SEMANTITRACE: STEALTHY COPYRIGHT TRACING IN BLACK-BOX MULTIMODAL RAG VIA SEMANTIC PARASITISM

Anonymous authors
Paper under double-blind review
Code is available at https://anonymous.4open.science/r/semanti_anonymous-4C14.

ABSTRACT

Commercial Multimodal Retrieval-Augmented Generation (RAG) platforms ingest web-scale corpora behind proprietary APIs, making visual copyright protection reliant on verifiable black-box provenance tracing. Prior defenses bridge the cross-modal gap with synthetic visual canaries featuring bizarre compositions or arbitrary acronyms. These globally anomalous patterns contradict natural data distributions, leaving them vulnerable to pre-ingestion out-of-distribution sanitization. To resolve this trade-off between stealth and auditable decodability, we introduce **SemantiTrace**, which replaces de novo canary synthesis with *semantic parasitism*. SemantiTrace embeds contextually coherent, micro-scale forensic signatures in benign hosts by mutating existing typography or synthesizing geometrically natural, branded scene elements. To survive black-box retrieval and trigger explicit textual generation, Cross-modal Dual-Guided Latent Diffusion couples visual retrieval pull with teacher-forced textual decoding on the latent flow. Off-canvas pixel drift is eliminated by a support-preserving output projection; Dual-Space Soft Latent Blending and spatially gated gradients remove hard mask-edge discontinuities and control boundary energy within physically plausible supports. SemantiTrace passes pre-ingestion anomaly filters with 1.8% rejection on DENSE-10K and MMQA, and 2.2% on WebQA. On WebQA, our parasite canaries achieve 81.70% query success after global filtering; after corpus-wide adversarial repair and index reconstruction on a 400,000-item gallery, they retain 68.80% query success and $287/500$ All-K anchors, with a 0.90-point loss in benign QA utility. Under zero-clean-baseline audits with private multi-candidate assignment and hidden OCR serving, the selected $K = 1$ design requires 11 black-box queries for at least 95% fixed-population detection coverage under complete ingestion, with an exact private-assignment test at $p < .01$ and simultaneous 99% empirical false-alarm upper bounds below 1% for its evaluated budgets.

1 INTRODUCTION

Retrieval-Augmented Generation as a Service (RaaS) grounds Multimodal Large Language Models in external knowledge repositories (Gao et al., 2023; Peng et al., 2025; Li et al., 2025; Jovanović et al., 2025; Liu et al., 2025a; Lv et al., 2025). Yet proprietary visual archives are frequently assimilated into commercial vector indices and generation pipelines without authorization. Black-box provenance protection is challenging: dataset owners can neither examine the private retrieval database nor inspect generator weights (Zhao et al., 2023; Abootorabi et al., 2025).

Copyright proprietors can inject verifiable visual canaries into candidate datasets before potential ingestion (Chen et al., 2025). Related text-generation watermarking protects a different output modality (Pan et al., 2025; Gu et al., 2025; Park et al., 2026). In multimodal retrieval, a viable watermark cannot merely be an imperceptible pixel pattern. It must traverse two consecutive bottlenecks: the modality gap, where visual embeddings must be prioritized by black-box multimodal retrievers, and indirect propagation, where retrieved visual evidence must force the textual generator to decode an explicit textual signature verbatim (Suri et al., 2025; Drushchak et al., 2025). Pioneering methods such as AQUA (Chen et al., 2025) synthesize artificial images with rare acronyms or

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.

