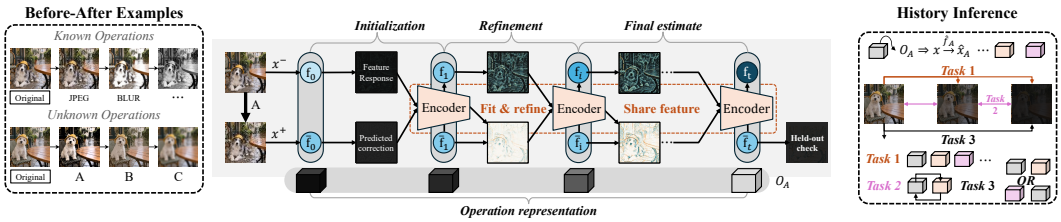

Figure 2: ChainPrint overview. Successive stages estimate one operation from before-and-after examples; cubes denote the selected model and its parameters. Nodes labeled Encoder summarize fitting and selection; pretrained weights and feature definitions remain fixed. Held-out checks guide the selection stages. The final function is reused to compare candidate histories. Feature-response and correction maps are schematic.

4.1 Estimating Reusable Operations

Our declared model library contains conventional processing models and fixed local-AI editing configurations. Conventional proposals are fitted to individual example pairs; editor configurations add further proposals. Among these functions $\mathcal{F}_g(S_g)$, we select the one with the lowest mean prediction error:

$$\widehat{f}_g^{\,0} = \arg\min_{f \in \mathcal{F}_g(S_g)} \frac{1}{K} \sum_{k=1}^{K} d_8(f(x_{gk}^-), x_{gk}^+), \qquad d_8(a,b) = \frac{\|255(a-b)\|_2^2}{3HW}. \tag{3}$$

Images are aligned $H \times W$ RGB arrays in $[0,1]$; d_8 is mean squared error (MSE) in 8-bit units. An editor proposal must produce every output required by a fit to be eligible. We use a frozen Stable Diffusion v1.5 inpainting model (Rombach et al., 2022); each configuration fixes its prompt, mask, and seed. The seed is reset on every application, defining a reusable function of the current image.

Refinement and correction are selected together. Refinement changes the base model; correction estimates recurring residual errors. For $K \geq 2$, rule m constructs $f_{g,m}^{(-k)}$ without pair k or any proposal fitted from it. Its mean held-out error is $V_g(m) = \frac{1}{K}\sum_{k=1}^{K} d_8\left(f_{g,m}^{(-k)}(x_{gk}^-), x_{gk}^+\right)$. Correction competes after initialization and during blockwise quantizer refinement; local filters then compete uncorrected with the complete current model. Each stage retains that model unless an alternative improves on its saved mean held-out error.

For base function $\widehat{f}_g$, stack a constant, predicted RGB values, and a 3×3 neighborhood of $x - \widehat{f}_g(x)$ per pixel into fixed features $\Phi_g(x) \in \mathbb{R}^{HW \times 31}$. Let $R_{gk} \in \mathbb{R}^{HW \times 3}$ stack residuals $255(x_{gk}^+ - \widehat{f}_g(x_{gk}^-))$. We fit $B_g \in \mathbb{R}^{31 \times 3}$ across pixels and examples:

$$\begin{aligned} B_g &= \arg\min_B \sum_{k=1}^{K} \|\Phi_g(x_{gk}^-)B - R_{gk}\|_F^2, \\ \widetilde{f}_g(x) &= Q_8\left(\widehat{f}_g(x) + \Phi_g(x)B_g/255\right), \end{aligned} \tag{4}$$

where the residual is reshaped to the image and Q_8 clips and rounds to the normalized 8-bit grid. The shared B_g gives input-dependent corrections, not a stored image. Correction is fitted within each held-out fold. Using all examples, we refit accepted block models and corrections and select pooled and local base models from existing proposals. The final $\widetilde{f}_g$ is fixed for inference; at $K = 1$, we retain the uncorrected initial model.

Fitting and selection use only operation examples; query endpoints and candidate histories are excluded. Held-out losses guide selection, not independent generalization estimates. Editor weights and feature definitions stay fixed (complete fitting and acceptance rules: Appendix D).

4.2 Composing and Ranking Histories

For each candidate, every function acts on the preceding predicted image:

$$\widehat{x}_{h,0} = x, \qquad \widehat{x}_{h,t} = \widetilde{f}_{h_t}(\widehat{x}_{h,t-1}), \qquad s(x,y,h \mid \mathcal{S}) = -\frac{d_8(\widehat{x}_{h,L_h}, y)}{255^2}. \tag{5}$$

Candidates are ranked by decreasing score. For (A, B, A), $\widetilde{f}_A(\widetilde{f}_B(\widetilde{f}_A(x)))$ applies the same A to different intermediate images. Functions are shared across candidates, so new candidate questions require no further fitting when examples stay fixed. Common prefixes within a query reuse cached predictions. Editor refusals leave histories extending that prefix without predictions. ChainPrint ranks available candidates and abstains if none remain; primary evaluation assigns incomplete queries zero credit in the original denominator (Section 5.1).

4.3 When Operation Examples Determine a History

History comparison depends on how operation uncertainty changes relative losses. Changes to shared operation estimates can move candidate predictions together; endpoint variation alone need not change the preferred history.

For a fixed query and finite candidate set, let $\mathfrak{F}_{\mathcal{S}}$ be a nonempty set of shared deterministic function collections $F = (f_1, \ldots, f_G)$ from a declared family that meet specified operation-example error tolerances. Assume every candidate is executable for every F in this set. Its endpoint is $z_h(F) = (f_{h_{L_h}} \circ \cdots \circ f_{h_1})(x)$ and its normalized loss is $\ell_h(F; y) = \|z_h(F) - y\|_2^2/D$, with $D = 3HW$. The theory compares these alternatives; the estimator returns one collection.

Uncertainty that cancels in comparison. Choose $F^0 \in \mathfrak{F}_{\mathcal{S}}$, set $z_h^0 = z_h(F^0)$, and let $\mathcal{V}_x = \operatorname{span}\{z_h^0 - z_k^0 : h, k \in \mathcal{H}\}$ collect candidate-difference directions. All pairwise loss differences remain unchanged if

$$z_h(F) = z_h^0 + c(F) \quad \text{for all } h,\ F \in \mathfrak{F}_{\mathcal{S}}, \qquad c(F) \perp \mathcal{V}_x. \tag{6}$$

A common displacement perpendicular to all candidate differences changes every candidate's squared loss equally. This condition is also necessary for relative-loss invariance over all $y \in [0,1]^D$ (Proposition B.1). For a fixed observation, the preferred candidate can remain unchanged even when loss differences vary.

Shared composition can produce this structure. Fix an orthogonal decomposition $\mathbb{R}^D = U \oplus V$ and suppose

$$f_g^F(u + v) = a_g(u) + n_g^F(v), \qquad u \in U, \quad v \in V. \tag{7}$$

The maps a_g are identical across compatible collections; n_g^F may vary with F. All operations preserve common domains for both components. If the maps n_g^F commute within each collection, candidates with matching counts of every operation share their V-component endpoint, including repeated operations. Its squared-error contribution cancels in their loss differences, while their U components can distinguish order because the maps a_g need not commute (Proposition B.2).

For example, let $A_a(s,t) = (s/2 + 1/4, at)$ and $B_b(s,t) = (s/2, bt)$ on $[0,1]^2$, with $a, b \in [0,1]$. The first-coordinate maps are fixed; examples with $t = 0$ leave a, b unrestricted. From $x = (0,1)$, AB (A then B) and BA predict $(1/8, ab)$ and $(1/4, ab)$. When AB generates $y = (1/8, a^\star b^\star)$, their normalized loss difference is $\ell_{BA} - \ell_{AB} = 1/128$ for every a, b. Uncertain predictions share the same second-coordinate error, so the correct order remains determined (derivation: Appendix B.1).

Uncertainty that changes comparison. For a local analysis, fix model families and discrete choices and parameterize the shared functions by $\theta = (\theta_g)_{g=1}^G$, with one block per operation. Let k uniquely minimize loss at $\widehat{\theta}$. Its advantage over competitor h is $m_h(\theta) = \ell_h(\theta) - \ell_k(\theta)$, with $m_h^0 = m_h(\widehat{\theta}) > 0$. For changes $\delta = \theta - \widehat{\theta}$, define

$$\mathcal{U}_{\mathcal{S}} = \{\delta : \|A\delta\|_2 \le \varepsilon,\ \|\delta\|_2 \le R\}, \qquad W_{\mathcal{S}}(v) = \sup_{\delta \in \mathcal{U}_{\mathcal{S}}} |v^\top \delta|. \tag{8}$$

Here $A\delta$ approximates changes in predicted example outputs, and R fixes the analyzed neighborhood. Valid residual and uniform approximation bounds determine ε so that this set contains the example-compatible changes within radius R (Appendix B.2). For any direction v, $W_{\mathcal{S}}(v)$ is the largest allowed magnitude of $v^\top \delta$.

Write the margin as $m_h(\widehat{\theta}+\delta) = m_h^0 + v_h^\top \delta + r_h(\delta)$, where v_h describes its linear response and $|r_h(\delta)| \le \rho_h$ uniformly bounds the remainder on $\mathcal{U}_\mathcal{S}$. The preferred history remains unchanged if its advantage exceeds every allowed decrease. In particular, k remains uniquely preferred throughout $\widehat{\theta}+\mathcal{U}_\mathcal{S}$ if

$$m_h^0 > W_\mathcal{S}(v_h) + \rho_h \qquad \text{for every } h \neq k. \tag{9}$$

The right-hand side bounds linear variation and approximation error. For differentiable margins, v_h can be taken as $\nabla m_h(\widehat{\theta})$. For affine margins and finite R, the condition with $\rho_h = 0$ is necessary and sufficient.

Examples need only constrain directions relevant to comparison. With exact affine example responses, changes in $\ker A$, meaning $A\delta = 0$, preserve those responses. If each affine margin's v_h lies in the row space of A, the span of its rows, then $v_h^\top \delta = 0$ for every such change: loss differences stay fixed despite nonunique parameters (Corollary B.4). Repeated operations contribute to the same parameter blocks, so their effects can reinforce or cancel. Thus equally well-fitted estimates may differ in ranking stability according to their uncertainty directions; Section 5.4 tests this effect.

From stable preference to correct recovery. Suppose the analyzed set contains the true operations on all relevant inputs, the recorded history is a candidate, and its output equals the observation. Its loss at the true functions is zero, so a candidate uniquely preferred throughout the set must be the recorded history.

Under the same observation assumption, the recorded history outranks a competitor if their true-output distance exceeds the sum of their endpoint-error bounds, all measured in root mean squared error (RMSE) (Proposition C.2). One such bound sums the endpoint effects of local discrepancies evaluated at true intermediate inputs and propagated through the remaining estimated operations (Proposition C.3). It also applies to quantized functions without requiring differentiability.

The guarantees require the stated structural conditions or valid uncertainty bounds. Held-out example losses alone do not bound errors on unobserved intermediate images, and the estimator does not compute these certificates.

5 Experiments

We test recovery, reuse under changed conditions, and the effect of operation uncertainty on history decisions.

5.1 Experimental Setup

We use 180 query-source groups, 60 each from FiveK (Bychkovsky et al., 2011), Flickr8k (Hodosh et al., 2013), and COCO (Lin et al., 2014); variants sharing content form one group. Operation examples, queries, and external adaptations' FiveK training and validation content are disjoint. The main comparison uses $K = 8$ examples per operation, $N = 16$ candidates, and aligned 256×256 images. Conventional histories combine compression, resampling, blur, and sharpening; every AI-containing history and candidate includes one local edit from the declared inpainting prior. Selection rules are fixed before evaluation, and estimated operations are shared across sources. External training targets and training/validation candidates exclude complete test sequences of operation categories.

Eight external task adaptations are Two-Stream CNN (Liao et al., 2020), TransDetect (Li et al., 2023), TMFNet (Niu et al., 2024), FHDM+SVM (Wei et al., 2025), SeqFakeFormer++ (Shao et al., 2025), MMNet (Xia et al., 2024), TSOM++ (Li et al., 2025), and Editprint (Wu et al., 2026). They share query images, candidates, and operation examples. Conventional external interfaces use eight operation slots versus ChainPrint's four; Figure 3(a) matches inventories. Joint fitting shares ChainPrint's complete model prior and endpoint score. Appendix I gives adaptation budgets, sequence exclusions, and the finite-budget joint optimizer.

Top-1 averages credit over uniform tie breaking; generation failures and incomplete rankings remain in the denominator with zero credit. We average variants within source groups, groups within image sources, and the three sources equally. Paired 95% confidence intervals use 10,000 source-group bootstrap draws conditional on fixed examples and models. Tables bold each column's best value, including ties.

Table 1: History recovery: Top-1 (%) across three image sources for two-, four-, and eight-step histories, $K = 8$, $N = 16$. Failures score zero.

Method	FiveK						Flickr8k						COCO					
	Conventional			AI-containing			Conventional			AI-containing			Conventional			AI-containing		
	2	4	8	2	4	8	2	4	8	2	4	8	2	4	8	2	4	8
Two-Stream CNN	0.0	13.3	5.0	6.7	6.7	0.0	0.0	13.3	6.7	5.4	5.0	1.7	0.0	11.7	5.0	5.8	5.0	1.7
TransDetect	4.8	6.6	5.0	5.6	4.2	7.3	4.8	5.8	7.0	4.7	5.3	4.7	4.9	7.3	6.0	6.0	9.4	6.0
TMFNet	0.0	0.0	5.0	1.2	6.7	8.3	0.0	1.7	8.3	1.7	8.3	6.7	0.0	1.7	8.3	1.7	1.7	11.7
FHDM (79D) + SVM	1.7	6.7	16.7	10.8	21.7	13.3	5.0	5.0	15.0	9.6	11.7	15.0	3.3	3.3	16.7	9.6	15.0	11.7
SeqFakeFormer++	0.0	1.7	5.0	8.3	8.3	3.3	0.0	3.3	5.0	8.8	8.3	3.3	0.0	3.3	5.0	9.6	8.3	3.3
MMNet	0.0	3.3	3.3	2.5	5.0	5.0	0.0	3.3	3.3	2.5	3.3	6.7	0.0	3.3	3.3	2.5	5.0	8.3
TSOM++	1.7	3.3	6.7	1.7	11.7	5.0	0.0	8.3	8.3	0.4	11.7	6.7	0.0	3.3	5.0	0.4	11.7	5.0
Editprint	36.7	35.0	36.7	20.0	31.7	33.3	41.7	31.7	26.7	18.3	25.0	31.7	43.3	30.0	38.3	17.9	21.7	30.0
Joint fitting	**96.7**	**99.2**	**100**	**67.1**	**56.7**	**60.0**	**100**	**100**	**100**	**95.0**	91.7	90.0	**100**	**100**	**100**	**92.9**	**90.0**	**86.7**
ChainPrint	93.3	**99.2**	**100**	**67.1**	**56.7**	56.7	**100**	**100**	**100**	93.3	**93.3**	**91.7**	**100**	**100**	**100**	89.6	88.3	85.0

Table 2: Pairwise accuracy (%) for order, repetition, and configuration changes within 16-candidate sets. Columns give chain length; AI histories contain one local edit.

Method	Conv. order			AI order			Conv. repeat			AI repeat			Conv. config.			AI config.		
	4	6	8	4	6	8	4	6	8	4	6	8	4	6	8	4	6	8
SeqFakeFormer++	46.7	26.1	43.9	46.1	25.0	76.7	46.7	43.3	43.3	52.2	27.8	54.4	62.8	41.7	46.7	52.2	49.4	39.4
TSOM++	31.1	62.2	55.0	53.9	65.6	27.8	72.8	63.3	63.9	75.6	59.4	41.7	52.8	42.2	68.9	45.6	48.9	48.3
Editprint	92.2	84.4	77.8	82.2	40.6	**81.1**	61.7	71.1	71.1	77.2	60.6	45.0	53.9	55.6	57.8	57.8	65.0	53.9
Joint fitting	**100**	**100**	**100**	**86.7**	85.6	79.4	**99.4**	**99.7**	**100**	**84.4**	85.0	**79.4**	**99.4**	**98.6**	**97.8**	**86.1**	85.0	78.3
ChainPrint	**100**	**100**	**100**	86.1	**86.1**	78.9	**99.4**	**99.7**	**100**	83.3	**85.6**	78.9	**99.4**	**98.6**	**97.8**	85.6	**85.6**	**78.9**

5.2 History Recovery Across Temporal Spans

Single-operation examples recover histories not demonstrated as complete sequences. Across eighteen equally weighted cells (Table 1), ChainPrint averages 89.68% Top-1 versus joint fitting's 90.32% and external adaptations' 3.4–30.5%. Eight-step accuracy is 100.0% for conventional histories and 77.8% with a local AI edit. Every AI candidate includes an edit, so presence detection cannot distinguish histories. Four- and eight-step AI rankings are complete for 143/180 and 140/180 queries and correct whenever complete. Incomplete rankings score zero: recovery therefore depends on both discrimination and ranking availability.

Under the same prior, ChainPrint's advantage over joint fitting is lower setup cost (Section 5.5). All paired intervals for accuracy differences include zero, establishing neither superiority nor equivalence. Both correctly rank the same 137 eight-step AI queries; ChainPrint completes three more and joint fitting five. Their 1.11-point gap thus comes from ranking availability.

The recovered distinctions extend beyond operation categories. Table 2 tests order, repetition, and configuration differences, with half credit for ties. Conventional configuration accuracy reaches 97.8–99.4% despite unchanged category sequences: operation names alone cannot resolve these distinctions. Eight-step AI order accuracy is 78.9% versus Editprint's 81.1%, including failures; pairwise accuracy differs from full-candidate ranking. Fixed-endpoint tests reuse the same estimates across spans. Their accuracy is nonmonotonic because operations and competitors determine endpoint differences (Appendix J).