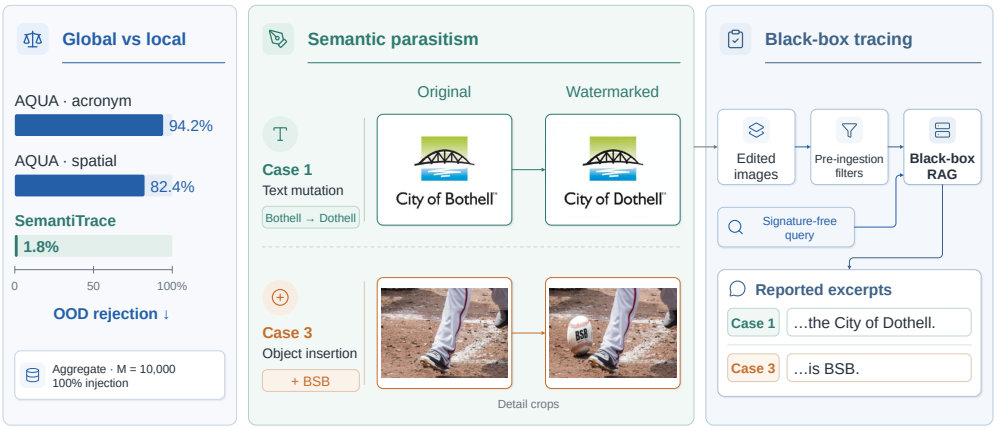


Figure 1: From synthetic watermarking to semantic parasitism. Left: pre-ingestion OOD rejection of AQUA's global acronym/spatial canaries and local SemantiTrace edits in the dense fidelity evaluation ($M = 10{,}000$, 100% injection; Table 10). Table 1 compares natural-scene AQUA and AnyText2 after filtering. Images and response excerpts reuse Cases 1 and 3 of Figure 17 (Appendix I). Image ingestion and signature-free queries follow separate paths.

unnatural spatial juxtapositions. While mathematically distinct, these synthetic canaries depart drastically from the image manifold: their structured scenes are almost entirely expunged by standard pre-ingestion Mahalanobis distance filters and vision-language anomaly inspectors (Figure 1, left). Although our natural-scene AQUA re-instantiations provide a stronger comparator (Tables 1 and 13), verifying canaries solely prior to automated sanitization fails to demonstrate real-world forensic auditability (Wang & Yu, 2025).

To resolve the trade-off between out-of-distribution detectability and verifiable downstream decodability, **SemantiTrace** replaces de novo canary synthesis with *semantic parasitism* (Figure 1). Rather than adding an alien synthetic composition, SemantiTrace edits existing benign hosts. It embeds verifiable textual signatures through micro-scale, contextually harmonious edits: a minimal character mutation on existing visible text, or a contextually plausible physical object hosting the textual trigger. Nesting the forensic trap within a legitimate image's semantic structure satisfies benign scene priors.

Operationalizing semantic parasitism requires addressing optimization challenges. Target-aware anchor mining identifies hosts in sparse retrieval neighborhoods with clean, editable canvases from segmented text or smooth structural regions. A frozen vision-language arbiter selects the canvas boundary and a contextually congruent textual trigger. Cross-modal Dual-Guided Latent Diffusion bridges the modality gap without alerting anomaly detectors. It applies simultaneous gradients to a flow-matching backbone: a retrieval loss pulls the visual representation toward benign audit probes, while a generative loss teacher-forces the surrogate vision-language model to decode the exact target signature. Dual-Space Soft Latent Blending removes hard mask-edge discontinuities, Spatially-Gated Gradient Scaling confines the gradient source, and the final output projection exactly preserves pixels outside the target support while controlling boundary energy (Equation 22).

Evaluations show 1.8% rejection on DENSE-10K and MMQA and 2.2% on WebQA (2.0% across the two audit cohorts) while establishing consistent zero-shot signature decoding across both open-weight and closed-source commercial multimodal engines. To avoid reliance on idealized matched clean systems during litigation, an exact conditional hypothesis test using randomized private assignment establishes mathematically provable provenance guarantees. Our contributions are:

1. **The Semantic Parasitism Paradigm:** We replace isolated synthetic generation with localized, context-aware host edits for multimodal dataset watermarking, achieving empirical pre-ingestion survival exceeding 98% across dense benchmark evaluations.
2. **Dual-Guided Flow Optimization:** We couple visual retrieval attraction with textual decoding on latent flows, using spatially gated gradients and soft latent blending to preserve visual fidelity while enforcing robust signature decodability.
3. **Statistically Rigorous Black-Box Auditing:** We introduce an exact paired hypothesis test and a prospective private-assignment audit requiring no un-ingested clean baseline,

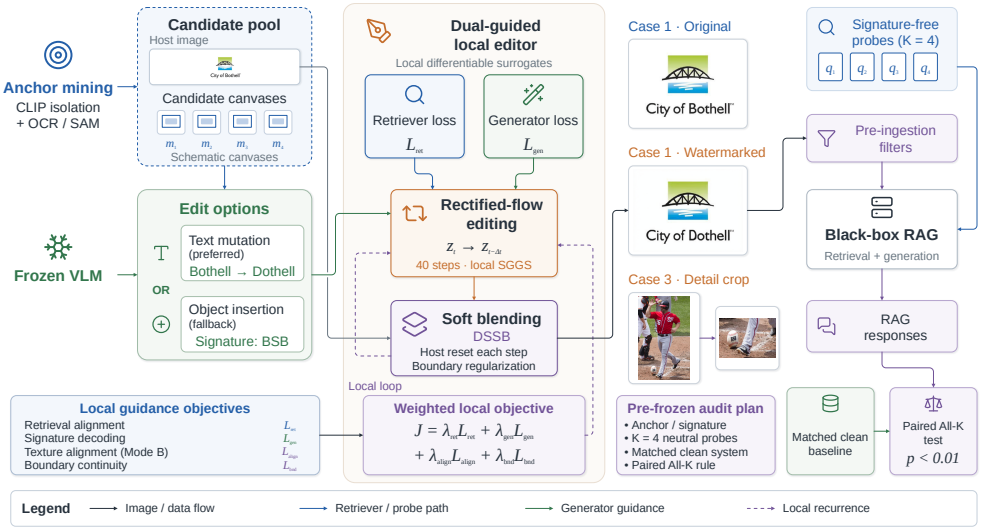


Figure 2: SemantiTrace: anchor mining, context-aware triggers, surrogate dual-guided editing with soft blending, and signature-free paired All-K auditing against a matched clean baseline. Examples reuse Cases 1 and 3 of Figure 17 (Appendix I); the baseball inset is an unaltered crop.

demonstrating rigorous false-alarm calibration and superior query efficiency under full-scale corpus repair and reindexing.

2 Related Work

Multimodal RAG and data tracing. RAG grounds language generation in retrieved knowledge (Lewis et al., 2020; Gao et al., 2023); multimodal extensions combine visual retrieval and VLM reasoning (Chen et al., 2022; Abootorabi et al., 2025). Their use of proprietary corpora motivates dataset-level tracing (Liu et al., 2025a; Chen et al., 2025). Textual canaries and data watermarks support provenance detection (Sablayrolles et al., 2020; Liu et al., 2025b; Sander et al., 2025), but invisible image watermarks and pixel perturbations do not necessarily survive the retrieval-to-text pathway (Wen et al., 2023; Alam et al., 2025; Sun et al., 2025). AQUA bridges this modality gap; its structured and natural-scene constructions have different OOD profiles (Fort et al., 2021; Liu & Qin, 2025).

Controllable generative editing. Latent diffusion (Rombach et al., 2022; Yao et al., 2025), local blending (Avrahami et al., 2022; 2023; Zhong et al., 2025), prompt-based control (Hertz et al., 2022; Bieske & Lorente, 2025), and external guidance (Bansal et al., 2024) enable constrained semantic edits. Guidance fields have also been used to steer generalist robot policies (Song et al., 2026). Building on these tools and SAM masks (Kirillov et al., 2023; Ravi et al., 2025), SemantiTrace couples retrieval and decoding objectives. Unlike pixel-level adversarial noise (Carlini et al., 2024) or globally anomalous synthesis, the resulting signature is a localized, plausible scene element. Appendix E.4 isolates the joint retrieval, decoding and sanitization objective against a same-host AnyText2 control with the full guidance objective.

3 Methodology: The SemantiTrace Framework

The owner edits source images before possible ingestion by a suspect RaaS platform, then audits its textual responses. Commercial retrievers, filters, and generators remain black boxes; optimization uses local differentiable surrogates (Figure 2).

3.1 Target-Aware Anchor Mining

We partition source dataset $\mathcal{D}$ into semantic clusters, selecting one anchor per cluster, favoring low retrieval competition and a locally editable canvas. With CLIP image embeddings Φ_{img}, the isolation

score is

$$\mathcal{S}_{iso}(I_i) = \frac{1}{K_{\text{nn}}} \sum_{I_k \in \mathcal{N}_{K_{\text{nn}}}(I_i)} \big(1 - \cos(\Phi_{img}(I_i), \Phi_{img}(I_k))\big). \tag{1}$$

Here $K_{\text{nn}} = 10$ denotes mining neighbors, distinct from the $K = 4$ audit probes. Higher isolation reduces local distractor competition.

OCR provides textual masks $\mathcal{M}_{text}$; SAM provides structural masks $\mathcal{M}_{struct}$. Textual canvases are preferred: character changes preserve physical geometry. Editability combines spatial prominence and OCR reliability:

$$\mathcal{S}_{text}(m_j) = \text{Area}(m_j)\, \text{Conf}_{OCR}(m_j). \tag{2}$$

For structural canvases, sufficiently large, smooth surfaces reduce blending artifacts. We compute the Laplacian on the continuous host image, not an image–mask product, and evaluate its variance on the eroded interior support $\Omega_{int}(m_j) = \text{erode}(m_j, \mathcal{K}_\rho)$, with $\rho = 3$ pixels:

$$\mathcal{S}_{struct}(m_j) = \frac{\text{Area}(m_j)}{\text{Var}_{x \in \Omega_{int}(m_j)}[\nabla^2 I_i(x)] + \epsilon}. \tag{3}$$