5.3 Reuse, Generalization, and Robustness

Matched examples update the operation knowledge. Under the common eight-operation inventory, ChainPrint and joint fitting reach 100% at $K = 2, 3, 4, 6, 8$, exceeding the strongest external adaptation by 65.0–66.1 points (Figure 3a); accuracy is already saturated at two examples. Panels (b,c) change settings or implementations of resampling, blur, or sharpening. Matched-versus-old comparisons change only example outputs: queries, candidates, example inputs, and fitting budgets are fixed, and both conditions refit. At $K = 8$, matched examples improve new-setting Top-1 by 53.9 points (95% interval: 46.7–61.1), reaching 100.0%, and new-implementation accuracy by 4.4 points (1.7–7.2), reaching 99.4%. Old examples retain much of the useful behavior under changed implementations, but adding old-setting examples does not close the configuration gap.

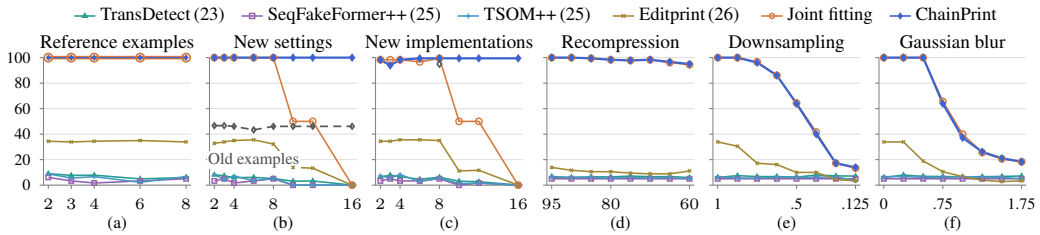

Figure 3: Eight-step conventional Top-1 (%): 180 source groups, $N = 16$. Horizontal axes: examples K (a–c), JPEG quality (d), scale factor (e), and blur σ in pixels (f). Panel (a) matches eight-operation inventories; (b–f) use four slots for ChainPrint and eight for other methods. Gray marks: ChainPrint with old examples. Endpoint processing (d–f) fixes $K = 8$. Larger-K availability: Appendix I.

Endpoint degradation tests a different limit. Panels (d–f) alter only the later image, holding the earlier image, examples, candidates, and operation models fixed. At JPEG quality 60, ChainPrint retains 95.0% Top-1 versus Editprint's 11.1%; strong downsampling and blur reduce it to 13.9% and 18.3%. With no new examples describing this added processing, the curves measure tolerance to changed observations: severe degradation can erase candidate differences.

A short list can retain useful histories even when the recorded chain is absent. On twelve source groups, the same functions rank up to 128 candidates. At 128, ChainPrint and joint fitting reach 100% Top-1 and Recall@5 for conventional histories present in the library; Editprint reaches 33.3% Recall@5. With the recorded history replaced, we replay actual candidate operations on eight independent images withheld from the ranker. Both methods' mean best-of-five error is 8.445×10^{-5} RGB MSE versus the mean library minimum of 8.390×10^{-5}, giving regret 5.457×10^{-7}. Using only the first choice adds 7.791×10^{-6}. Thus five recommendations approach the best available behavior; only the evaluator knows their best member, and similarity does not establish occurrence.

For AI histories, ChainPrint and joint fitting reach 91.7% and 66.7% Recall@5, with failures counted. Missing replays leave full-cohort AI recommendation errors undefined; identity retrieval alone does not establish usefulness when the recorded history is absent (Appendix I.1).

5.4 Ablations and Mechanism Analysis

Paired examples and executable composition provide the largest gains. On eight-step AI histories, shuffling before-and-after pairings lowers Top-1 by 71.8 points (95% interval: 65.9–77.3): unchanged images lose their demonstrated correspondence. Replacing sequential composition with additive changes lowers it by 41.4 points (33.1–49.4). Composition lets each operation respond to earlier changes; Table 2 separately tests order. Shared correction changes no source-level Top-1 outcome.

Table 3: Ablations: Top-1 (%), $K = 8$, $N = 16$. Columns give chain length. The first four rows vary refinement/correction; later rows alter full ChainPrint.

	Conventional			AI-containing		
Variant	2	4	8	2	4	8
Initial estimates	**97.8**	**99.7**	**100**	**84.2**	**80.0**	77.2
Initial + correction	**97.8**	**99.7**	**100**	**84.2**	**80.0**	77.2
Initial + refinement	**97.8**	**99.7**	**100**	83.3	79.4	**77.8**
ChainPrint	**97.8**	**99.7**	**100**	83.3	79.4	**77.8**
In-sample selection	**97.8**	**99.7**	**100**	83.3	79.4	**77.8**
Additive composition	35.3	12.9	10.6	47.2	46.6	36.4
Shuffled pairs	3.3	6.2	6.2	5.7	6.0	6.0
+ operation names	**97.8**	**99.7**	**100**	83.3	79.4	**77.8**

Refinement's AI differences range from -0.83 to $+0.56$ points, with intervals including zero. Coarse names and in-sample selection also leave Top-1 unchanged (Table 3).

On eight-step mechanism queries ($K = 8$), refinement increases unanimous correct agreement by 14.6 points; correction leaves it unchanged. All default estimates already reach 100% Top-1. Refinement also changes which alternative collections pass the compatibility check: agreement describes this retained set, not all remaining uncertainty (Appendix J).

Equal fit and overall variation need not imply equal stability. For the two-dimensional control $F_{a,b}(u, v) = ((1 + a)u + bv, v)$, we match example count, fit-error distribution, compatible-pool size, and overall prediction spread while changing which parameter the examples constrain. Across all 305 compatible estimates, constraining a preserves correct ranking in 24/32 conditions along u, versus 0/32 for b; along v, this reverses to 0/32 versus 16/32.

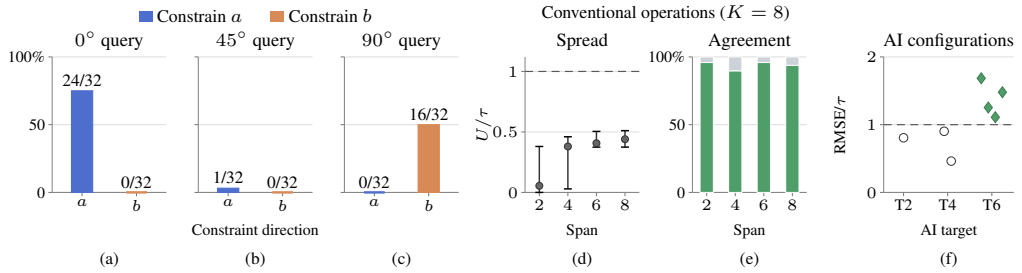

Figure 4: Operation differences and history decisions, $\tau = 1/255$. (a–c) Conditions correctly ranked by all 305 compatible estimates, out of 32 per constraint/direction. (d) Maximum conventional pairwise spread U/τ: median and interquartile range of 48 source–seed observations. (e) Correct agreement at $K = 8$; gray is unresolved. (f) Reference-relative RMSE for seven AI configurations; eleven are unevaluable. Diamonds mark four configurations of one operation above τ, preserving all nine correct decisions. Full diagnostics: Appendices K, I, and L.2.

This reversal isolates the direction of uncertainty relative to candidate comparison, as predicted by Section 4.3 (Figure 4a–c; full control: Appendix K).

Image experiments distinguish nearby agreement from substantial differences. Spread U is the maximum difference between retained estimates of any operation, measured by mean RMSE over independent evaluation images. Conventional estimates yield 89.6–95.8% correct agreement across spans at $K = 8$; every correct-agreement observation has $U \leq 1/255$. These are at most nine sampled model collections, not the full compatible set. Singletons at $K = 1$ leave agreement unresolved. Varying K also changes the retained collections, so it does not isolate uncertainty-set contraction (Appendix I).

Changing the editor's parameters tests larger differences. Seven of eighteen example-compatible alternatives have complete measurements on eight evaluation images. Four exceed $1/255$ reference-relative RMSE while each preserves all nine relevant correct eight-step decisions. All four concern one operation: differing image effects can preserve correct history choices, without establishing prevalence across operations (Appendix L.2).

5.5 Accuracy and Computational Cost

On six fixed timing sources, ChainPrint's setup costs 393.404 s versus 2086.499 s for joint fitting, a 5.30× reduction; ranking sixteen histories costs 17.518 versus 17.346 s/query (Table 4). For 100 queries sharing examples and unchanged operations, $S/100 + t$ is 21.452 versus 38.211 s/query; both methods recover all timing queries. The saving comes from estimating reusable functions, with similar individual ranking costs. Measurements use an A6000, batch size one, and exclude prior training (timing, memory, and cache controls: Appendix I.2).

Table 4: Eight-step AI histories on six timing sources. Top-1 (%); setup S and query t in seconds. $Q = 100$ reports $S/100 + t$.

Method	Top-1	Setup	Query	Q=100
Two-Stream CNN	0.0	**0.159**	**0.020**	**0.022**
TransDetect	10.3	0.494	0.092	0.097
TMFNet	0.0	0.510	0.054	0.059
FHDM (79D) + SVM	16.7	0.722	0.033	0.041
SeqFakeFormer++	0.0	0.411	0.106	0.110
MMNet	0.0	0.459	0.141	0.146
TSOM++	0.0	0.409	0.194	0.198
Editprint	33.3	9.958	0.289	0.389
Joint fitting	**100**	2086.499	17.346	38.211
ChainPrint	**100**	393.404	17.518	21.452

6 Conclusion

A history decision can be determined before its operations are fully identified. ChainPrint estimates executable functions from paired examples and reuses them across candidate histories. Experiments attribute the main recovery gains to preserving pairings and composing functions sequentially. The theory and directional control explain which uncertainty affects candidate comparisons; a real-editor instance preserves correct eight-step decisions despite differing operation effects. Recovery remains conditional on the model library, candidates, and aligned images. Within this setting, examples need to resolve decision-relevant differences, providing a basis for turning partial operation knowledge into reusable evidence about complex histories.

AI USE STATEMENT

We used ChatGPT and Codex to assist with organizing and drafting manuscript passages, including appendix material, refining the exposition of the methodology, and identifying relevant literature. The tools also assisted with research ideation and experimental design, implementation and debugging, and mathematical derivations and proof writing. AI tools generated illustrative image assets for the figures. The AI-edited images used in the experiments were generated with the frozen Stable Diffusion v1.5 inpainting model described in Section 4.1 and Appendix D. The authors reviewed the AI-assisted material and checked the cited literature. The authors take responsibility for the final text, claims, proofs, results, and figures.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the task and available information. Appendix D gives the complete estimator, model priors, and selection rules. Appendix I documents data roles, adaptation budgets, random seeds, aggregation, ties, and failure handling. Theoretical assumptions and proofs accompany the results in the appendices. Additional measurements and diagnostic scope are provided in Appendix J. These descriptions distinguish reported observations from unmeasured or undefined quantities.

REFERENCES

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25005–25017, 2022. doi: 10.52202/068431-1813.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Brooks_InstructPix2Pix_Learning_To_Follow_Image_Editing_Instructions_CVPR_2023_paper.html.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011. URL https://data.csail.mit.edu/graphics/fivek/.

Zhipeng Chen, Yao Zhao, and Rongrong Ni. Detection of operation chain: JPEG-resampling-JPEG. *Signal Processing: Image Communication*, 57:8–20, 2017. doi: 10.1016/j.image.2017.04.008.

Zhipeng Chen, Jie Zhu, and Jun Zhang. The forensicability of operation detection in image operation chain. *IEEE Access*, 10:68557–68569, 2022. doi: 10.1109/ACCESS.2022.3185994.

Coalition for Content Provenance and Authenticity. C2PA and content credentials explainer. C2PA Specifications. URL https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html. Version 2.2. Accessed September 13, 2026.

Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. doi: 10.1109/TIFS.2019.2916364.

Priya Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3fc2c60b5782f641f76bcefc39fb2392-Paper.pdf.

Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *SIGGRAPH 2001 Conference Proceedings*, 2001. URL https://mrl.cs.nyu.edu/publications/image-analogies/.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, August 2013. doi: 10.1613/jair.3994.

Shi-Min Hu, Kun Xu, Li-Qian Ma, Bin Liu, Bi-Ye Jiang, and Jue Wang. Inverse image editing: Recovering a semantic editing history from a before-and-after image pair. *ACM Transactions on Graphics*, 32(6): 194:1–194:11, 2013. doi: 10.1145/2508363.2508371.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Yuanman Li, Jiaxiang You, Jiantao Zhou, Wei Wang, Xin Liao, and Xia Li. Image operation chain detection with machine translation framework. *IEEE Transactions on Multimedia*, 25:6852–6867, 2023. doi: 10.1109/TMM.2022.3215000.

Yunfei Li, Yuezun Li, Baoyuan Wu, Junyu Dong, Guopu Zhu, and Siwei Lyu. Texture, shape, order, and relation matter: A new transformer design for sequential DeepFake detection, 2025. URL https://arxiv.org/abs/2404.13873v5.

Xin Liao, Kaide Li, Xinshan Zhu, and K. J. Ray Liu. Robust detection of image operator chain with two-stream convolutional neural network. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):955–968, 2020. doi: 10.1109/JSTSP.2020.3002391.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014. URL https://arxiv.org/abs/1405.0312.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Hannes Mareen, Dante Vanden Bussche, Fabrizio Guillaro, Davide Cozzolino, Glenn Van Wallendael, Peter Lambert, and Luisa Verdoliva. Comprint: Image forgery detection and localization using compression fingerprints. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, volume 13644 of *Lecture Notes in Computer Science*, pp. 281–299. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-37742-6_23.

Owen Mayer and Matthew C. Stamm. Forensic similarity for digital images. *IEEE Transactions on Information Forensics and Security*, 15:1331–1346, 2020. doi: 10.1109/TIFS.2019.2924552.

Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer. Image provenance analysis at scale. *IEEE Transactions on Image Processing*, 27(12):6109–6123, 2018. doi: 10.1109/TIP.2018.2865674.

Yakun Niu, Lei Tan, Lei Zhang, and Xianyu Zuo. TMFNet: Two-stream multi-channels fusion networks for color image operation chain detection, 2024. URL https://arxiv.org/abs/2409.07701.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Rui Shao, Tianxing Wu, and Ziwei Liu. Robust sequential DeepFake detection. *International Journal of Computer Vision*, 133(6):3278–3295, January 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02339-6.

Ashutosh Srivastava, Tarun Ram Menta, Abhinav Java, Avadhoot Gorakh Jadhav, Silky Singh, Surgan Jandial, and Balaji Krishnamurthy. ReEdit: Multimodal exemplar-based image editing. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 929–939, February 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Srivastava_ReEdit_Multimodal_Exemplar-Based_Image_Editing_WACV_2025_paper.html.

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and M. Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, Nov 2020. ISSN 2334-3303. doi: 10.1109/tac.2020.2966717.

Sebastiano Verde, Cecilia Pasquini, Federica Lago, Alessandro Goller, Francesco De Natale, Alessandro Piva, and Giulia Boato. Multi-clue reconstruction of sharing chains for social media images. *IEEE Transactions on Multimedia*, 25:9491–9505, 2023. doi: 10.1109/TMM.2023.3253389.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6830–6839, 2023. doi: 10.1109/CVPR52729.2023.00660.

Yang Wei, Haowei Liu, Xiaochen Yuan, Xiuli Bi, and Bin Xiao. Let images speak more: An efficient method for detecting image manipulation history. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):10665–10678, 2025. doi: 10.1109/TCSVT.2025.3571767.

Haiwei Wu, Kemou Li, Yuanman Li, and Jiantao Zhou. Editprint: General digital image forensics via editing fingerprint with self-augmentation training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35483–35493, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Editprint_General_Digital_Image_Forensics_via_Editing_Fingerprint_with_Self-Augmentation_CVPR_2026_paper.html.

Ruiyang Xia, Decheng Liu, Jie Li, Lin Yuan, Nannan Wang, and Xinbo Gao. MMNet: Multi-collaboration and multi-supervision network for sequential deepfake detection. *IEEE Transactions on Information Forensics and Security*, 19:3409–3422, 2024. ISSN 1556-6021. doi: 10.1109/tifs.2024.3361151.

Xu Zhang, Zhaohui H. Sun, Svebor Karaman, and Shih-Fu Chang. Discovering image manipulation history by pairwise relation and forensics tools. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1012–1023, 2020. doi: 10.1109/JSTSP.2020.2999827.

A Task Scope and Available Information

Operation examples and queries. The task observes groups of before-and-after examples, two query image versions, and a finite set of candidate histories. Each example group describes one operation under a shared configuration. Operation names, true parameters, and implementation metadata are withheld at inference. A method may estimate parameters within a declared model library; those model structures and any frozen editor configurations constitute its prior knowledge. The examples need not uniquely determine an operation on every possible input.

One application is an investigation with archived image versions and operation examples, but no executable access to the original tool version or its exact configuration. Available workflow records or hypotheses supply candidate histories. The task assumes these candidates and example groups are available. Benchmark construction may execute the true operations to generate observations and evaluate predictions. Inference receives no ground-truth assignment from anonymous identifiers to generating configurations. The declared model library may include a generating configuration, but its correspondence to an example group must be inferred. Different image content for examples and queries tests transfer across content; the basic interface itself does not require this separation.

Recovery spans and compound questions. For a propagation sequence $x^{(0)}, \ldots, x^{(T)}$, choosing $x = x^{(s)}$ and $y = x^{(T)}$ asks for the intervening subchain. The predecessor gives a one-step interval, an intermediate version a longer interval, and the initial version the complete history. The earlier observation need not be a camera-original image. Candidate histories may differ in order, length, repeated operations, or selected unknown positions. Combining such differences asks a compound question through the same ranking interface. Each query is ranked independently; one-to-one assignment across queries requires an additional correspondence constraint.

Recorded history present or absent. When the recorded history is a candidate, Top-1 accuracy measures how often it ranks first, Recall@k measures its inclusion among the first k, and mean reciprocal rank summarizes its position. When it is absent, recommendation quality compares candidate outputs with the recorded history's outputs on independent evaluation images (Appendix H.1). These images and true operation outputs are unavailable at inference. Behavioral similarity does not establish which history occurred; detecting absence is a separate decision problem.

Distinct histories can yield the same later image. For correct recovery, the relevant separation is between their true outputs on the earlier image, $T_h(x)$ and $T_{h^\star}(x)$, rather than only between predicted outputs. Appendix C gives a sufficient recovery condition when true-output separation exceeds combined endpoint-error bounds. Appendix H specifies aggregation, ties, and failure handling for the reported evaluations.

B When Examples Suffice to Stabilize Candidate Histories

The distinction between model identification and task sufficiency also appears in data informativity, which studies properties shared by all systems consistent with observed data (van Waarde et al., 2020). Task-based model learning instead fits predictive models for their downstream optimization objective (Donti et al., 2017). Our analysis concerns whether example-compatible operation models preserve comparisons between composed histories.

Operation examples may leave operation models nonunique while determining how candidate histories compare. We first characterize exact invariance of relative squared losses across compatible collections of operation functions. We then give local conditions under which changes in these losses cannot alter a fixed query's preferred history. ChainPrint selects operation models by example prediction error and does not evaluate these conditions during inference.

B.1 Exact Invariance of Relative Losses

Fix an earlier image x and a nonempty finite candidate set $\mathcal{H}$. Let $\mathfrak{F}_{\mathcal{S}}$ be a nonempty set of collections $F = (f_1, \ldots, f_G)$ whose example prediction errors meet specified tolerances within a declared model family. Each f_g is deterministic and shared across every occurrence of operation g. Assume every

candidate composition is defined at x for every $F \in \mathfrak{F}_{\mathcal{S}}$. For a later image y, define

$$z_h(F) = \big(f_{h_{L_h}} \circ \cdots \circ f_{h_1}\big)(x), \qquad \ell_h(F; y) = \frac{1}{D}\|z_h(F) - y\|_2^2,$$

where the images are flattened into $D = 3HW$ scalar values. Choose a reference collection $F^0 \in \mathfrak{F}_{\mathcal{S}}$, write $z_h^0 = z_h(F^0)$, and define the space spanned by candidate endpoint differences:

$$\mathcal{V}_x = \operatorname{span}\{z_h^0 - z_k^0 : h, k \in \mathcal{H}\}.$$

Proposition B.1 (Invariance of relative squared losses)**.** *Let $\mathcal{Y} \subseteq \mathbb{R}^D$ satisfy $\operatorname{aff}(\mathcal{Y}) = \mathbb{R}^D$. The following statements are equivalent:*

1. *For every $F \in \mathfrak{F}_{\mathcal{S}}$, every $h, k \in \mathcal{H}$, and every $y \in \mathcal{Y}$,*
$$\ell_h(F; y) - \ell_k(F; y) = \ell_h(F^0; y) - \ell_k(F^0; y).$$
2. *For every $F \in \mathfrak{F}_{\mathcal{S}}$, there exists a common displacement $c(F)$ such that*
$$z_h(F) = z_h^0 + c(F) \quad \textit{for every } h \in \mathcal{H}, \qquad c(F) \in \mathcal{V}_x^{\perp}. \tag{10}$$

Proof. Fix $F \in \mathfrak{F}_{\mathcal{S}}$. For each pair h, k,

$$D\big(\ell_h(F; y) - \ell_k(F; y)\big) = \|z_h(F)\|_2^2 - \|z_k(F)\|_2^2 - 2\langle y, z_h(F) - z_k(F)\rangle.$$

Suppose the first statement holds. The displayed affine function of y then agrees with its counterpart for F^0 on $\mathcal{Y}$. Since $\operatorname{aff}(\mathcal{Y}) = \mathbb{R}^D$, their linear coefficients and constant terms agree. The linear coefficients give $z_h(F) - z_k(F) = z_h^0 - z_k^0$ for every pair. Choose $k_0 \in \mathcal{H}$ and set $c(F) = z_{k_0}(F) - z_{k_0}^0$. It follows that $z_h(F) = z_h^0 + c(F)$ for every h. Substituting these expressions into the equality of constant terms gives

$$\begin{aligned} 0 &= \|z_h^0 + c(F)\|_2^2 - \|z_k^0 + c(F)\|_2^2 \\ &\quad - \|z_h^0\|_2^2 + \|z_k^0\|_2^2 \\ &= 2\langle c(F), z_h^0 - z_k^0\rangle. \end{aligned}$$

Thus $c(F)$ is orthogonal to $\mathcal{V}_x$. Conversely, a common displacement leaves $z_h(F) - z_k(F)$ unchanged, and its orthogonality to $\mathcal{V}_x$ leaves the difference of squared norms unchanged. The displayed loss identity therefore gives the first statement for every $y \in \mathcal{Y}$. □

The affine-hull condition holds for any nonempty open subset of $\mathbb{R}^D$ and for the complete normalized 8-bit image grid $\{0, 1/255, \ldots, 1\}^D$, which contains the origin and every coordinate unit vector. Thus the proposition does not require continuous-valued observations or differentiable operation functions. It characterizes exact relative-loss invariance over the observation set. For any fixed $y \in \mathcal{Y}$, this invariance preserves the entire candidate ranking, including ties; in particular, a unique preferred candidate remains unchanged across $\mathfrak{F}_{\mathcal{S}}$. The condition is sufficient but not necessary for a fixed query's preferred candidate to remain unchanged: loss differences can vary without changing which candidate has the lowest loss.

Cancellation under shared composition. The following structure produces the common orthogonal displacement directly from shared operation composition.

Proposition B.2 (Cancellation from shared composition)**.** *Let $\mathbb{R}^D = U \oplus V$ be a fixed orthogonal decomposition, and let P_U and P_V be the corresponding orthogonal projections. Suppose the state domain is $\mathcal{X}_U \oplus \mathcal{X}_V$, where $\mathcal{X}_U \subseteq U$ and $\mathcal{X}_V \subseteq V$ contain $u = P_U x$ and $v = P_V x$, respectively. Write f_g^F for the function representing operation g in collection $F \in \mathfrak{F}_{\mathcal{S}}$. Assume that every collection satisfies*

$$f_g^F(u' + v') = a_g(u') + n_g^F(v') \quad \textit{for all } u' \in \mathcal{X}_U,\ v' \in \mathcal{X}_V,$$

where $a_g : \mathcal{X}_U \to \mathcal{X}_U$ is the same for every F and $n_g^F : \mathcal{X}_V \to \mathcal{X}_V$. For each fixed F, assume that the maps n_g^F commute pairwise on the common invariant domain $\mathcal{X}_V$. No commutativity is required between maps from different collections, and the maps a_g need not commute. Suppose every candidate has the same operation counts $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_G)$, so that $|\{t : h_t = g\}| = \nu_g$ for every $h \in \mathcal{H}$ and every g. Repeated operations are allowed and reuse the same function.

Let $a_h = a_{h_{L_h}} \circ \cdots \circ a_{h_1}$. Then, for every $F \in \mathfrak{F}_{\mathcal{S}}$ and $h \in \mathcal{H}$,

$$z_h(F) = a_h(u) + N^F_{\boldsymbol{\nu}}(v), \qquad N^F_{\boldsymbol{\nu}} = (n^F_G)^{\nu_G} \circ \cdots \circ (n^F_1)^{\nu_1},$$

where a zero power denotes the identity on $\mathcal{X}_V$. For any $y \in \mathbb{R}^D$, set $y_U = P_U y$ and $y_V = P_V y$. Then

$$D\ell_h(F; y) = \|a_h(u) - y_U\|_2^2 + \|N^F_{\boldsymbol{\nu}}(v) - y_V\|_2^2.$$

Consequently, every pairwise loss difference is independent of F.

Proof. Composition preserves the separated form because each component maps its domain into itself. For each fixed F, pairwise commutativity on $\mathcal{X}_V$ allows the maps in the second component to be reordered without changing their composition. This component therefore depends only on the shared counts $\boldsymbol{\nu}$, giving the claimed endpoint expression. Orthogonality of U and V yields the loss decomposition. Its second term is identical across candidates and cancels in every pairwise comparison.

To connect this construction to Proposition B.1, use the reference collection F^0 and set

$$c(F) = N^F_{\boldsymbol{\nu}}(v) - N^{F^0}_{\boldsymbol{\nu}}(v) \in V.$$

The endpoint expression gives $z_h(F) = z^0_h + c(F)$ for every h. Moreover, $z^0_h - z^0_k = a_h(u) - a_k(u) \in U$, so $\mathcal{V}_x \subseteq U$. Since $U \perp V$, we have $c(F) \in \mathcal{V}_x^{\perp}$. Thus Equation 10 holds. □

If candidate counts differ, the result applies within each group of histories with matching counts; it does not imply invariance between groups. This is a sufficient construction of relative-loss invariance. The separated form and commutativity conditions are not assumed to hold for the JPEG or local AI editing executors used in our experiments.

A two-dimensional construction. Consider the continuous state space $[0, 1]^2$ and the declared operation family

$$A_a(s, t) = (s/2 + 1/4, at), \qquad B_b(s, t) = (s/2, bt), \qquad a, b \in [0, 1].$$

The family fixes the first-coordinate functions. Suppose all operation examples use inputs with $t = 0$ and the corresponding outputs

$$A_a(s, 0) = (s/2 + 1/4, 0), \qquad B_b(s, 0) = (s/2, 0).$$

These example predictions do not depend on a or b, so every collection $F = (A_a, B_b)$ in this family has zero example error. Thus the examples leave both parameters unrestricted: $\mathfrak{F}_{\mathcal{S}} = \{(A_a, B_b) : a, b \in [0, 1]\}$.

Take $x = (0, 1)$ and $\mathcal{H} = \{AB, BA\}$, with AB denoting A followed by B. Both candidates use each operation once, with the same a, b shared across candidates. Their endpoints are

$$z_{AB}(F) = B_b(A_a(x)) = (1/8, ab),$$
$$z_{BA}(F) = A_a(B_b(x)) = (1/4, ab).$$

Let $F^\star = (A_{a^\star}, B_{b^\star})$ be the generating collection, where $a^\star, b^\star \in [0, 1]$, and let the observed endpoint be $y = z_{AB}(F^\star) = (1/8, a^\star b^\star)$. For this two-coordinate construction, $D = 2$, so

$$\ell_{AB}(F; y) = \frac{(ab - a^\star b^\star)^2}{2},$$
$$\ell_{BA}(F; y) = \frac{1}{128} + \frac{(ab - a^\star b^\star)^2}{2}.$$

Consequently,

$$\ell_{BA}(F; y) - \ell_{AB}(F; y) = \frac{1}{128} \qquad \text{for every } F \in \mathfrak{F}_{\mathcal{S}}.$$

The absolute losses can change with the operation collection, while AB remains the unique minimum-loss candidate.

This construction satisfies Proposition B.2: the uncertain second-coordinate maps multiply by a and b and commute, while the fixed first-coordinate maps preserve a difference between the two orders. More explicitly, choose any reference collection $F^0 = (A_{a_0}, B_{b_0})$. For both candidates,

$$z_h(F) = z_h(F^0) + c(F), \qquad c(F) = (0, ab - a_0 b_0).$$

Their reference difference space is $\mathcal{V}_x = \mathrm{span}\{(1,0)\}$. The common displacement is therefore orthogonal to every candidate difference, as required by Proposition B.1. This is a continuous two-dimensional theoretical construction without Q_8 clipping or quantization; it is not an image experiment.

B.2 Local Conditions for Ranking Stability

We now fix the later image and allow candidate loss differences to vary within a declared parameter neighborhood.

Shared operation models and candidate losses. Let $\theta = (\theta_g)_{g=1}^{G} \in \mathbb{R}^p$ parameterize deterministic operation models f_{g,θ_g}. Every occurrence of operation g, within and across candidate histories, uses the same parameter block. Fix the model family, its discrete choices, and an estimate $\widehat{\theta}$ obtained from the operation examples; the later image is not used to refit this estimate. For a fixed query (x, y) and finite candidate set $\mathcal{H}$, define

$$z_h(\theta) = f_{h_{L_h},\theta_{h_{L_h}}} \circ \cdots \circ f_{h_1,\theta_{h_1}}(x), \qquad \ell_h(\theta) = \frac{1}{D}\|z_h(\theta) - y\|_2^2. \tag{11}$$

Here D counts scalar image values. Let k be the unique minimum-loss candidate at $\widehat{\theta}$. For each $h \neq k$, define the loss margin $m_h(\theta) = \ell_h(\theta) - \ell_k(\theta)$ and $m_h^0 = m_h(\widehat{\theta}) > 0$. The margin is the difference between the two candidates' losses, so a positive value favors k. Baseline ties are outside the strict stability statement below.

Uncertainty constrained by examples. Let $F_{\mathcal{S}}(\theta)$ stack the models' responses on their respective example inputs, using a fixed normalization, and let $Y_{\mathcal{S}}$ stack the corresponding observed outputs. An affine response matrix or a response Jacobian A defines

$$\mathcal{U}_{\mathcal{S}} = \{\delta \in \mathbb{R}^p : \|A\delta\|_2 \leq \varepsilon,\ \|\delta\|_2 \leq R\}, \qquad \delta = \theta - \widehat{\theta}. \tag{12}$$

Here $\varepsilon \geq 0$ and $R \geq 0$. The radius R is declared, not inferred from good example fits. The operation models and compositions must be defined throughout the analyzed neighborhood; parameter coordinates, normalization, and radius must remain consistent. Changing parameter units requires transforming both constraints; arbitrary rescaling cannot strengthen the claim. This set is neither a posterior distribution nor a confidence region.

For example, suppose a putative true parameter $\theta^\star$ lies within radius R, the baseline and true example residual norms are at most τ_0 and τ, respectively, and throughout this ball

$$\|F_{\mathcal{S}}(\widehat{\theta} + \delta) - F_{\mathcal{S}}(\widehat{\theta}) - A\delta\|_2 \leq \rho_{\mathcal{S}}. \tag{13}$$

The triangle inequality then places $\theta^\star - \widehat{\theta}$ in $\mathcal{U}_{\mathcal{S}}$ when $\varepsilon = \tau_0 + \tau + \rho_{\mathcal{S}}$. The same inclusion holds for every parameter in the ball whose example residual norm is at most τ, using the same baseline and uniform remainder bounds. For affine responses, $\rho_{\mathcal{S}} = 0$. For differentiable responses with $A = J_{F_{\mathcal{S}}}(\widehat{\theta})$, a uniform vector Taylor remainder bound is required. If the response Jacobian is $\beta_{\mathcal{S}}$-Lipschitz throughout the ball in the induced Euclidean operator norm, then $\rho_{\mathcal{S}} = \beta_{\mathcal{S}} R^2/2$ is valid. A Jacobian at the fitted parameter alone does not provide this control. Residual tolerances must use the same norm as the stacked responses; a mean squared error (MSE) cannot be substituted directly for a norm tolerance. Example fit alone does not establish that true operations belong to the declared function family.

Exact affine margins. For a direction $v \in \mathbb{R}^p$, define its worst allowed variation by

$$W_{\mathcal{S}}(v) = \sup_{\delta \in \mathcal{U}_{\mathcal{S}}} |v^\top \delta| = \sup_{\delta \in \mathcal{U}_{\mathcal{S}}} v^\top \delta. \tag{14}$$

The equality follows from central symmetry. An affine margin is a substantive assumption: even affine endpoint predictions generally produce quadratic losses. The proposition applies when the quadratic terms cancel in the loss difference; otherwise the remainder analysis below is needed.

Proposition B.3 (Stability from example constraints)**.** *Let $R < \infty$. If each margin is exactly affine throughout $\mathcal{U}_S$, $m_h(\widehat{\theta} + \delta) = m_h^0 + v_h^\top \delta$, then k remains the unique minimum-loss candidate for every allowed parameter if and only if*

$$m_h^0 > W_S(v_h) \qquad \text{for every } h \neq k. \tag{15}$$

For an individual competitor, equality permits a tie, and a smaller baseline margin permits a reversal.

Proof. The set $\mathcal{U}_S$ is compact and symmetric. Hence the minimum of $v_h^\top \delta$ is attained and equals $-W_S(v_h)$. The smallest allowed margin is exactly $m_h^0 - W_S(v_h)$. Applying this identity to every competitor proves the claim. □

Corollary B.4 (Which parameter directions matter)**.** *Without the radius constraint, that is, for $R = \infty$,*

$$W_S(v) = \begin{cases} \varepsilon\sqrt{v^\top (A^\top A)^\dagger v}, & v \in \operatorname{range}(A^\top), \\ +\infty, & v \notin \operatorname{range}(A^\top), \end{cases} \tag{16}$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. For globally affine margins on this set, the same strict stability criterion applies.

Proof. If v has a nonzero component in $\ker A$, scaling that component leaves $A\delta = 0$ and makes $v^\top \delta$ unbounded. Otherwise, writing $Q = A^\top A$, the generalized Cauchy–Schwarz inequality gives $|v^\top \delta| \leq \sqrt{v^\top Q^\dagger v}\, \|A\delta\|_2$. For $v \neq 0$ and $\varepsilon > 0$, equality is attained at $\delta = \varepsilon Q^\dagger v / \sqrt{v^\top Q^\dagger v}$. The zero cases are immediate. Thus finite extrema are attained; an unbounded direction can reverse any finite baseline margin. □

Consequently, $\ker A \neq \{0\}$ need not prevent stability. Examples may leave operation models nonunique while constraining every relevant v_h. As a small illustration, take $A = (1, 0)$, $\varepsilon = 0.02$, and $R = 1$. The affine margins $0.1 + \delta_1$ and $0.1 + \delta_2$ have identical baseline margins and unit gradient norms, but their minima are 0.08 and -0.9. Only the first is stable: its variation follows the direction constrained by the examples. This is an illustrative margin calculation, not an image experiment or a claim of correct history recovery.