This excludes mask-edge discontinuities from the texture estimate. Within each semantic cluster, the preferred textual route selects

$$I_{anc} = \arg \max_{I_i \in \mathcal{D}_{cluster}} \left[\mathcal{S}_{iso}(I_i) \max_{m_j \in \mathcal{M}_{text}} \mathcal{S}_{text}(m_j) \right]. \tag{4}$$

If no viable textual canvas exists, we substitute $\mathcal{M}_{struct}$ and $\mathcal{S}_{struct}$. The winning image's top-scoring masks form $\mathcal{M}_{cand}$ for the VLM (rationale: Appendix B.1). HNSW approximates large-gallery neighbor search, preserving cluster-constrained selection (Appendix F.1).

3.2 Context-Aware Canary Generation

A frozen VLM receives the selected anchor I_{anc} and candidate bounding boxes, jointly choosing a canvas m_{opt}, visual trigger T_{trig}, and textual signature S_{trap}. **Mode A (preferred)** mutates existing text while preserving word length, casing, and scene geometry, e.g., "CAFE" to "CXFE". **Mode B (fallback)** inserts a plausible object when editable text is absent, e.g., a green book titled "XCO" on an empty desk. Each host receives exactly one of the two modes. Appendix I.2 gives the exact prompt.

3.3 Cross-modal Dual-Guided Semantic Injection

We intervene in a latent flow-matching backbone (Black Forest Labs, 2025). Let $z_0 = \mathcal{E}_{vae}(I_{anc})$ and $z_t = (1-t)z_0 + tz_1$, where $z_1 \sim \mathcal{N}(0, \mathbf{I})$ and denoising runs from $t = 1$ to 0. The clean estimate is $\hat{z}_0 = z_t - tv_\theta(z_t, t, c)$, decoded as $\hat{x}_0 = \mathcal{D}_{vae}(\hat{z}_0)$. The retriever and generator guidance losses are

$$\mathcal{L}_{ret} = 1 - \cos(\Phi_{img}(\hat{x}_0), \Phi_{text}(Q_{probe}))\,, \tag{5}$$

$$\mathcal{L}_{gen} = -\sum_{i=1}^{|S_{trap}|} \log P_{\mathcal{G}_{surr}}(s_i \mid s_{<i}, \hat{x}_0, Q_{probe}). \tag{6}$$

These respectively pull the host toward the probe and make the surrogate VLM explicitly decode the signature. For object insertion, a DINOv2-g/14 encoder (Oquab et al., 2024) supplies intermediate feature maps with normalized channel Gram matrices $\mathbf{G}_l$. Over the target boundary band ∂m_{opt}, manifold alignment matches the host's texture and lighting characteristics:

$$\mathcal{L}_{align} = \sum_{l \in \mathcal{L}} w_l \left\| \mathbf{G}_l(\hat{x}_0 \odot \partial m_{opt}) - \mathbf{G}_l(I_{anc} \odot \partial m_{opt}) \right\|_F^2. \tag{7}$$

The boundary loss $\mathcal{L}_{bnd} = \|\nabla^2 \hat{x}_0 \odot \partial M_{pixel}\|_2^2$ separately penalizes seams. With $\mathcal{J} = \lambda_{ret}\mathcal{L}_{ret} + \lambda_{gen}\mathcal{L}_{gen} + \lambda_{align}\mathcal{L}_{align} + \lambda_{bnd}\mathcal{L}_{bnd}$, the guided velocity is

$$\hat{v}_\theta = v_\theta + \eta \nabla_{z_t} \mathcal{J}, \qquad z_{t-\Delta t}^{guided} = z_t - \Delta t\, \hat{v}_\theta. \tag{8}$$

The positive velocity correction yields a negative guidance increment under reverse Euler integration. To avoid costly, unstable per-step differentiation through the complete velocity predictor, we treat its

field as locally stationary: $\partial \hat{z}_0 / \partial z_t \approx \mathbf{I}$. We then replace the raw guidance gradient with Spatially-Gated Gradient Scaling (SGGS). Writing $\mathbf{g}_{pixel} = \nabla_{\hat{x}_0} \mathcal{J}$ and $\mathbf{J}_{\mathcal{D}_{vae}}$ for the decoder Jacobian, the stabilized gradient is

$$\widetilde{\nabla}_{z_t} \mathcal{J} = \mathbf{J}_{\mathcal{D}_{vae}}^{\top} \left[m_{opt} \odot \left(\gamma_t \frac{\mathbf{g}_{pixel}}{\|\mathbf{g}_{pixel}\|_2 + \epsilon} \right) \right], \qquad \gamma_t = 1 - t. \tag{9}$$

The spatial gate confines the gradient source to the canvas, while the warm-up suppresses early noise-dominated updates. SGGS is an empirical constrained-update rule; Appendix B.4 gives the full gradient chain and a stability experiment.