Nonlinear margins and repeated operations. For differentiable compositions, set $v_h = \nabla m_h(\widehat{\theta})$. If a valid bound holds throughout $\mathcal{U}_S$,

$$|m_h(\widehat{\theta} + \delta) - m_h^0 - v_h^\top \delta| \leq \rho_h, \tag{17}$$

then $m_h^0 > W_S(v_h) + \rho_h$ is sufficient for strict stability. Indeed, subtracting the worst linear variation and remainder bounds the margin from below. A Hessian operator-norm bound β_h on the entire radius-R ball gives $\rho_h = \beta_h R^2/2$. A pointwise derivative or sampled perturbations do not establish this uniform bound.

Shared parameters determine the relevant directions. Let $x_{h,0} = x$, $x_{h,t} = f_{h_t,\theta_{h_t}}(x_{h,t-1})$, and $J_{h,t,g} = \partial x_{h,t}/\partial \theta_g$. The chain rule, evaluated along the baseline trajectory, gives

$$J_{h,0,g} = 0, \qquad J_{h,t,g} = D_x f_{h_t}\, J_{h,t-1,g} + \mathbf{1}_{\{h_t = g\}} D_{\theta_g} f_g. \tag{18}$$

Repeated occurrences therefore accumulate into the same block, with contributions that may reinforce or cancel after subsequent operations. Each derivative is evaluated at its own intermediate input. Assigning independent parameters to repeated occurrences changes the uncertainty being analyzed and discards these shared contributions. Writing $J_h = \partial z_h/\partial \theta$, with all terms evaluated at $\widehat{\theta}$, yields

$$v_h = \frac{2}{D}\left[J_h^\top (z_h - y) - J_k^\top (z_k - y)\right]. \tag{19}$$

What stability does and does not establish. Stability concerns the preferred candidate under allowed operation models. To recover the recorded history additionally requires a shared true parameter $\theta^\star \in \widehat{\theta} + \mathcal{U}_S$ representing the operations on all relevant inputs, the recorded history as a candidate $h^\star \in \mathcal{H}$, and a noiseless observation $y = z_{h^\star}(\theta^\star)$. Its loss is then zero, so a stable unique minimizer must be $k = h^\star$. Observation noise requires additional candidate separation; a missing

recorded history precludes this conclusion. These statements concern raw endpoint MSE; score clipping and tolerance-based ties require separate checks. Quantization, clipping boundaries, and discrete model changes require separate control: a smooth approximation does not automatically certify the actual executor. Without justified model coverage and uniform remainder bounds, the derivative calculations provide sensitivity diagnostics only.

C Conditional Properties of Operation Composition

The following elementary properties connect operation estimation to history ranking for a paired observation (x, y) and a finite candidate set. They do not address recovery from the later image alone. The properties are conditional: they do not guarantee general identifiability, and fitting operation examples does not establish the required approximation bounds. All configurations and estimation decisions are fixed, so the true operations and their estimates are deterministic functions. We work on a common image domain $\mathcal{X} \subseteq [0,1]^D$ on which every operation and estimate maps $\mathcal{X}$ into itself; D is the number of scalar image values.

C.1 Consistency Under Local Renaming

Let $\mathcal{E}$ denote the complete operation-estimation rule, including model selection, shared correction fitting, and its adoption decision. For example group S_g, write $\widetilde{f}_g = \mathcal{E}(S_g)$. The same rule and tie-breaking order apply to every group, without using its numerical identifier. A permutation π moves S_g to index $\pi(g)$ and replaces every occurrence of g in a candidate history by $\pi(g)$. We write $\widetilde{f}_h = \widetilde{f}_{h_{L_h}} \circ \cdots \circ \widetilde{f}_{h_1}$ for the corresponding estimated composition.

Proposition C.1 (Local-renaming consistency)**.** *For corresponding histories, consistent renaming preserves predictions and scores:*

$$\widetilde{f}_{\pi h}(x \mid \pi\mathcal{S}) = \widetilde{f}_h(x \mid \mathcal{S}), \qquad s(x, y, \pi h \mid \pi\mathcal{S}) = s(x, y, h \mid \mathcal{S}). \tag{20}$$

Proof. The estimator receives the same pixels for group $\pi(g)$ after renaming as for group g before renaming. Each corresponding candidate therefore retrieves the same functions in the same order. Their compositions and endpoint errors are unchanged. □

C.2 Prediction Accuracy and Candidate Separation

Use the root-mean-square error (RMSE) distance $d(u, v) = \|u - v\|_2/\sqrt{D}$. Ranking by negative pixel mean squared error (MSE) is equivalent to ranking by increasing d. Suppose the recorded (generating) history $h^\star$ belongs to $\mathcal{H}$ and $y = T_{h^\star}(x)$. For each candidate, assume

$$d\big(\widetilde{f}_h(x), T_h(x)\big) \le \epsilon_h, \qquad \Delta_h = d\big(T_h(x), T_{h^\star}(x)\big). \tag{21}$$

Proposition C.2 (Sufficient condition for correct ranking)**.** *The generating history ranks strictly above a competitor h whenever $\Delta_h > \epsilon_h + \epsilon_{h^\star}$. If this condition holds for every $h \neq h^\star$, it is the unique minimum-error candidate.*

Proof. The triangle inequality gives

$$d\big(\widetilde{f}_h(x), y\big) \ge \Delta_h - \epsilon_h > \epsilon_{h^\star} \ge d\big(\widetilde{f}_{h^\star}(x), y\big). \tag{22}$$

□

These bounds explain how small composition errors relative to candidate separation can ensure recovery. The condition is sufficient, not necessary. Distinct histories with identical outcomes have $\Delta_h = 0$; the paired image observation cannot distinguish those histories by their forward predictions. The strict-ranking conclusions concern raw MSE or RMSE distances; score clipping and tolerance-based ties require separate implementation checks. If instead $d(y, T_{h^\star}(x)) \le \nu$, the same triangle inequalities give the sufficient condition $\Delta_h > \epsilon_h + \epsilon_{h^\star} + 2\nu$. Here ν must be a known or assumed observation-error bound; the condition alone is not an empirical robustness result.

C.3 Transporting Local Errors Through Estimated Suffixes

Fix a candidate history h and an earlier image x. Its true trajectory is $x_0 = x$ and $x_t = T_{h_t}(x_{t-1})$. The estimated suffix composes the remaining operation models after a given step:

$$\widetilde{F}_{h,>t} = \widetilde{f}_{h_{L_h}} \circ \cdots \circ \widetilde{f}_{h_{t+1}}, \qquad \widetilde{F}_{h,>L_h} = \mathrm{Id}. \tag{23}$$

In particular, $\widetilde{F}_{h,>0} = \widetilde{f}_h$. At step t, measure the local estimation error on the true input and its effect after this suffix:

$$\eta_{h,t}(x) = d\big(\widetilde{f}_{h_t}(x_{t-1}), x_t\big), \tag{24}$$

$$b_{h,t}(x) = d\Big(\widetilde{F}_{h,>t}\big(\widetilde{f}_{h_t}(x_{t-1})\big), \widetilde{F}_{h,>t}(x_t)\Big), \qquad B_h(x) = \sum_{t=1}^{L_h} b_{h,t}(x). \tag{25}$$

The suffix may suppress or amplify a local discrepancy. This definition measures its actual effect without assuming that the operations are smooth.

Proposition C.3 (Telescoping composition bound)**.** *For every $h \in \mathcal{H}$ and $x \in \mathcal{X}$,*

$$d\big(\widetilde{f}_h(x), T_h(x)\big) \le B_h(x). \tag{26}$$

Proof. Set $z_t = \widetilde{F}_{h,>t}(x_t)$ for $0 \le t \le L_h$. Then $z_0 = \widetilde{f}_h(x)$, $z_{L_h} = T_h(x)$, and $z_{t-1} = \widetilde{F}_{h,>t}(\widetilde{f}_{h_t}(x_{t-1}))$. Thus $d(z_{t-1}, z_t) = b_{h,t}(x)$, and the triangle inequality along $z_0, \ldots, z_{L_h}$ proves the bound. □

A response modulus bounds how much the suffix output can change for a given input change. It expresses the bound in terms of local error magnitudes. For $r \ge 0$ and $v \in \mathcal{X}$, let

$$\widehat{\omega}_{h,>t}(r; v) = \sup_{\substack{u \in \mathcal{X} \\ d(u,v) \le r}} d\big(\widetilde{F}_{h,>t}(u), \widetilde{F}_{h,>t}(v)\big). \tag{27}$$

By definition,

$$B_h(x) \le \sum_{t=1}^{L_h} \widehat{\omega}_{h,>t}\big(\eta_{h,t}(x); x_t\big). \tag{28}$$

The supremum is finite because outputs lie in $[0,1]^D$. The modulus may have jumps, including for quantized operations; no Lipschitz or differentiability assumption is required.

An alternative stepwise bound uses the true-operation response modulus

$$\omega_g(r) = \sup_{\substack{u,v \in \mathcal{X} \\ d(u,v) \le r}} d\big(T_g(u), T_g(v)\big). \tag{29}$$

Let $\widehat{x}_0 = x$ and $\widehat{x}_t = \widetilde{f}_{h_t}(\widehat{x}_{t-1})$. If $d(\widetilde{f}_g(u), T_g(u)) \le \eta_g$ at every predicted input of operation g encountered in this trajectory, then $e_t = d(\widehat{x}_t, x_t)$ satisfies

$$e_0 = 0, \qquad e_t \le \eta_{h_t} + \omega_{h_t}(e_{t-1}). \tag{30}$$

This follows by inserting $T_{h_t}(\widehat{x}_{t-1})$ and applying the triangle inequality. Unlike $\eta_{h,t}$, its approximation term concerns predicted inputs.

If $y = T_{h^\star}(x)$ and $h^\star \in \mathcal{H}$, the observable distance margin between a competitor and the recorded history satisfies

$$d\big(\widetilde{f}_h(x), y\big) - d\big(\widetilde{f}_{h^\star}(x), y\big) \ge \Delta_h - B_h(x) - B_{h^\star}(x). \tag{31}$$

Consequently, $\Delta_h > B_h(x) + B_{h^\star}(x)$ is a sufficient condition for strictly ranking $h^\star$ above h. This connects local estimation to candidate separation through the remaining estimated operations.

In a controlled experiment with access to the true operations, the true trajectories, $\eta_{h,t}$, $b_{h,t}$, and Δ_h can be computed to examine local errors, their growth through a suffix, and candidate separation. These are evaluation diagnostics, not quantities supplied to the inference rule. The endpoint effects of different errors may cancel, making the sum of their magnitudes conservative. Leave-one-content-out example errors do not bound errors on true or predicted intermediate inputs. Accepting a shared correction does not guarantee a larger ranking margin. The analysis concerns the observed later image; it supplies no guarantee for behavior or regret on independent probe images.

D IMPLEMENTATION OF THE COMPLETE ESTIMATOR

Inputs and selection order. Each anonymous operation receives K aligned RGB pairs (x^-_{gk}, x^+_{gk}) at 256×256 resolution. The main comparison uses $K = 8$. ChainPrint's original conventional bank contains four conventional operation instances. The AI-containing evaluation and the matched example-count comparison in Figure 3(a) use eight instances: four conventional and four local-AI. After anonymous permutation, eight-instance banks are processed in two batches of four and four-instance banks in one batch, with the same model prior and independent selection rule for every instance; batch membership supplies no operation-type information. The registered example-count sweep varies $K \geq 2$ without changing the model families, search grids, or acceptance tolerances. The separate $K = 1$ mechanism diagnostic uses the explicit fallback below.

Selection first compares pooled conventional and editor proposals with optional correction. It then considers block threshold alternatives and block gain alternatives, each with and without correction, followed by unit-gain filtering and integer local filtering. Thus correction is considered within the initial and block stages, rather than applied once after all refinement. The two local stages compare uncorrected alternatives with the complete current model. Fitting uses only the operation examples and the declared editor prior. Query endpoints, candidate histories, rendering parameters, and semantic labels do not enter model selection. Conventional model evaluation runs on the CPU, principally in float64; the separable optimizer uses float32. Pretrained weights remain fixed.

Pooled proposals and correction. Appendix F specifies the conventional single-pair operation model library and its numerical searches. Each example pair supplies a cascade proposal and a JPEG proposal. These $2K$ proposals precede the eight editor configurations specified below. We select the model with the lowest full-image MSE, averaged equally over examples in 8-bit units. Exact ties preserve proposal order. Selection returns an existing fitted model $\widehat{f}_g$ without jointly refitting its base parameters.

For held-out example h, we remove its two conventional proposals and the corresponding loss column. An editor proposal is eligible only if it is available for every remaining example; its availability on h cannot affect selection. We reselect the base model and fit its 31×3 residual matrix B_g on the other $K - 1$ pairs. The features are an intercept, predicted RGB values, and the 3×3 RGB neighborhood of the input-minus-prediction difference. Boundaries are replicated. We solve the float64 normal equations with a symmetric pseudoinverse of relative tolerance 10^{-10}. The corrected model $\widetilde{f}_g$ adds the residual in byte units, clips and rounds the result, and divides by 255. We accept shared correction at this stage only if all K held-out predictions are available, no held-out example has higher error, and mean error strictly decreases. If accepted, we refit B_g on all K pairs for the base model selected using all examples.

Block alternatives. Each fold aggregates statistics from exactly its $K - 1$ included examples. The block models use a full-resolution YCbCr 8×8 DCT and fit one quantization step $q \in \{1, \ldots, 255\}$ per channel–frequency coordinate. The threshold parameter is shared across the operation and follows the ordered grid

$$b \in (0.5, 0, 0.125, 0.25, 0.375, 0.625, 0.75, 0.875).$$

For $b = 0.5$, quantization uses nearest-even rounding. Otherwise, it uses $q\ \mathrm{sign}(c)\lfloor |c|/q + b \rfloor$. We minimize coefficient error first over q and then over b, retaining the first minimum at each step. The threshold stage compares ordinary and threshold-adjusted quantizers, each with and without a rounded mean RGB offset. The gain stage estimates a nonnegative gain for each frequency: $a = \max(C, 0)/E$ for template energy $E > 0$ and template–target product sum C; we set $a = 1$ when $E = 0$. It compares gain quantizers with and without the same offset.

Both stages compare the currently selected model, all uncorrected alternatives, and then all alternatives with shared correction, in that order. Each alternative is fitted without the held-out example, including its correction. A stage uses the previously saved held-out losses for the current model; it does not repeat earlier selection stages. A replacement requires mean held-out MSE to improve by more than 10^{-10}. Unlike the pooled correction stage, this does not require improvement on every example. We refit the selected alternative and any correction on all K pairs. If held-out losses for the current model are unavailable, all later stages retain that model.

Local filtering alternatives. Each local stage fits K independent single-pair proposals at positions $(r + c) \bmod 4 \neq 0$. Unit-gain proposals use the nine-tap separable model in Appendix F. We train each proposal for 128 Adam (Kingma & Ba, 2015) steps at learning rate 0.03, copy the terminal model, and replace only its gain with one.

Integer filtering operates on rounded bytes. Its radius $\rho \in [0, 3.999]$, gain $a \in [-4, 4]$, and threshold $\tau \in [0, 255]$ are shared across RGB channels. With $Q = 2^{24}$ and $n = \lfloor \rho \rfloor$, the kernel assigns weight $w = \lfloor Q/(2\rho + 1) \rfloor$ to offsets $-n, \ldots, n$. The two neighboring offsets each receive $\lfloor [Q - (2n + 1)w]/2 \rfloor$. We apply three horizontal passes followed by three vertical passes, with replicated boundaries. Each weighted sum s is quantized as $\lfloor (s + Q/2)/Q \rfloor$. For low-pass difference δ, the output is

$$\mathrm{clip}_{[0,255]}(x + \mathrm{trunc}(\delta \ \mathrm{round}(100a)/100) \, \mathbf{1}\{|\delta| > \lfloor \tau \rfloor\}) \, .$$

We fit this filter with straight-through gradients (Bengio et al., 2013). The two starting radii are $\rho = 0.125$ and 0.5; gain is initialized by least squares and threshold at zero. Each start uses 128 Adam steps with learning rate $0.03(129 - t)/128$. We retain the earliest minimum error on fitting positions across iterates and starts.

For either local stage, selection with held-out example h excludes the proposal fitted to h and its loss column. We select using the other $K - 1$ examples, then score the selected proposal on h. Adoption requires mean held-out improvement exceeding 10^{-10}. Finally, we select among all K proposals by their mean error over all K examples. Ties favor the proposal from the earliest example, and the selected proposal is not jointly refitted. These held-out losses guide model selection; they do not independently estimate validation performance.

A single example in the mechanism diagnostic. At $K = 1$, no held-out content remains for the preceding selection rules. We therefore select the uncorrected initial model by its error on the sole example, using its two conventional proposals and all available configurations of the same editor prior. Correction and later refinement are disabled. All four estimation variants consequently return the same initial model at this count; they are not four distinct ablations. This fallback defines the one-example diagnostic only. Appendix I specifies how it enters the finite set of compatible estimates.

Declared editor prior. The eight editor configurations form a Cartesian product of two prompts, two boxes, and two random seeds. The prompts are "a patch of red flowers" and "a small blue ceramic vase." The boxes are $(48, 64, 144, 176)$ and $(96, 80, 208, 208)$ in (x_0, y_0, x_1, y_1) coordinates, with eight-pixel inward feathering. The seeds are 112 and 317. Stable Diffusion 1.5 inpainting (Rombach et al., 2022)[1] uses Diffusers 0.36.0, CUDA fp16, 25 steps, strength 1, guidance 7.5, and an empty negative prompt. Each call resets a CPU generator to the configuration's seed. We enlarge the input to 512^2 by bilinear interpolation and edit the region where the mask is nonzero. We then downsample by antialiased bilinear interpolation and blend with the original alpha mask. Bitwise identity across devices is not assumed. Selecting a prompt, mask, and seed therefore relies on this matched finite prior; it does not discover an unknown editor.