Dual-space soft blending. We Gaussian-feather the pixel mask ($\sigma = 3$) and bilinearly downsample it to M_{latent}. Each step blends the guided state with the host's unperturbed interpolation:

$$z_{t-\Delta t} = M_{latent} \odot z_{t-\Delta t}^{guided} + (\mathbf{1} - M_{latent}) \odot z_{t-\Delta t}^{orig}. \tag{10}$$

Here $z_{t-\Delta t}^{orig} = (1 - (t - \Delta t)) z_0 + (t - \Delta t) z_1$ uses the same deterministic host interpolation at every step. Editing starts at $t_{start} = 0.80$ for 40 steps of a 50-step uniform Euler schedule ($\Delta t = 0.02$). Unlike a hard binary latent cut, Gaussian feathering avoids mask-edge step discontinuities; the pixel-space transition field $\partial M_{pixel} = \|\nabla M_{pixel}\|_2$ localizes the boundary penalty. Repeated background resetting, soft transitions, and $\mathcal{L}_{bnd}$ suppress perceptual drift. Appendices B.3 and E.2 give the full formulation and hard/soft blending comparison.

3.4 Verification via Cluster-Robust Hypothesis Testing

For each of $N = 500$ anchors, the auditor issues $K = 4$ diverse, signature-free queries and records response $R^{(i,j)}$. Let $Z_{ij} = \mathbb{I}[\mathrm{Norm}(S_{trap}^{(i)}) \subseteq_{\mathrm{tok}} \mathrm{Norm}(R^{(i,j)})]$, where $\subseteq_{\mathrm{tok}}$ means a contiguous sequence of complete normalized tokens (Appendix A.1). We report unconditional end-to-end query success and anchor-level extraction frequency:

$$\mathrm{CGSR}_{query} = \frac{1}{NK} \sum_{i,j} Z_{ij}, \qquad \mathrm{CER}_i = \frac{1}{K} \sum_{j} Z_{ij}. \tag{11}$$

Unlike success conditional on retrieval, this metric includes filtering and retrieval failures. All baselines use the same protocol and un-truncated global retrieval ranks.

Correlated probes of one image are collapsed into an anchor-level unit. The **Union** indicator is $Y_i^{union} = \max_j Z_{ij}$ (sensitivity), whereas the primary **All-K Intersection** indicator is $Y_i^{inter} = \prod_j Z_{ij}$ (cross-query reproducibility). Against a matched, un-ingested clean system, let b and c count positive and negative discordant anchor pairs. The operational null is $\mathcal{H}_0 : p_{suspect,i} \leq p_{clean,i}$ for every anchor. Under cross-anchor conditional independence, the conservative one-sided paired test is

$$p_{\mathrm{McNemar}} = \sum_{k=b}^{b+c} \binom{b+c}{k} 2^{-(b+c)}. \tag{12}$$

Stateless requests, permuted dispatches, and low-temperature decoding mitigate identifiable dependence. Mapping non-ingestion to this null also assumes that an innocent system has no systematically higher background collision rate than its matched clean baseline.

The paired test allows semantic heterogeneity. Let $u_i = P(Y_i^{suspect} = 1, Y_i^{clean} = 0)$ and $v_i = P(Y_i^{suspect} = 0, Y_i^{clean} = 1)$. Under $\mathcal{H}_0$, $u_i \leq v_i$, so conditional on an anchor being discordant, its positive-discordance probability is $u_i/(u_i + v_i) \leq 1/2$. Given the discordant anchor set and cross-anchor independence, the positive count is stochastically dominated by $\mathrm{Binomial}(b + c, 1/2)$, justifying Equation 12 without assuming identical anchor success probabilities.

For zero clean All-K matches, $\prod_i (1 - p_i) \leq (1 - \bar{p}_{\mathrm{clean}})^N$ gives the conservative 99% bound $\bar{p}_{\mathrm{clean}} \leq 1 - 0.005^{1/500} \leq 1.06\%$ on the portfolio collision rate. Appendix A.4 calibrates complete audit decisions under the non-ingestion configurations and gives simultaneous bounds without requiring equal collision probabilities in every family. When no matched service exists, the owner privately draws one of eight presealed candidates per permitted host. Queries omit every candidate; H_i retains candidates present in all four answers. Conditional on an assignment-independent null transcript, the selected-hit count has generating function $\prod_i (1 - |H_i|/8 + (|H_i|/8)z)$, giving an exact private-assignment tail without independent service responses (Appendix G).

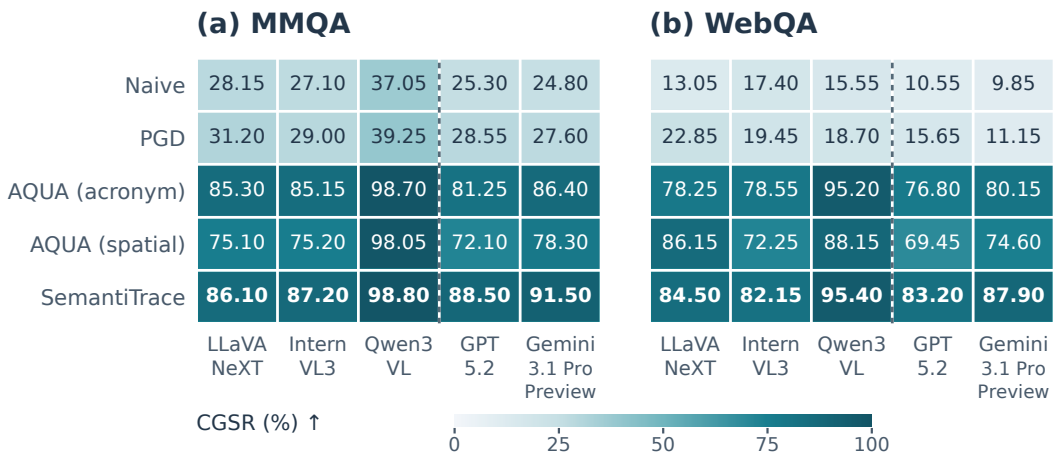


Figure 3: Pre-sanitization query-level CGSR (%, higher is better): five generators, both datasets. AQUA (acronym) is the restricted structured control; the dashed divider separates open-weights from commercial generators. Table 1 compares native queries; Tables 9 and 9 (Appendix C) give complete global ranks and p-values.

4 Experiments

4.1 Setup and Evaluation Protocol

We study **stealthiness**, **extraction efficacy**, and **statistical robustness** under a frozen end-to-end evaluation protocol. Appendix A specifies the cohorts, failure handling and audit sampling rules; Appendix B.3 gives the tensor interfaces and preprocessing. Each MMQA (Talmor et al., 2021) and WebQA (Chang et al., 2022) corpus receives $N = 500$ canaries, each tested with $K = 4$ probes. The default retriever is CLIP-ViT-L/14 (Radford et al., 2021) (top-$k = 5$); generators are LLaVA-NeXT-7B, InternVL3-8B, Qwen3-VL-8B-Instruct, GPT-5.2, and Gemini 3.1 Pro Preview, with $T \leq 0.2$. Request and model configurations are frozen (Appendix A.1).

Generation uses FLUX.2-klein-base-9B and a Qwen3-VL-8B-Instruct surrogate. The guidance weights are $(\lambda_{ret}, \lambda_{gen}, \lambda_{align}, \lambda_{bnd}) = (2.5, 4.0, 1.0, 1.5)$ with $\eta = 0.20$. Calibration uses only the surrogate on an independent 50-image development split; the commercial-model sensitivity analysis is post hoc. Baselines include natural-image canaries (Naive), PGD ($L_\infty = 8/255$), AQUA's acronym/spatial variants, natural-scene AQUA with native acronym queries, and matched local editors based on AnyText2 (Tuo et al., 2024). Appendix A.3 gives scheduler, mining, baseline, and metric details. Scoring, repair and initializer definitions are in Appendix B.6.

The pre-ingestion adversary combines a Mahalanobis detector on CLIP features (99th-percentile threshold) and a Qwen3-VL-32B global anomaly filter. We distinguish the **dense fidelity evaluation** ($M = 10{,}000$, 100% injection) from the **forensic audit cohort** ($N = 500$ per dataset). The dense pool has 98.20% survival; audit-cohort survival is $491/500 = 98.20\%$ on MMQA and $489/500 = 97.80\%$ on WebQA. Post-filter recall and end-to-end extraction retain all $NK = 2000$ queries in their denominator. Observable string matches also include signature-specific background collisions; Appendix A.3 separates those from the in-context retrieval–decoding pathway.