Evaluation availability and ties. After fitting, we compose each candidate history on the predicted intermediate images and rank candidates by endpoint RGB MSE in $[0, 1]$. If the editor's checker refuses an output, that prefix and every history extending it become unavailable. The implementation retains the candidate set and can return a diagnostic ranking of available predictions; if none are available, it abstains. The primary comparison in Section 5.1 requires scores for the complete candidate set. Queries with unavailable candidates receive zero recovery credit and remain in the denominator. Records marked complete must contain finite scores for every candidate; violations are rejected. Thus available-only diagnostic rankings are not used to improve the reported primary accuracy. Tie groups use absolute tolerance 10^{-12} relative to each group's lowest loss. Metrics are averaged over uniform orderings within ties. Availability is reported separately from recovery accuracy.

[1] Model `stable-diffusion-v1-5/stable-diffusion-inpainting`, revision `8a4288a76071f7280aedbdb3253bdb9e9d5d84bb`.

E RETROSPECTIVE FOUR-STEP COMPARISON

The retrospective study uses 45 source-content groups from 12 devices, 16 distinct four-step candidates per group, and eight examples per operation. The operations are blur, downsampling followed by size restoration, sharpening, and JPEG. Each candidate generates one query endpoint, giving 720 queries per encoder: 495 with JPEG in the recorded history and 225 without it. JPEG uses quality 76 and 4:4:4 chroma. Pillow and FFmpeg native MJPEG are evaluated separately. Corrected and uncorrected predictions share all inputs, candidate sets, and selected base models.

		Top-1 (%) ↑		Endpoint MSE ↓	
Encoder	History	No correction	ChainPrint	No correction	ChainPrint
Pillow	With JPEG	98.6475	98.6475	0.68245	0.67115
Pillow	Without JPEG	99.3552	99.3552	0.29097	0.27831
Native MJPEG	With JPEG	62.6578	66.7470	4.67457	4.18848
Native MJPEG	Without JPEG	99.3552	99.3552	0.29097	0.27831

Table A1: Retrospective correction comparison: 45 source-content groups, 12 devices, $L = 4$, $N = 16$, and $K = 8$. MSE compares the predicted endpoint of the recorded history with the observed endpoint, in squared 0–255 pixel units. Means use content–device averaging; endpoint queries are dependent.

Table A1 shows that shared correction improves Top-1 with native MJPEG and leaves Top-1 unchanged with Pillow. For JPEG-containing histories, native-MJPEG Recall@3 increases from 81.6146% to 85.9522%; Pillow Recall@3 remains 100%. The 45 groups share 174 example contents across 360 assignments, so shared examples connect all groups. We report descriptive paired observations and do not treat the 495 JPEG-containing endpoints as independent samples or claim population-level significance. A separate earlier external comparison uses different methods and 60 different source contents; Appendix G reports its results.

F CONVENTIONAL MODELS AND RETROSPECTIVE IMPLEMENTATION

The complete ChainPrint estimator reuses the conventional single-pair proposal routines described in this appendix. We also specify how the retrospective study pooled these proposals and applied shared correction. Appendix D describes the complete estimator: it adds eight declared editor configurations and further model alternatives, evaluated in a fixed selection sequence.

F.1 IMAGES AND CANDIDATE MODELS

The proposal routines take aligned, equal-size RGB arrays in $[0, 1]$ at 256×256 resolution and are applied separately to every operation group. The retrospective study and the original conventional ChainPrint comparison use $G = 4$ conventional groups. The AI-containing evaluation and the matched example-count comparison in Figure 3(a) use $G = 8$ anonymous instances, comprising four conventional and four local-AI operations; the latter comparison retains conventional query histories. The half-resolution chroma model requires height and width divisible by 16. The resampling estimator models a reduction in resolution followed by interpolation back to the original size. It does not estimate registration between arbitrary-size or geometrically misaligned query versions. The local AI editor, including its declared prompts, masks, and random seeds, is specified in Appendix D.

For each operation group, every available example pair produces two proposals: one from the selection cascade below and one from a JPEG search. The cascade evaluates model families in sequence, using family-specific validation rules. The fitting positions are $T = \{(r, c) : (r + c) \bmod 4 \neq 0\}$ and the validation positions are $V = T^c$, with zero-based coordinates. These masks select output positions; the full observed input remains available to compute predictions. Unless stated otherwise, validation uses RGB MSE in 8-bit units over all positions in V.

Generic responses. The spatial model uses fourteen symmetric basis responses over a 9×9 neighborhood. Each basis sums neighbor-minus-center differences for offsets with the same unordered pair of absolute coordinates. Boundaries are replicated, and the ridge coefficient is 10^{-3}. Pointwise quantization searches levels 2–256. We fit both models on T and compare them on V; ties favor the spatial model. If the spatial model wins, we first consider a nonnegative scalar gain, then a gain with a

Stage	Parameter fitting	Decision and retained parameters
Generic response	Fit spatial and pointwise models on T; then consider spatial gain, dead-zone, and color responses.	Compare on V in the order described below. Refit the selected generic response on the full example.
Block response	Fit DCT steps on the top 192 rows. Fit a separate generic baseline on that crop, including its internal T/V selection.	Compare full-resolution chroma, then half-resolution chroma, with the generic baseline on the bottom 64 rows. Retain full-example step estimates if selected.
Block RGB offset	For each block variant, fit a rounded mean RGB residual on the top 192 rows using its top-row step estimates.	Compare on the bottom 64 rows against the best uncorrected model. Refit steps and offsets on the full example if selected.
Resampling	If no block model was selected, search sizes and filters on a deterministic subset of T.	Replace the generic response only if MSE on V decreases. Retain the fitted size and filter without refitting.
Separable filter	If the retained response is spatial and neither block nor resampling was selected, optimize on all of T.	Replace the generic response only if MSE on V decreases. Retain the terminal iterate without refitting.
JPEG	Search quality and subsampling on 1,024 positions sampled from T.	Always retain a separate JPEG proposal. Also replace the cascade proposal if JPEG reproduces every output byte in both T and V.

Table A2: Per-example proposals for 256×256 images, constructed in the order shown. Later stages retain the current model on equal validation error. The text specifies the initial spatial–pointwise tie rule and the exact JPEG replacement rule.

hard dead-zone threshold. Each replaces the previous model only if validation MSE strictly decreases. Gain and threshold fitting exclude output channel values at 0 or 1. The dead-zone search considers all distinct cutoffs in the absolute response and fits the nonnegative gain analytically. Ties in explained squared error favor the largest active set. We then compare a cross-channel affine model with the selected generic model. The affine model uses the 42 channelwise spatial responses and three center values, centered ridge regression with coefficient 10^{-3}, and a fitted intercept. Spatial regression excludes saturated output channel values. Affine regression requires all three output channels at a fitting position to lie strictly between 0 and 1. Refitting on the full example preserves these exclusions; validation includes saturated outputs.

Block responses. The block models use Pillow RGB-to-YCbCr conversion and an orthonormal 8×8 DCT. Each of the 192 channel–frequency coordinates has an independently fitted integer quantization step from 1–255, chosen to minimize coefficient error on the fitting blocks. The half-resolution variant averages and rounds each input 2×2 chroma block. It estimates target chroma using the least-squares inverse of its fixed bilinear upsampler, then applies the DCT at that resolution. During fitting, this inverse uses only the target crop from the top rows. Full-resolution chroma replaces the generic baseline only if RGB MSE on the bottom rows decreases. Half-resolution chroma must improve on both models. The best offset variant must improve on all three uncorrected models; ties between offset variants favor full-resolution chroma.

Resampling, separable filtering, and JPEG. Resampling enumerates bicubic interpolation before Lanczos, with square latent sizes from 1×1 through 255×255. The same filter reduces and restores resolution. Its fitting subset starts with 1,024 evenly spaced integer indices across the flattened image and discards indices in V. JPEG uses a different subset: 1,024 evenly spaced entries from the flattened list of positions already in T. Both subsets use NumPy's integer `linspace`, including its endpoints, and all three channels at each sampled position. JPEG enumerates subsampling 0, 1, and 2, then quality 1–100 within each subsampling choice. It uses Pillow with optimization and progressive encoding disabled. Neither search uses validation pixels to choose these parameters.

The separable response applies a shared symmetric nine-tap kernel three times horizontally and three times vertically, rounding after each pass. It then fits a gain and hard threshold to the resulting difference from the input. Kernel logits start at $-r^2/(2 \cdot 0.5^2)$ for radii $r = 0, \ldots, 4$. The threshold starts at zero, and gain is initialized by least squares on T. We optimize the kernel, gain, and threshold for 128 Adam (Kingma & Ba, 2015) steps at learning rate 0.03, using straight-through

rounding (Bengio et al., 2013) and a sigmoid surrogate for the threshold gate. The kernel is softmax-normalized, and the threshold is constrained to be nonnegative. Validation uses the exact rounded response from the terminal iterate.

Discrete parameter searches retain the first exact minimum. This favors fewer pointwise quantization levels, smaller DCT steps, bicubic over Lanczos, smaller latent sizes within each filter, and earlier JPEG subsampling and quality choices. A model family replaces another only if error strictly decreases, with two exceptions: the initial spatial model wins a tie with pointwise quantization, and an exact JPEG match replaces the cascade proposal. Comparisons use no tolerance. The two proposals from an example may therefore be identical; we retain such duplicates. Selection uses before-and-after operation examples without access to the true rendering parameters.

Before ranking queries, we fit local operation parameters from the operation examples using a fixed operation model library and fixed fitting procedures. No global retraining is required. We rank histories by comparing their composed image predictions with the observed later image.

F.2 Retrospective Pooled Selection and Correction

The retrospective procedure pools only conventional proposals. For each operation group, we order the 16 proposals as the eight cascade proposals followed by the eight JPEG proposals, preserving example order within each family. We apply every proposal to all eight input images and record a 16×8 matrix of full-image RGB MSE in 8-bit units. Selection minimizes the equally weighted mean over the eight examples; exact ties favor the first proposal. The selected proposal keeps its fitted base parameters, without averaging or joint refitting.

For a predicted image $v = \widehat{f}_g(x)$, the feature row at pixel p is

$$\phi_g(x)_p = \left[1;\ v_p;\ \mathrm{vec}\{(x - v)_{p+\delta} : \delta \in \{-1, 0, 1\}^2\}\right]^\top. \tag{32}$$

The feature vector includes all three RGB channels, giving 31 features. Neighborhood extraction replicates boundary pixels. The design matrix Φ_{gk} contains one row per pixel. The residual matrix is $R_{gk} = 255(x^+_{gk} - v_{gk})$, reshaped to $HW \times 3$. We compute the least-squares solution as

$$B_g = \left(\sum_k \Phi_{gk}^\top \Phi_{gk}\right)^\dagger \left(\sum_k \Phi_{gk}^\top R_{gk}\right). \tag{33}$$

We accumulate sufficient statistics in float64 and compute the symmetric pseudoinverse with relative tolerance 10^{-10}. Shared correction uses no additional ridge penalty. Corrected predictions add $\Phi_{gk} B_g$ in 8-bit units, clip to $[0, 255]$, round to integer bytes, and divide by 255. Validation uses these clipped and rounded predictions.

In this conventional pool, fold k removes both proposals fitted to example k and its loss column. We select among the remaining 14 proposals by mean MSE on the other seven examples; ties preserve their original order. For the selected base model, we fit residual coefficients using all pixels of those seven examples. The selected base model may differ across folds. Let e^0_k and e^1_k be its uncorrected and corrected full-image MSE on the held-out example. We adopt shared correction exactly when

$$e^1_k \le e^0_k \quad \text{for every } k, \qquad \frac{1}{K}\sum_k e^1_k < \frac{1}{K}\sum_k e^0_k. \tag{34}$$

If correction is accepted, we return to the base proposal selected from all 16 entries using all eight examples. We then refit the correction on all pixels of all eight examples. Otherwise, we retain that base proposal without correction. The eight folds decide whether to use shared correction; they do not independently estimate test performance. This implementation uses $K = 8$ examples per operation. The complete estimator retains this residual model, with the expanded pool and stage-specific acceptance rules in Appendix D.

F.3 Composition and Computational Cost

A cache stores the predicted image for each distinct prefix of a candidate history. For P distinct nonempty prefixes, composition requires P operation applications; without reuse, it requires

$\sum_{h\in\mathcal{H}} L_h$. Each completed candidate also requires one image-distance computation. The cost of an application depends on the selected operation model, so prefix counts alone do not measure runtime savings.

The completed four-step study converts negative-MSE scores to normalized log scores, with probabilities clamped below 10^{-300}. Normalization, floating-point rounding, and score-dependent tie tolerances can change candidate order. These scores are not calibrated probabilities of physical histories. The retrospective scoring rules are specified below. The complete estimator instead ranks raw MSE on $[0, 1]$ images, computed in float64, with the availability and tie rules in Appendix D. Operation fitting and shared correction are timed separately from query evaluation.

G Earlier External Comparison

A separate completed comparison uses 60 source-content groups from 12 devices, four-step histories, 16 candidates, and eight examples per operation. Table A3 reports the held-out condition for histories containing JPEG: 671 endpoint queries, evaluated by all four methods with the same inputs and candidate sets. These earlier operation estimators predate the pooled selection and shared correction in Table A1. Their 60 source contents do not overlap with the query inputs from that comparison's 45 source contents.

Method in the earlier comparison	Top-1 (%) ↑	Recall@3 (%) ↑
Earlier base estimator	49.7802	70.3608
Earlier codec-refined estimator	97.6248	100.0000
DINOv2 with task adapter	9.6537	25.5650
Comprint with task adapter	5.5245	20.9139

Table A3: Earlier external comparison: 60 source-content groups, 12 devices, $L = 4$, $N = 16$, and $K = 8$, for JPEG-containing histories in the held-out condition. Means use content–device averaging. The methods and queries differ from those in the retrospective shared-correction comparison.

The DINOv2 (Oquab et al., 2024) and Comprint (Mareen et al., 2023) encoders receive explicit four-step task adaptation and larger fitting budgets. Their results describe these adapters, not reproductions of the authors' complete forensic methods. Training compute is not matched. The table compares the earlier estimators under that study's protocol; it does not provide a paired comparison of the current ChainPrint correction with external methods. The current estimator and external adaptations are compared under the three-source protocol in Table 1.

H Ranking Metrics, Averaging, and Uncertainty

Each query consists of an image pair. When the candidate set contains the recorded history, let r_i be its rank for query i, with n queries. We evaluate exact recovery and retrieval using

$$\text{Top-}k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{r_i \le k\}, \qquad \text{MRR} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{r_i}. \tag{35}$$

Top-1 measures exact recovery of the recorded history. With one relevant history per query, Top-k accuracy, Hit@k, and Recall@k coincide. Mean reciprocal rank (MRR) summarizes the recorded history's position in the ranking.

For the current ChainPrint and joint-fitting evaluation, we sort raw $[0, 1]$ MSE losses in ascending order. Each tie group contains losses within absolute tolerance 10^{-12} of its minimum. External adapters use the same grouping rule, applying tolerance 10^{-12} to negative raw compatibility scores in their own units. These scores are not pixel errors. The completed four-step results retain their earlier normalized-score rule: $\tau = 10^{-7} + 10^{-6}\max(1, \max_j |s_j|)$, anchored at each group's highest score. If the recorded history lies in a tied block at ranks $a, \ldots, b$, uniform random tie breaking gives reciprocal-rank contribution

$$\frac{1}{b-a+1}\sum_{r=a}^{b}\frac{1}{r} \tag{36}$$

and Top-k contribution $\max(0, \min(b, k) - a + 1)/(b - a + 1)$. This averages reciprocal ranks directly, rather than taking the reciprocal of the average tied rank.

Across all 2,880 method–query records in the retrospective study, replacing saved normalized scores with raw negative MSE leaves Top-1, Recall@3, and MRR for the recorded history unchanged. It changes 29 complete candidate orders, 10 strict top-three candidate sets, and 95 approximate tie partitions. Agreement on these recovery metrics therefore does not imply identical full rankings. This comparison does not evaluate the new fixed-tolerance protocol.

Candidate count is reported alongside ranking scores. For a uniformly random ranking over M candidates with one correct history, chance Top-1 is $1/M$, chance Top-k is $\min(k, M)/M$, and chance MRR is H_M/M, where $H_M = \sum_{j=1}^{M} 1/j$.

Current three-source evaluation. We first average repeated query conditions within each source-content group, then average groups within FiveK, Flickr8k, and COCO, and finally give the three image sources equal weight. This rule applies to recovery metrics and to the per-query recommendation errors in Appendix H.1; minima are computed before aggregation. The primary recovery metric retains generation failures and incomplete candidate rankings as zero-credit queries. Replay failures have separately reported coverage and are not silently converted to zero error. Paired 95% intervals use 10,000 bootstrap draws of source groups within each image source, with the same draws for each method. They describe variation over source content conditional on the fixed operation examples and selected models.

Historical device-based evaluation. The earlier studies in Appendices E and G retain their original averaging: queries within source image, source-image means within device, then equally weighted device means. That device-based rule does not apply to the current three-source evaluation. Variation across optimization seeds and variation across sampled source groups describe different sources of uncertainty.

When the recorded history is absent from the candidate set, detecting its absence and evaluating recommended alternatives require separate reference labels and metrics. Both use the same observed image pair and operation examples. The candidate set specifies the histories available for retrieval.

H.1 Evaluating Recommended Histories

Exact recovery identifies the recorded operation sequence. Recommendation quality measures how closely candidate histories reproduce its effects. For query i, let $h_i^\star$ be the recorded history and let $\mathcal{P}_i = \{z_{ij}\}_{j=1}^{m_i}$ contain $m_i \geq 1$ independent evaluation images. These images represent inputs to the recovered subchain, including processed earlier images where applicable, and are withheld from every method.