4.2 Extraction Efficacy and Transfer

Figure 3 shows that retrieval-oriented PGD does not reliably bridge the modality gap: its CGSR stays below 40%. SemantiTrace reaches 82.15–98.80% across the evaluated open-weights generators and 83.20–91.50% on the held-out generator configurations. Unlike spatial hallucinations, its plausible local triggers can be read as scene evidence by aligned generators.

The retrieval benefit also transfers beyond the optimization-time CLIP index. On unseen SigLIP (Zhai et al., 2023), DINOv2 with `dino.txt`, and ConvNeXt/OpenCLIP interfaces, pre-filter Recall@5 is 89.45–95.40%. Caption-indexed retrieval reaches 91.80% on MMQA and 88.50% on WebQA, whereas PGD's noise is not transcribed into signatures. Appendix D gives these interface definitions together with OCR-first and ColPali evaluations.

Table 1: Complete-pipeline comparisons after global filtering on both fixed 500-host portfolios. CGSR uses all 2,000 queries per dataset; All-K and rejection count anchors out of 500. Native AQUA probes supply displayed initials, never the complete verification phrase.

	MMQA			WebQA		
Construction / query	CGSR (%)↑	All-K↑	Rejected↓	CGSR (%)↑	All-K↑	Rejected↓
AQUA structured / restricted	3.10	0	471	2.15	0	471
AQUA natural / restricted	64.20	248	90	58.30	211	110
AQUA aligned structured / native	58.65	239	161	54.50	223	188
AQUA natural / native	71.40	273	90	66.95	239	110
AnyText2 + DSSB	69.85	262	21	64.30	227	28
AnyText2 + full dual guidance	82.45	353	16	78.45	291	18
SemantiTrace	**86.90**	**400**	**9**	**81.70**	**372**	**11**

Restricted-query and native-query natural AQUA use identical constructed images and filter outcomes. Aligned structured construction includes every scene landmark used by the query. AQUA rows are our re-instantiations of its published constructions. The guided AnyText2 row includes both modes and all native rendering costs; it receives the same retrieval, generation, alignment, boundary, mask and gradient-gate settings.

Table 2: Post-sanitization audit on WebQA/GPT-5.2. Union/All-K count successful anchors out of 500; CGSR uses all 2,000 probes. Paired All-K is primary. Native AQUA uses its own clean arm (7/2,000 query hits; 1/500 All-K).

Method	Query CGSR (%)↑	Union↑	All-K↑	$p_{\text{All-}K}$↓
Clean baseline	0.05	1	0	–
AQUA structured / restricted	2.15	18	0	1.0000
AQUA natural / native	66.95	389	239	1.36×10^{-70}
SemantiTrace	**81.70**	**469**	**372**	$\mathbf{1.04 \times 10^{-112}}$

4.3 Stealthiness and Post-Sanitization Verification

Table 10 in Appendix C.2 measures intrinsic editing fidelity without dilution by sparse deployment. SemantiTrace has lower LPIPS than PGD (0.044 vs. 0.084), while its FID_{10K} stays close to the clean reference (13.85 vs. 12.45). AQUA's global anomalies are frequently rejected (94.2% for acronyms; 82.4% for spatial canaries), compared with 1.8% for SemantiTrace on this dense evaluation (Figure 1, left).

Filtering changes practical extraction sharply. The restricted structured AQUA ablation falls from 76.80% to 2.15% on WebQA, whereas the stronger natural/native control retains 66.95% (Table 1). SemantiTrace changes from 83.20% to 81.70%. Against the closest full-guidance editor, same-host AnyText2 with full dual guidance (AT-full), its WebQA increments are 3.25 points in query success and 16.20 points in All-K. The manifold, waterfall, and extraction-distribution visualizations are in Appendix C.2.

Union still detects AQUA's sparse surviving hits ($18/500$ vs. $1/500$, paired $p = 3.81 \times 10^{-5}$; Table 2); All-K is our pre-frozen reproducibility criterion. SemantiTrace retains $372/500$ All-K successes versus zero clean matches, giving $b = 372, c = 0$ and $p = 2^{-372}$. Its binomial-reference 99% interval is $[69.05\%, 79.27\%]$; an independent, heterogeneous-anchor Hoeffding interval is $[67.12\%, 81.68\%]$, also separated from the clean zero-event bound. The nested rates $74.40\% \leq 81.70\% \leq 93.80\%$ distinguish reproducibility, query efficacy, and sensitivity.