For a fixed nonnegative output loss ℓ, define

$$D_i(h) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell\big(T_h(z_{ij}), T_{h_i^\star}(z_{ij})\big)\,. \tag{37}$$

Here T_h applies the actual operations, using the configurations recorded when constructing the benchmark. These configurations are unavailable at inference. Pixel MSE on aligned images in $[0, 1]$ is the output loss used in this evaluation. The evaluation-image distribution, preprocessing, and treatment of stochastic operations are fixed for each evaluation.

For a nonempty candidate set $\mathcal{H}_i$, let R_{ik} be the first k candidates returned by the method, where $1 \leq k \leq |\mathcal{H}_i|$. Define

$$E_{ik} = \min_{h \in R_{ik}} D_i(h), \qquad D_i^{\text{best}} = \min_{h \in \mathcal{H}_i} D_i(h), \qquad \text{Regret}_{ik} = E_{ik} - D_i^{\text{best}} \geq 0. \tag{38}$$

Each candidate's error is averaged over evaluation images before either minimum is taken. Absolute error E_{ik} measures the output error of the best recommended history. The candidate-set minimum D_i^{best} measures the lowest error available in the candidate set. Regret measures the additional error incurred by the ranking. We report all three when the recorded history is absent. Zero Regret can coexist with a large absolute error.

When the recorded history is present and its replay is deterministic, $D_i^{\text{best}} = 0$ for losses that vanish on identical outputs, so absolute error and Regret coincide. For $k > 1$, the metric evaluates the best option among the recommendations. It does not identify that option for the investigator. At $k = |\mathcal{H}_i|$, Regret is trivially zero.

Methods share candidate sets, evaluation images, losses, and k. For tied ranking scores, metrics are averaged over uniformly random tie breaking. True output losses never determine the returned order. We first compute these quantities per query, then average within source groups, within each image source, and equally across FiveK, Flickr8k, and COCO, as specified for the current protocol in Appendix H. In the reported results, $E@1$, $E@5$, and B denote the aggregated values of E_{i1}, E_{i5}, and D_i^{best}, respectively. Similarity on the evaluation-image distribution establishes neither functional equivalence on every possible input nor that a recommended history occurred.

I Experimental Protocol and Additional Controls

This appendix specifies the data roles, operation inventories, aggregation rules, and additional controls. The conventional model-collection analysis and AI retrieval analysis use existing test-source groups and are reported as retrospective diagnostics.

Source roles and operations. The main evaluation uses 180 query-source groups, sixty per image source. Separate FiveK content supplies 45 training groups and eight validation groups for the external task adaptations. The operation-example pool reserves 24 further source groups, excluded from those roles and the main queries. For the main $K = 8$ comparison, eight contents passing all declared editing configurations are selected before method scoring; all attempts are retained. Query failures are retained without replacement. These exclusions cover the current source inventory and available duplicate checks, rather than certifying absence from every historical development image.

Each of six conditions has four conventional and four local-AI operation instances. Conventional settings combine JPEG qualities 68/76/84 with 4:4:4 chroma, bicubic downsampling and resizing back at ratios .5/.625/.75, and Gaussian blur radii .4/.8/1.2. Sharpening uses radius 1.2, amount 160, and threshold 2. Two groups of four AI configurations come from the declared eight-configuration editor prior. Operation identities and generating parameters remain hidden from inference. The generating inventory is distinct from the complete model prior used by ChainPrint and Joint fitting.

Queries, training, and scoring. The main evaluation has 540 conventional and 1,080 AI-containing query units. The AI two-step condition contains four registered variants per source; they are averaged within source before aggregation. The four- and eight-step AI conditions each contain one query unit per source group. For each length and variant, the external adaptations' training and validation candidates exclude complete test sequences of operation categories. These learned comparisons use the disclosed task adaptations and frozen, validation-selected checkpoints. TMFNet and TransDetect are paper reimplementations; Editprint uses released weights with a task adaptation. The completed export pins checkpoint and training-contract hashes, but does not establish equivalence to the original authors' training budgets or provide independent training-seed replication. Their results concern these task adaptations, rather than reproductions of the authors' original benchmark protocols.

External adaptation budgets. Each neural adaptation uses 270 training and 48 validation episodes per head, six epochs, and batches of up to four episodes. The schedule contains 70 batches per epoch, at most 420 parameter updates per head; unavailable AI endpoints can leave a batch empty. Two-Stream CNN uses SGD with learning rate 10^{-3}, a doubled bias learning rate, momentum 0.9, and weight decay 5×10^{-4}. The other neural adaptations use AdamW (Loshchilov & Hutter, 2019) with weight decay 10^{-4} and gradient clipping at 0.1. Their learning rate is 10^{-3} except for TSOM++, which uses 2×10^{-5}. All use seed 2026091807. After each completed epoch, validation expected Top-1 is averaged over repeated episodes within each source–length cell, then equally across sources and lengths; the earliest maximizing checkpoint is frozen before formal evaluation. Four separate AI heads correspond to the four two-step variants; head zero is fixed for the four- and eight-step tests. FHDM+SVM evaluates 221 RBF models per head on the grid $\log_2 C \in [-2, 4]$ and $\log_2 \gamma \in [-4, 4]$, both in increments of 0.5, with tolerance 0.001 and seed 2026091802. The same validation criterion selects the model, retaining the first lexicographic grid point on exact ties. These are fixed task-adaptation budgets, not the original papers' native training schedules.

Joint fitting under the complete prior. Joint fitting estimates each operation jointly across its example contents; operations are fitted independently, and query images and candidate histories do not enter fitting. It retains the same complete model families, parameter domains, forward operators, and endpoint score as ChainPrint. Raw-model selection minimizes equally weighted full-image example MSE in eight-bit units. The joint quantized-separable fit uses 128 Adam updates at learning rate 0.03. For the unit-gain proposal, a separate run of this joint fit uses the included example contents; its terminal model is copied and its gain set to one. Integer-local fitting shares parameters across contents, with two initial radii, 0.125 and 0.5, and 128 Adam updates per start at learning rate $0.03(129 - t)/128$ for step $t = 1, \ldots, 128$. Exact training-mask MSE selects the iterate and start, including initialization. For the main $K = 8$ comparison, shared correction refits the complete raw-family competition on seven contents in each of eight held-content folds. It is adopted only if no held-out error increases and the mean strictly decreases; the final model and accepted correction are refitted on all eight contents. Closed-form and grid fits retain their declared objectives; the nonconvex fits use finite budgets and do not guarantee a global optimum.

The primary score uses absolute 10^{-12} tie groups anchored at the group's lowest loss, with uniform expected Top-1 and reciprocal-rank credit. Queries with unavailable candidates receive zero credit and remain in the denominator. Records marked complete must contain finite scores for every candidate; violations are rejected. Generation and method availability are reported separately. This primary metric differs from a diagnostic ranking restricted to executable candidates. Intervals resample source groups within each source, use common draws for method differences, and condition on the selected models and operation examples.

Adaptation and degradation. The example-count sweep uses a separately registered pool of sixteen example contents, disjoint from training, validation, and query content. Counts are $2, 3, 4, 6, 8, 10, 12, 16$, using nested subsets; each fitted model uses only the examples available at its own count. The final common-inventory comparison reuses the existing model for each corresponding K, after checking that it corresponds to the intended operation and ordered list of examples. This comparison performs no new fitting and does not substitute the $K = 8$ model at smaller counts. For ChainPrint, each of the K selection folds excludes its held-out content from proposal generation, parameter fitting, and proposal selection, then evaluates predictions on that content. The selected model is fitted or selected on all K examples as specified in Appendix D; fitting at a larger count and reusing that model at a smaller count is not permitted. The migration conditions in Figure 3(b,c) each contain sixty sources per image dataset, with twenty assigned to each of resampling, blur, and sharpening. Changed settings use a resampling ratio of .375, a Gaussian blur radius of 1.6, or an unsharp-mask amount of 240 with radius 1.2 and threshold 2. Changed implementations preserve the nominal settings but replace Pillow operations with independent NumPy implementations: antialiased Catmull–Rom resampling, a discrete Gaussian with reflected boundaries, and an unsharp mask using that Gaussian. Resampling uses half-pixel coordinates and folds out-of-range contributions to the edge; Gaussian kernels extend to four standard deviations, rounded up. Sharpening applies the original threshold, and outputs are rounded and clipped to eight-bit RGB. Matched-example and old-example conditions share example inputs, changed-operation queries, candidate definitions, and fitting budgets. Only the selected operation's example outputs differ, and both conditions refit. The separate earlier four-category setting diagnostic also included JPEG; its queries and paired estimates are not pooled with the three-category migration curves.

The later-image conditions use JPEG qualities $95, 90, 85, 80, 75, 70, 65, 60$; scale factors $1, .875, .75, .625, .5, .375, .25, .125$; and Gaussian blur standard deviations $0, .25, .5, .75, 1, 1.25, 1.5, 1.75$ pixels. Downsampled images are restored to the common size. The three interventions act separately on the same reference endpoint. Scale one and blur zero are identity controls; JPEG quality 95 is not an identity operation. Conventional and AI-containing results and their execution coverage are retained separately.

Operation inventories in the example-count diagnostic. The common-inventory comparison in Figure 3(a) supplies eight operation instances to every method: four conventional and four local-AI instances. It uses the same 180 sources, conventional eight-step query histories, and sixteen candidates. At $K = 2, 3, 4, 6, 8$, every source has complete example coverage. ChainPrint and joint fitting both achieve Top-1 and MRR of 1.000 at all five counts; their source-paired difference is zero. ChainPrint's paired Top-1 advantage over Editprint ranges from 65.0 to 66.1 percentage points, with

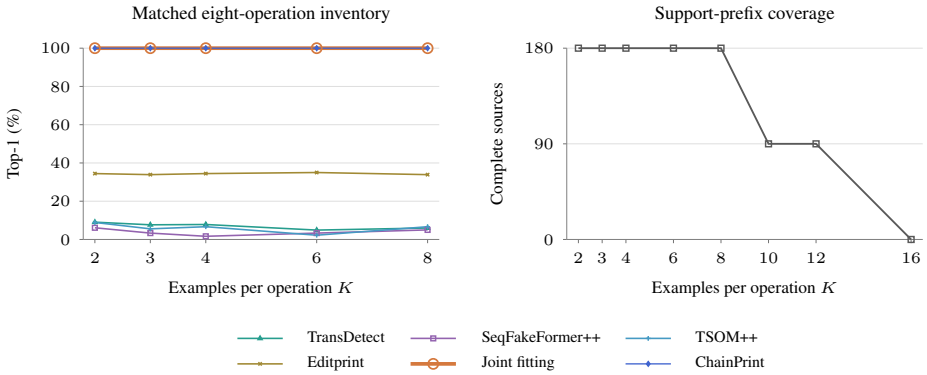

Figure A1: Matched reference-example comparison and support coverage. Left: conventional eight-step Top-1 (%) with an eight-operation support inventory, $N = 16$ candidates, and 180 registered source queries from three datasets. All six methods have complete support at $K \in \{2, 3, 4, 6, 8\}$. ChainPrint and Joint fitting coincide at 100%; the smaller blue diamonds lie within the orange circles at the same coordinates. Editprint reaches 33.9–35.0%, giving a ChainPrint–Editprint difference of 65.0–66.1 percentage points. Right: complete registered support prefixes for the same six methods, with a fixed denominator of 180 sources. Coverage falls to 90/180 at $K = 10, 12$ and 0/180 at $K = 16$; these counts are not performance estimates. Performance means are shown without confidence intervals.

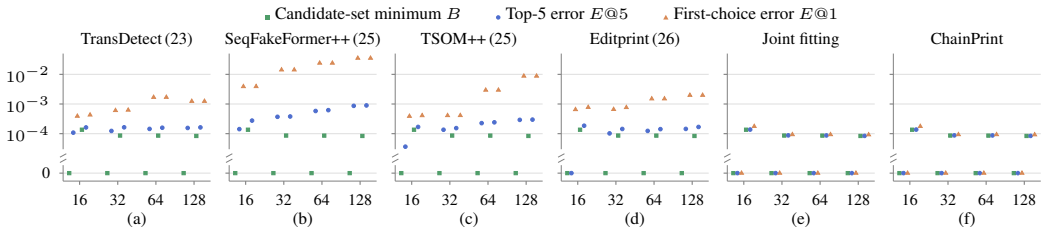

Figure A2: Eight-step conventional retrieval with $K = 8$. For each candidate-set size, left and right groups include or exclude the recorded history. Green squares, blue circles, and orange triangles show the library minimum B, best-of-five error $E@5$, and first-choice error $E@1$. Positive RGB MSE values share a logarithmic axis; exact zeros appear in a separate zero strip. Means use twelve source groups and eight independent replay probes.

every 95% source-bootstrap interval above zero. These results describe this fixed query and candidate design; the flat accuracy curve does not measure a benefit from further examples.

Figure A1 reports this comparison beside example availability at larger counts. Only 90 of 180 queries have the required eight-operation example sets at $K = 10$ and 12, and none do at $K = 16$. Those shortfalls retain zero credit under the full denominator, but are shown as coverage rather than joined to the performance curve. The migration and endpoint-processing curves in Figure 3(b–f) retain the original four-operation ChainPrint and eight-slot comparison interfaces. The migration curves also retain the larger-K availability shortfalls. Their matched-versus-old contrasts compare each method with itself on the same queries, without establishing cross-method sample efficiency under matched inventories.

I.1 History Retrieval and Recommendation Quality

Returning useful histories. Figure A2 increases candidate-set size while keeping the inspection budget at five histories. Twelve preselected source groups provide paired sets with the recorded history present or replaced; all operation examples remain available. At 128 candidates, both ChainPrint and joint fitting reach 100% Recall@5 on the conventional diagnostic; Editprint, the best external adaptation, reaches 33.3%. Table A4 reports Top-1 and Recall@5 from the same rankings.

With the recorded history present or absent, we replay it and each candidate on eight independent evaluation images and compare their outputs. Let B be the minimum replay error over all candidates, $E@5$ the minimum among the returned five, and $E@1$ the first-choice error. With the recorded conventional history absent from 128 candidates, ChainPrint achieves $E@5 = 8.445 \times 10^{-5}$ against

Table A4: Recorded-history identification with the recorded history present, on twelve source groups with $K = 8$ and eight-step histories. Top-1 and Recall@5 (R@5), in %, are reported separately for conventional and AI-containing queries. Conventional rows share the queries in Figure A2; incomplete candidate rankings receive zero credit.

	$N = 16$		$N = 32$		$N = 64$		$N = 128$	
Method	Top-1	R@5	Top-1	R@5	Top-1	R@5	Top-1	R@5
Conventional								
TransDetect	9.4	35.7	4.7	21.4	1.6	9.3	0.7	4.2
SeqFakeFormer++	8.3	50.0	0.0	8.3	0.0	8.3	0.0	8.3
TSOM++	25.0	50.0	8.3	41.7	8.3	33.3	0.0	16.7
Editprint	33.3	100	8.3	66.7	8.3	58.3	8.3	33.3
Joint fitting	100	100	100	100	100	100	100	100
ChainPrint	100	100	100	100	100	100	100	100
AI-containing								
TransDetect	1.7	25.3	0.8	6.1	0.4	2.1	0.0	0.6
SeqFakeFormer++	0.0	25.0	0.0	0.0	0.0	0.0	0.0	0.0
TSOM++	0.0	33.3	0.0	8.3	0.0	0.0	0.0	0.0
Editprint	33.3	91.7	25.0	66.7	0.0	41.7	0.0	25.0
Joint fitting	100	100	100	100	75.0	75.0	66.7	66.7
ChainPrint	91.7	91.7	91.7	91.7	91.7	91.7	91.7	91.7

the shared candidate-set minimum $B = 8.390 \times 10^{-5}$ in RGB MSE. Joint fitting has the same means. Top-5 regret is 5.457×10^{-7}, while the first-choice gap $E@1 - E@5$ is 7.791×10^{-6}, quantifying the additional error from inspecting only the first history. A small $E@5 - B$ means that inspecting five returned histories approaches the best available approximation; a large B means that even ideal ranking cannot supply a close alternative. Here, the returned five closely approach the library's best available behavior, while inspecting the first choice alone leaves a larger error. The evaluator identifies the best of five by replay; the ranking method does not receive that answer, and a close behavioral match does not establish that an alternative history actually occurred. For AI-containing histories, complete-cohort replay errors remain undefined because some candidate replays are unavailable; they are not pooled into Figure A2. Recorded-history identification remains measurable with failures counted: at 128 candidates, ChainPrint and joint fitting achieve 91.7% and 66.7% Recall@5 on the twelve-source AI diagnostic (Table A4). A separate post-hoc analysis asks whether the returned histories include one whose behavior can be verified from all eight probes. Only three of the twelve sources have any fully evaluated candidate, placing a 25% ceiling on this evidence-based rate. With the recorded history absent and 128 candidates, ChainPrint's verified Top-5 utility is zero at the four RMSE thresholds from $1/255$ to $8/255$. It reaches 16.7% at $12/255$ and 25% at the three thresholds from $16/255$ to $32/255$. Appendix I.1 reports every threshold and the comparison methods. This measures what the available evidence can validate, without substituting for the undefined AI replay errors or regret.

Recorded-history retrieval protocol. Twelve source groups, four per source, use nested candidate sets of 16/32/64/128 histories. The absent-history condition replaces only the complete recorded history with a registered alternative; operation examples and other candidates remain available. For each candidate, its true replay error is averaged over eight independent evaluation inputs before computing the minima. Figure A2 plots the absolute values B, $E@5$, and $E@1$ on a common logarithmic scale for positive RGB MSE, with exact zeros placed in a separate strip. The derived differences $E@5 - B$ and $E@1 - E@5$ retain their original definitions; distances on the logarithmic axis represent ratios, not these additive differences. Top-1 and Recall@5 are defined only when the recorded history is present. Undefined replay values retain their failure reasons and coverage rather than being replaced by fitted-model predictions.

Verified utility for AI-containing histories. Missing replays prevent us from computing complete-cohort AI behavioral errors or regret. We therefore analyze a narrower question: does a returned set contain a candidate whose usefulness can be verified from the available independent replays? This post-hoc analysis uses ten methods, the same twelve source groups, $N = 16, 32, 64, 128$, and separate present- and absent-history conditions. The metric, all thresholds, and aggregation rules were fixed before this calculation on previously inspected data. No new editor calls, model fitting, or candidate scoring are used.

Table A5: Post-hoc verified utility for AI-containing retrieval at $N = 128$. Each method cell reports Top-1/Top-5 utility (%) across all twelve sources. Columns give the RMSE threshold $a/255$. Bound rows give the method-independent fraction with a verified useful candidate in the library. All eight thresholds are shown.

History	Method	1	2	4	8	12	16	24	32
Present	ChainPrint	25.0/25.0	25.0/25.0	25.0/25.0	25.0/25.0	25.0/25.0	25.0/25.0	25.0/25.0	25.0/25.0
	Joint fitting	16.7/16.7	16.7/16.7	16.7/16.7	16.7/16.7	16.7/16.7	16.7/16.7	16.7/16.7	16.7/16.7
	Editprint	0/0	0/0	0/0	0/0	0/0	0/0	0/0	8.3/16.7
	Availability bound	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Absent	ChainPrint	0/0	0/0	0/0	0/0	8.3/16.7	16.7/25.0	25.0/25.0	25.0/25.0
	Joint fitting	0/0	0/0	0/0	0/0	8.3/16.7	8.3/16.7	16.7/16.7	16.7/16.7
	Editprint	0/0	0/0	0/0	0/0	0/0	0/0	0/0	8.3/16.7
	Availability bound	0	0	0	0	25.0	25.0	25.0	25.0

A candidate is verified useful at threshold $a/255$ only when errors on all eight independent evaluation images (probes) are available and their mean RGB MSE is at most $(a/255)^2$. Thus, the threshold applies to the square root of mean MSE, rather than the mean of probe-wise RMSE. We evaluate every $a \in \{1, 2, 4, 8, 12, 16, 24, 32\}$. For $k = 1$ or 5, verified Top-k utility is the probability that the method's original Top-k contains at least one such candidate. The original score ties retain uniform ordering. If a boundary tie contains m candidates, g are verified useful, and q positions remain, its contribution is $1 - \binom{m-g}{q}/\binom{m}{q}$. An already included useful candidate gives probability one. An incomplete original ranking receives zero; candidates are never reranked after dropping missing replays. Each reported rate averages the twelve sources equally, including every incomplete source. Probes, candidates, thresholds, and nested library sizes are not independent repetitions.

Only three source groups have any candidate with all eight evaluation-image errors: one group from FiveK and two from Flickr8k. The remaining nine have no verifiable candidate, so the rate cannot exceed 25% with this evidence. The library's verified-availability bound is the fraction of sources containing at least one verified useful candidate at the given threshold. It is a bound on what can be verified, not the true best-in-library error B. Of 12,384 probe-level errors, 9,364 are available and 3,020 are missing. The editor's safety checker and refusal outcomes remain unchanged; missing outputs are not imputed or replaced. These missing outcomes remain unverified; zero utility does not imply a large behavioral error or an incorrect history.

Table A5 reports all thresholds at the largest library size; the accompanying data include all four sizes and ten methods. With the recorded history present, ChainPrint attains the 25% evidence bound at every threshold and library size. Joint fitting does so for $N \leq 64$, but reaches 16.7% at $N = 128$ because one of these three sources has an incomplete ranking. That difference cannot be attributed to better ordering among completed rankings. Across the 96 original ranking records per method, ChainPrint has twelve incomplete records and joint fitting fourteen; the external adaptations have none. With the history absent, no candidate is verified useful at thresholds up to $8/255$, for any library size. At $N = 128$, ChainPrint's Top-5 rate rises to 16.7% at $12/255$ and 25% at $16/255$ and above. The eight thresholds expose this dependence on acceptable error without choosing a favorable cutoff. These are descriptive proportions on the fixed twelve sources; they neither recover the missing $E@1$, $E@5$, or B values nor define AI regret.

I.2 Finite-Collection Mechanism and Component Comparisons

The diagnostic in Figure A3(a–e) uses conventional histories from twelve registered test-source groups, four per image source. It examines the two-, four-, six-, and eight-step suffixes ending at the same physical eight-step endpoint, with sixteen candidates per query. Because these source groups belong to the existing test set, this is a test-subset diagnostic rather than an independent confirmation. For each of four example-sampling seeds, a fixed random permutation of the sixteen-example pool supplies nested subsets of $K = 1, \ldots, 8$ examples. Each count uses only its selected examples for fitting and compatibility checks.

For each estimation variant, we first fit its default collection of operation models to all K examples. We then form K alternative collections, each using uncorrected initial models fitted and selected from one of those examples. Repeated parameter collections are removed, retaining the default first

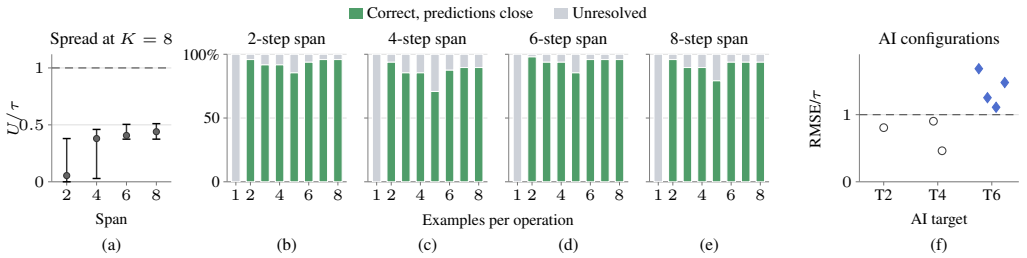

Figure A3: Operation variation and history agreement. (a) Conventional prediction spread U/τ at $K = 8$, with $\tau = 1/255$ RGB RMSE: median and interquartile range over 48 source–example-seed observations. (b–e) Conventional state fractions, averaging four seeds per source, for spans sharing an eight-step endpoint. Larger-spread correct agreement and incorrect agreement both occur zero times; every $K = 1$ query is unresolved. (f) Mean independent-image effect divided by τ for seven AI configurations with all eight probe pairs available. Four exceed the threshold and preserve the correct choice on all nine relevant queries; all four concern the same operation. Eleven configurations lack complete pairs. This effect measure differs from U (Appendix L.2).

and then the alternatives in example order. For operation g, let

$$e_g(f) = \left[\frac{1}{K}\sum_{k=1}^{K} \mathrm{MSE}\Big(f(x_k^-), x_{gk}^+\Big)\right]^{1/2},$$

where RGB values are in $[0, 1]$. The example error pools per-example MSE before taking the square root. An alternative collection is compatible only if every operation satisfies $e_g(f_g) \le e_g(f_g^{\mathrm{default}}) + 1/255$ with available predictions on all selected examples. This construction yields at most $K + 1 \le 9$ distinct collections before compatibility filtering. It samples a finite set of models and does not enumerate all models allowed by the theory. Query endpoints, recorded histories, and evaluation inputs are not used to generate or retain estimates.

At $K = 1$, all four variants use the uncorrected initial-model fallback in Appendix D. The default and sole alternative are identical, so deduplication leaves one collection and the state is Unresolved. This point records insufficient comparable estimates, rather than evidence of either agreement or disagreement between distinct models.

For each operation and pair of retained collections, we average output RGB RMSE over eight independent evaluation inputs. The query-level difference is the maximum over collection pairs and the operations in the recorded suffix; this evaluator-only operation set does not affect fitting or ranking. The primary threshold is $1/255$. Panel (a) displays $U/(1/255)$ at $K = 8$ for each span, using the median and interquartile range over all 48 repeated source–example-seed observations. These describe the recorded variation and are not confidence intervals; the 48 observations do not represent 48 independent sources. All 48 values are available for each of these four spans. A separate registered case retains eight model collections, each with the same positive nearest-wrong-history loss gap of 1.551×10^{-5} MSE. Its whole-collection probe difference is $U = 0.0017484$, below the primary threshold. At $K = 8$, the case is the first registered source/seed combination with at least eight compatible collections; if none qualifies, it is the combination with the largest count, using the same fixed order to resolve ties. Case selection uses model counts, not winners or correctness; the state proportions include all registered queries.

Differ and Close denote correct agreement with prediction differences above or at most the threshold. Wrong denotes agreement on an incorrect history. Unresolved includes different winners, ties, unavailable candidate or relevant probe predictions, and fewer than two comparable estimates. Except for the identical $K = 1$ fallback, the four estimation variants independently repeat fitting and selection on the same inputs. Their state proportions are interpreted jointly with coverage and retained-model counts. Additional thresholds $0.5/255$ and $2/255$ reuse the recorded prediction differences. The four example-sampling runs are averaged within each source group, followed by equal averaging across groups and the three image sources. Intervals use 10,000 paired source-group bootstrap samples; example seeds, probes, and compatible estimates do not count as independent source groups.

Efficiency. Methods use the same registered timing queries, hardware, batch size, warm-up, and synchronization rules. Setup includes example encoding, fitting, selection, and example-dependent

Table A6: Retrospective comparison of four estimation variants on the same twelve conventional source groups, with four example-sampling seeds, $K = 8$, and sixteen candidates for the eight-step span. States follow Figure A3 at threshold $1/255$; Top-1 evaluates each variant's default estimate. Coverage and median collection count use the same retained finite model sets. This fixed-subset diagnostic does not independently confirm a default-method effect.

		History decisions (%)				Estimate sets	
Variant	Top-1	Differ	Close	Wrong	Unres.	≥ 2 (%)	Median
Initial estimates	100	0.0	79.2	0.0	20.8	100	9
Initial + correction	100	0.0	79.2	0.0	20.8	100	9
Initial + refinement	100	0.0	93.8	0.0	6.2	100	8
ChainPrint	100	0.0	93.8	0.0	6.2	100	8

Table A7: Complete timing and sampled memory profiles for $K = 8$, $N = 16$, and eight-step histories. Top-1 is in %; all time columns are in seconds, with $Q = 1$ equal to $S + t$ and $Q = 100$ equal to $S/100 + t$. RAM and VRAM are the maximum sampled peaks over setup and query, in GiB. Each row uses the six fixed timing sources, with zero missing queries and zero incomplete candidate rankings.

Method	Top-1	Setup	Query	$Q = 1$	$Q = 100$	RAM	VRAM
Conventional							
Two-Stream CNN	0.0	0.158	0.020	0.178	0.021	1.166	0.510
TransDetect	1.0	0.492	0.093	0.585	0.098	1.305	0.730
TMFNet	0.0	0.512	0.054	0.566	0.059	1.172	0.947
FHDM (79D) + SVM	16.7	0.677	0.032	0.708	0.038	0.191	0.000
SeqFakeFormer++	0.0	0.409	0.105	0.514	0.109	1.576	0.535
MMNet	0.0	0.456	0.144	0.599	0.148	1.537	0.836
TSOM++	50.0	0.410	0.209	0.620	0.213	1.648	1.043
Editprint	16.7	9.923	0.289	10.212	0.389	1.439	4.998
Joint fitting	100	2086.499	0.175	2086.674	21.040	4.361	3.832
ChainPrint	100	393.404	0.175	393.579	4.110	2.921	3.832
AI-containing							
Two-Stream CNN	0.0	0.159	0.020	0.179	0.022	1.166	0.510
TransDetect	10.3	0.494	0.092	0.587	0.097	1.305	0.730
TMFNet	0.0	0.510	0.054	0.564	0.059	1.172	0.947
FHDM (79D) + SVM	16.7	0.722	0.033	0.756	0.041	0.193	0.000
SeqFakeFormer++	0.0	0.411	0.106	0.517	0.110	1.578	0.535
MMNet	0.0	0.459	0.141	0.600	0.146	1.538	0.836
TSOM++	0.0	0.409	0.194	0.603	0.198	1.662	0.926
Editprint	33.3	9.958	0.289	10.247	0.389	1.439	4.998
Joint fitting	100	2086.499	17.346	2103.845	38.211	4.361	3.832
ChainPrint	100	393.404	17.518	410.922	21.452	2.921	3.832

preparation. Query time includes processing the new image pair and producing scores for every candidate. Query-dependent fitting, if used by a comparison, belongs in query time. The $Q = 100$ column is the amortized quantity $S/100 + t$, not a cached throughput measurement. The reuse comparison fixes models, candidates, and random states and checks equal scores. Timing uses an NVIDIA RTX A6000 with driver 595.71.05, batch size one, and four CPU threads. The fixed cohort contains six source groups, two per image source, for each condition; all 120 method–condition–source query records are present. TMFNet, TransDetect, and Editprint were measured in a later registered window on the same frozen hardware and cohort, so the ten methods were not timed in one jointly randomized window. Weights and the single warm-up forward pass are loaded or executed before measurement; setup still includes the actual example-dependent fitting or encoding. Historical training and checkpoint selection are outside these setup and query costs. The means below describe these fixed sources without a timing confidence interval or training-seed replication. RAM is the sampled process-tree resident-set peak and VRAM is sampled NVIDIA process occupancy, including resident weights and caches; sampling intervals are 20 and 200 ms, respectively.

The native-method reuse controls in Table A8 match the primary timing records exactly by source, query, condition, and measured setup/query times. For each condition, we average times within a source, then sources within each image source, and finally the three image sources equally; the reported ratio divides the resulting reuse-off time by the reuse-on time. All 24 paired controls preserve candidate availability and scores exactly. ChainPrint's reuse-off/on ratio is 1.010 for AI-containing queries and 0.951 for conventional queries; these observations do not show a uniform speed benefit.

Table A8: Paired intermediate-prediction reuse control on the same six sources per condition. Times are in seconds; Off/on is the ratio of the corresponding source-averaged times. Score error is the largest absolute difference across every candidate score in the paired queries.

Method	Condition	Reuse on	Reuse off	Off/on	Score error
Joint fitting	Conv.	0.175	0.167	0.956	0
Joint fitting	AI	17.346	17.548	1.012	0
ChainPrint	Conv.	0.175	0.167	0.951	0
ChainPrint	AI	17.518	17.696	1.010	0

J Additional Analysis of the Reported Results

Matched-prior recovery and execution coverage. Under the same operation prior and endpoint score, mean Top-1 is 89.68% for ChainPrint and 90.32% for joint fitting across the six source-averaged conditions. ChainPrint's practical advantage in this comparison is its lower measured estimation cost: on the six fixed timing sources, example-dependent setup takes 393.4 s versus 2086.5 s for joint fitting (Section 5.5). Joint fitting matches ChainPrint on four- and eight-step conventional histories and four-step AI-containing histories, and scores higher on the remaining three conditions. On eight-step AI-containing histories, both methods correctly rank the same 137 queries. ChainPrint completes three further queries and joint fitting five, yielding the 1.11-point gap through ranking coverage. The source-paired intervals include zero in every condition, so we do not claim an accuracy advantage or statistical equivalence. The component comparisons below locate the information responsible for recovery in the paired examples and their sequential application.

Fixed-endpoint recovery spans. The main table uses separately generated histories of each length. A complementary evaluation fixes an eight-step endpoint and recovers the last one, two, four, six, or eight steps from the corresponding earlier version. One-step recovery uses four candidates; the other spans use sixteen. For the conventional histories, ChainPrint reaches 97.5%, 98.1%, 98.6%, and 100.0% Top-1 across the two-, four-, six-, and eight-step spans; one-step recovery reaches 99.7%. The estimates are reused across these suffixes without refitting to the query. Accuracy does not decrease monotonically with span, so chain length alone does not explain difficulty: which operations and competing histories occupy the span also matters.

Examples under changed operations. Changing the examples adapts the estimated operations to new settings and implementations. Figure 3(a) first tests example count with the same eight-operation inventory, query images, and candidates for all methods. ChainPrint and joint fitting both reach 100% Top-1 on all 180 sources at $K = 2, 3, 4, 6, 8$. ChainPrint exceeds the strongest external adaptation, Editprint, by 65.0–66.1 percentage points across these counts. The comparison removes the inventory difference in this diagnostic, but accuracy is already saturated at two examples; it does not show an improvement from adding more. Counts above eight are reported as availability diagnostics because their required example sets are incomplete. Panels (b,c) change either the configuration or the implementation of resampling, blurring, or sharpening, with the three categories equally represented within each image source. Each matched-versus-old comparison fixes the changed-operation queries, example input images, candidate definitions, and fitting budget; only the example outputs change. For new settings, matching examples yield 100.0% Top-1 at all eight example counts, while old-setting examples yield 43.3–46.7%. At $K = 8$, the paired gain is 53.9 percentage points (95% interval: 46.7–61.1). For new implementations, matching examples reach 99.4%, versus 95.0% with old examples at $K = 8$, a 4.4-point gain (1.7–7.2). Joint fitting also reaches 100.0% and 99.4% with matching examples in these two conditions. The smaller implementation gap shows that old examples already preserve much of the useful operation knowledge, while matched examples resolve additional differences. More old-setting examples do not close the tested configuration gap. The migration and endpoint-processing curves retain their original operation interfaces and all failures; Appendix I separates these protocols from the common-inventory comparison in panel (a). Their cross-method curves do not isolate sample efficiency under matched inventories.

Endpoint processing. Figure 3(d–f) separately applies JPEG recompression, downsampling, and Gaussian blur to the later image, keeping the recorded history, earlier image, examples, and candidates fixed. At JPEG quality 60, ChainPrint achieves 95.0% Top-1, a 5.0-point decline from the same clean queries. Editprint, the highest-scoring external adaptation at this level, falls from 33.9% to 11.1%.

Averaged equally over the eight downsampling and blur levels, ChainPrint exceeds Editprint by 49.0 and 44.1 points, respectively, but is 0.3 and 0.4 points below joint fitting. At the strongest tested downsampling and blur, ChainPrint falls to 13.9% and 18.3%, respectively. Here the operation models remain fixed: the curves measure tolerance to an altered observation, whereas the matched-example comparison tests adaptation to a changed operation. Keeping those questions separate prevents robustness to endpoint degradation from being mistaken for learning a new transformation.

Estimation controls and finite-collection agreement. Refinement and shared correction have a different empirical role. Shared correction changes no source-level Top-1 outcome, either before or after refinement, across all six conditions. Refinement without correction changes AI-containing Top-1 by -0.83, -0.56, and $+0.56$ points at two, four, and eight steps; all three paired intervals include zero. Using in-sample rather than held-out selection, and adding coarse operation names, also leave the source-level Top-1 outcomes unchanged here. The observed recovery gains therefore support paired operation examples and executable composition, while these additional estimation choices show no consistent accuracy gain in the tested setting.

On the same eight-step mechanism queries at $K = 8$, adding refinement increases the fraction of unanimously correct choices by 14.6 points relative to initial estimates, while adding correction changes it by zero (Table A6). The full method matches refinement alone, and every variant's default estimate has 100% Top-1 on this fixed subset. The fraction retaining at least two comparable estimates and the median retained-set size accompany these changes. A higher agreement rate can arise because the retained collections change, so it is not by itself evidence that a component resolves more of the true operation uncertainty. Incorrect agreement is reported separately; ties, failures, and fewer than two comparable estimates remain unresolved.

Additional computation-pruning evaluation. A separate frozen new-content evaluation tests lower-bound pruning of candidate computations (Appendix L). Across the two evaluated model banks, refusal-adjusted unique edit requests fall by 36.61% and measured arm time by 39.24% under the shared non-retry policy. Preferred-history sets agree on all five queries with complete baseline scores in each bank, out of twenty-four registered queries; both banks share those five queries. This is limited evidence for reducing evaluation cost while preserving the preferred history, not a new accuracy comparison or a guarantee about the full ranking.

K A Controlled Experiment on Example Direction

We completed a deterministic two-dimensional experiment to test whether the direction constrained by operation examples affects history ranking when example fit and overall operation disagreement are matched. The unknown operation and the known translation are

$$F_{a,b}(u, v) = ((1 + a)u + bv, v), \qquad T_d(u, v) = (u + d, v), \tag{39}$$

with true parameters $(a, b) = (0.2, 0)$. The generating history is $h_0 = T_d \circ F_{a,b}^r$; its rival is $h_1 = F_{a,b}^r \circ T_d$. Each estimate shares the same parameters across repeated occurrences and both candidates. For query x, the observed endpoint is $y = h_0(x; 0.2, 0)$, and candidate loss is $\ell_j(a, b) = \frac{1}{2}\|h_j(x; a, b) - y\|_2^2$.

The fixed grid contains 3,721 estimates: 61 equally spaced values each for $a \in [0.05, 0.35]$ and $b \in [-0.15, 0.15]$. One arm uses example inputs $\pm(1, 0)$, constraining a; the other uses $\pm(0, 1)$, constraining b. Outputs come from the same true operation. Compatibility uses only these examples, with maximum absolute coordinate error at most 0.01 (numerical tolerance 10^{-11}). Each arm retains 305 estimates, which we enumerate completely. Both arms have two examples, identical sorted example-RMSE values, maximum example RMSE 0.0070710678, and global behavior diameter 0.1503329638, measured by RMS output distance across probes $\pm(1, 0), \pm(0, 1)$. These antipodal examples provide one independent parameter constraint per arm.

We evaluated all 96 paired conditions crossing unit-query directions $0°, 45°, 90°$, translations $d \in \{0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.2\}$, and repetition counts $r \in \{1, 3, 5, 7\}$, giving history lengths $2, 4, 6, 8$. Within each pair, the query, observed endpoint, candidates, and true candidate separation remain fixed. We compute the margin $m = \ell_1 - \ell_0$ after composing each candidate; a condition is stable and correct only if every compatible estimate has $m > 10^{-12}$. Ties are recorded separately.

Table A9: Conditions with strictly correct ranking throughout the compatible pool. Denominators count prescribed grid conditions, not independent test images.

Query direction	Constrain a	Constrain b
Along u (0°)	24/32	0/32
45°	1/32	0/32
Along v (90°)	0/32	16/32

Table A10: Additional operation estimates on the existing evaluation content. Top-1 (%); conventional and AI columns each average nine cells, and Overall averages all eighteen. Robustness averages the eight-step conditions under the same corrected endpoint score.

Method	Conventional	AI-containing	Overall	Robustness
Frozen ChainPrint	99.17	80.19	89.68	77.04
Shared-radius estimates	99.54	80.93	90.23	77.13
Joint fitting	99.54	81.11	90.32	77.34

Table A9 shows that the useful constraint direction depends on the query; neither example arm is uniformly better. The frequent failures at 45° are retained. For $x = (0, 1)$, $r = 1$, and $d = 0.6$, both arms have true candidate endpoint separation 0.0848528137 in coordinate RMSE. Their minimum margins are nevertheless -0.010962 when constraining a and $+0.000150$ when constraining b. The former pool contains 212 correct estimates, 92 reversals, and one tie; all 305 estimates in the latter pool are correct. Thus, in this construction, example count, fit, pool size, and overall disagreement do not individually determine ranking stability.

Finite sampling can conceal this instability: across 20 fixed seeds, subsets of 8 and 32 estimates unanimously selected the generating history while missing a full-pool reversal in 471 and 81 of 3,840 condition-arm–seed observations per budget, respectively. Such observations are not independent image trials or population success probabilities.

For queries along v, the repeated-operation margin simplifies to

$$m = \tfrac{1}{2} d a s_r^2 (da + 2b), \qquad s_r = \sum_{j=0}^{r-1} (1 + a)^j. \tag{40}$$

Since $a, d > 0$, repetition changes only a positive factor and preserves the ranking sign; the lengths in this branch therefore repeat the same stability distinction. The construction isolates the directional mechanism in a completely enumerated finite model class. Figure 4(a–c) summarizes this control; panels (d–f) examine the corresponding question with image-operation estimates.

L Additional Estimation and Efficiency Results

Shared-radius estimates. An additional development study changes the operation estimates while retaining the earlier queries, candidates, and aggregation rules. Table A10 reports the frozen model, the resulting shared-radius estimates, and Joint fitting on the same eighteen cells. The updated estimates improve the overall mean by 0.56 percentage points relative to frozen ChainPrint and remain 0.09 points below Joint fitting. These are post hoc results on previously examined content, so Tables 1–3 retain the frozen model and their matched component comparisons.

Pruning candidate evaluation. A final, separately frozen evaluation measures whether a lower bound can avoid candidate computations without changing the preferred history. For a local edit, pixels outside its propagated influence region provide a lower bound on endpoint error. Candidates whose bound exceeds the best completed loss, with the frozen numerical tolerance, need not be rendered fully. Unproved dependencies and multiple local edits receive the conservative bound zero. The procedure outputs the preferred history and its optimal tie set; it does not produce a complete ranking.

Eight source contents are selected before any new endpoint or candidate scoring, with two-, four-, and eight-step histories and sixteen candidates per query. The twenty-four queries use both the frozen shared-radius development estimates from Table A10 and the joint-fitting estimates, with complete

Table A11: Final new-content efficiency evaluation, summed across both model banks. Time excludes model loading and generation of the observed endpoint. Unique requests exclude refusal-related keys from both arms; actual calls include attempted requests.

Evaluation	Unique requests	Actual calls	Time (s)	Non-retry skips	Peak MiB
Complete	224	276	311.44	16	3281.6
Pruned	142	156	189.23	25	3281.6

and pruned evaluation on the same physical A6000 GPU. The source exclusion certificate checks the pinned formal, development, support, and earlier-pilot manifests. The unavailable historical full-source index limits this exclusion claim to the documented asset collection. Arm order is balanced by source; accepted image caches are separate, while refused or unresolved requests share an immutable non-retry record.

After removing refusal-related reuse, the required request set falls from 224 to 142, a 36.61% reduction. Measured arm time falls by 39.24% under the shared non-retry policy; this time comparison includes the effect of that policy as well as pruning. Endpoint generation uses 24 calls and 26.94 seconds, and model loading takes 1.27 seconds. Seventeen of the twenty-four observed endpoints are available. Each model bank has five queries with sixteen exact baseline scores, and the pruned and complete optimal sets agree on all five; both banks use the same five queries from two source contents. No lower-bound violation is observed at the checked exact scores. Failures remain in the registered denominator and are not replaced. The four frozen gates—zero observed optimal-set mismatches, zero checked lower-bound violations, at least 30% request reduction, and at least 20% measured time reduction—are met. The limited complete-baseline coverage bounds the direct consistency evidence; this systems result does not establish a new accuracy improvement. A postprocessing correction removed the erroneous requirement that the pruned arm also evaluate all sixteen candidates; the raw records and all execution settings remained unchanged.

L.1 A Controlled Local-Edit Witness

This diagnostic constructs alternatives to the fixed local editor whose outputs agree on the eight operation examples. For the editor f, the pair is

$$f_{\pm}(x) = \mathrm{clip}_{[0,1]}\left(f(x) \pm \alpha \tanh\left(w^{\top}\phi(x)/0.05\right) Md\right),$$

where $\phi(x)$ consists of 4×4 RGB mean pools and a constant (49 features), M is the edit mask, and d is a fixed RGB direction. A seeded random vector is projected onto the null space of the feature matrix for the eight operation examples and normalized to obtain w. The five directions and seven strengths $\alpha \in \{0, 1, 2, 4, 8, 16, 32\}/255$ were fixed before the new probe and query outcomes, without using query histories to construct the perturbations. This is a constructed extension that preserves the editor outputs on the operation examples, rather than uncertainty obtained by fitting multiple natural editor models. Appendix L.2 separately evaluates actual guidance-scale alternatives selected using the operation examples; the two diagnostics retain their different model classes and coverage denominators.

Twelve existing source contents yield 36 query blocks at spans two, four, and eight; all 1,260 paired direction–strength records are retained. Directions, strengths, and spans are repeated measurements of these source contents. The maximum output change on the operation examples is 1.61×10^{-15}. All 700 records with no local edit in the recorded suffix retain the corresponding zero-strength candidate scores exactly. The recomputed zero-strength baseline differs from the archived baseline by at most 1.6117×10^{-4} in a candidate score, exceeding the declared 10^{-12} score-drift tolerance. Candidate availability and winning histories have zero mismatches, but the construction therefore does not reproduce the archived scores exactly.

Figure A4 shows the complete eight-step condition, where every recorded history contains one local edit. At strength $4/255$, 16 of the 60 repeated pairs meet the different-effects and correct-history criterion, spanning six source contents. At each strength $8/255$, $16/255$, and $32/255$, all five directions from those six sources meet it (30/60 pairs). Five further sources retain the correct history but lack a complete probe bank, and one source has unavailable candidate scores; these cases remain in the denominator at every strength. The shorter spans contain one and three AI-containing sources at lengths two and four, respectively. Their complete probe coverage is zero and one source, so they do not supply a broader independent confirmation.

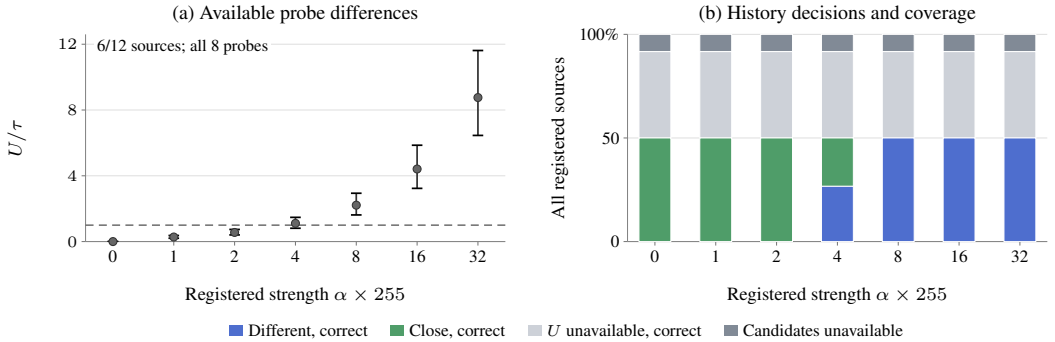

Figure A4: Constructed local-edit alternatives on eight-step histories, with $K = 8$ and $N = 16$. All seven registered strengths are shown at categorical x positions; $\tau = 1/255$ is fixed. (a) Mean whole-image probe difference U/τ and observed minimum–maximum over the six sources with all eight probes available, each with five directions. The dashed line is $U = \tau$; the segments are observed ranges, not confidence intervals. (b) State fractions retain all twelve sources, first averaging their five directions. Missing probe differences and missing candidate banks are separate categories. Blue indicates different predicted edit effects with the same correct history choice; green indicates close effects with the same correct choice.

These records provide an existence witness within the stated constructed model class: operation examples can admit different probe predictions while a finite candidate-history decision stays correct. They do not estimate the prevalence of ambiguity among naturally fitted AI models, guarantee agreement for every example-compatible model, or establish performance on new source contents.

L.2 Decision Stability Across Real Editor Configurations

From example compatibility to independent effects. This diagnostic tests whether distinct configurations of the fixed local editor can agree with the operation examples yet differ on independent images without changing an existing history decision. It evaluates actual guidance-scale alternatives, complementing the explicitly constructed output perturbations in Appendix L.1. Six distinct operation targets are identified by the pixel values of their complete example sets; repeated appearances of the same set do not count as independent targets. For each target, the default configuration is the complete configuration with lowest mean example MSE among the eight declared configurations, with ties resolved in their declared order. The example-only precheck varies the default guidance scale of 7.5 by offsets in $\{-0.5, -0.25, -0.125, 0.125, 0.25, 0.5\}$ while keeping the model weights, prompt, mask, seed, and other generation settings fixed. An alternative requires all eight example outputs and an example RMSE no more than $1/255$ above that of the default, where example RMSE is the square root of the mean of the eight image MSEs. All eighteen non-default alternatives passing this test for the six targets enter the following evaluation; independent images and query outcomes do not select the alternatives.

Eight evaluation images, disjoint from the operation examples and fixed before outcome evaluation, measure the effect of each alternative relative to its default. For each image we compute whole-image RGB RMSE on $[0, 1]$ pixels, then average the eight RMSEs within the configuration. An effect is evaluable only with all eight paired outputs and is considered different when this mean strictly exceeds $\tau = 1/255$. This configuration-to-default effect is distinct from the finite-collection prediction spread U in Figure 4(d). The image set, alternatives, threshold, and stopping rule remain fixed throughout the evaluation.

Pairing effects with history decisions. The decision evaluation reuses twelve existing eight-step queries, each with sixteen candidate histories. For each target and alternative, only that target's editor configuration is replaced in the relevant candidate compositions; the remaining operators, queries, candidates, and endpoint score remain unchanged. For every relevant query with scores for all candidates, the registered comparison checks whether the alternative preserves the default's selected history and whether its winner is unique. We additionally report whether both configurations select the recorded true history. Configurations and queries are repeated measurements of their targets, and the existing queries do not constitute new-content accuracy confirmation.

Seven alternatives, covering targets T2, T4, and T6, have complete eight-image effects (Table A12). Four exceed the frozen threshold, all from T6; their mean RGB RMSEs are 0.006607, 0.004913,

Table A12: All eighteen example-compatible editor alternatives across six targets. Independent effects require eight paired outputs; "–" denotes an unevaluable effect, never zero. Chain coverage gives complete/relevant configuration–query pairs; every complete pair retains the same correct winner. A witness requires a complete above-threshold effect and complete, correct winner agreement on the relevant queries. NE denotes an unevaluable witness.

Target	Guidance offset	Images	RGB RMSE	RMSE/τ	Chain coverage	Witness
T1	-0.500	7/8	–	–	2/3	NE
T1	-0.250	7/8	–	–	2/3	NE
T1	-0.125	7/8	–	–	2/3	NE
T1	$+0.125$	7/8	–	–	2/3	NE
T1	$+0.250$	7/8	–	–	2/3	NE
T1	$+0.500$	7/8	–	–	2/3	NE
T2	-0.125	8/8	0.003163	0.807	9/9	No
T3	$+0.125$	7/8	–	–	2/3	NE
T4	-0.125	8/8	0.003538	0.902	9/9	No
T4	$+0.125$	8/8	0.001813	0.462	9/9	No
T5	-0.250	7/8	–	–	2/3	NE
T5	-0.125	7/8	–	–	2/3	NE
T5	$+0.125$	7/8	–	–	2/3	NE
T5	$+0.250$	7/8	–	–	2/3	NE
T6	-0.250	8/8	0.006607	1.685	9/9	Yes
T6	-0.125	8/8	0.004913	1.253	9/9	Yes
T6	$+0.125$	8/8	0.004353	1.110	9/9	Yes
T6	$+0.250$	8/8	0.005800	1.479	9/9	Yes

0.004353, and 0.005800 at guidance offsets -0.25, -0.125, $+0.125$, and $+0.25$, respectively. Each of these four alternatives retains the same correct winner on all nine relevant eight-step queries. Figure 4(f) displays all seven complete effects, grouping alternatives by target to make this concentration explicit. The other eleven alternatives have only seven of eight paired image outputs because the corresponding default has an inherited rejection on one fixed image; they are not classified as above or below the effect threshold.

Across all eighteen alternatives, 85 of 96 registered configuration–query pairs are completely evaluable, and all 85 retain the same correct winner. Ten incomplete pairs inherit an unavailable baseline candidate bank for one query, and one additional pair is incomplete after an unretried safety-checker rejection. All 96 pairs remain in the coverage denominator. The fixed safety checker and its refusals are retained, with no image replacement, request retry, threshold adjustment, or additional parameter search.

Thus, distinct image effects and the same correct long-history decision coexist for one real editor target. The four alternatives share this target; they do not estimate the prevalence of the mechanism across targets or provide new retrieval-accuracy or regret estimates.