Natural-scene, output-only auditing and corpus-wide repair. Figure 4 connects the same 500-host strata to owner-permitted fields, eight private candidates and four assignment-independent queries. Failed permissions, catalogues and renders remain in every denominator; hidden visual (PV) and raw-OCR (PO) services receive no candidate hints. The full 400,000-item repair experiment retains $1376/2000$ query hits and $287/500$ All-K anchors, versus $1302/2000$ and $246/500$ for AT-full, at $877/1000$ normal-QA accuracy. When permissions, private assignment, hidden PO and the same repair hold *jointly*, the rates are $56.80\%/52.50\%$, the All-K counts $220/170$, and 95%-coverage costs 64/92 output queries with no clean arm. Fresh repaired 8/64-ingestion audits give $165/400$ versus $125/400$ declarations, with a 10.00-point paired gain; the overall normal-QA loss is 0.60

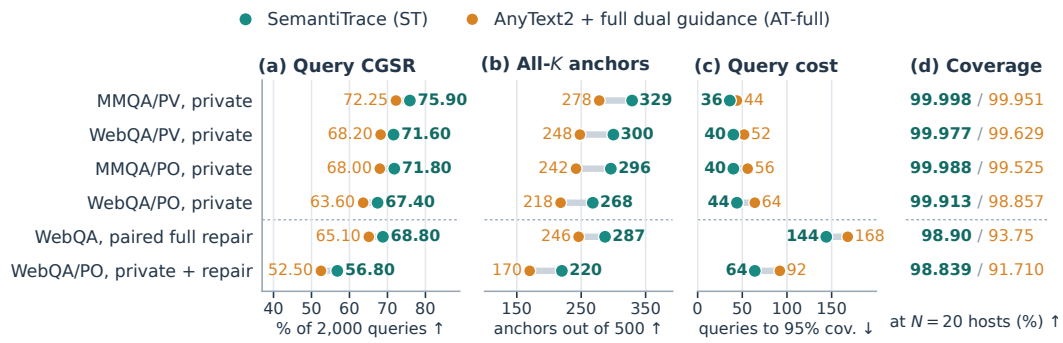


Figure 4: Deployment comparison of SemantiTrace (ST) and the closest complete-objective editor (AT-full): 500 hosts/2,000 probes per primary portfolio. PV/PO are hidden visual/OCR-text services; (c) output queries for 95% coverage; (d) coverage at $N = 20$ hosts. Both editors share permissions, candidates, assignments, build budget and hidden stack in private rows; these need no clean queries, including joint repair. Paired repair counts both arms. Coverage is exact on each fixed transcript population. Tables 41 and 46 report fresh reingestion audits.

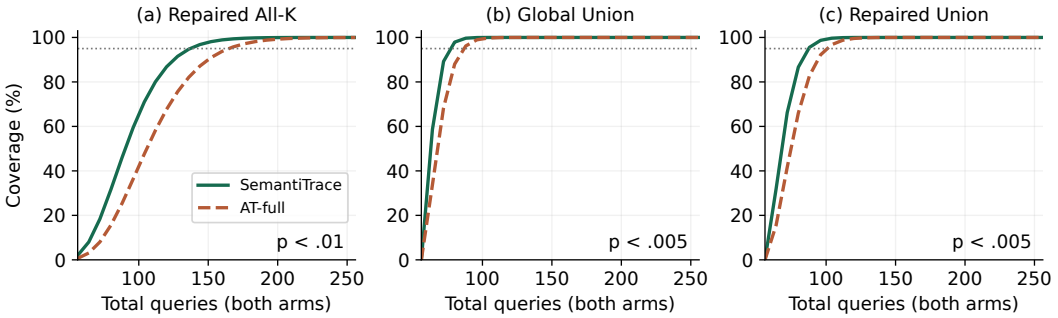


Figure 5: Complete-audit query costs against the closest full-guidance editor. (a) Primary repaired All-K, $p < .01$. (b,c) Independently development-selected Union for both editors, common $p < .005$, under global filtering/full-gallery repair. Curves give exact finite-cohort coverage, counting both arms. Rules: Tables 44 and 51 (Appendices G.5 and H.2).

points with simultaneous 99% upper bound 1.231 points (Appendix G.6). The raw-OCR schema, original-clue strata and an image-to-answer example are given in Appendices D.1 and D.2.

Endpoint-fair query efficiency. At fixed $K = 4$, independently selected Union costs 80/88 paired queries globally and 88/104 after repair (Figure 5b,c), with both arms billed and separately calibrated false alarms (Appendix H.2). Allowing $K \in \{1, 2, 4\}$ selects one-probe Union for both editors. The paired global/repair costs and their 2,000-host confirmation are in Appendix H.3. For the joint private deployment, a separate development set also selects $K = 1$. On 4,000 new hosts, ST needs 11 queries versus 13 for both AT-full and independently tuned AnyText2, a 15.38% saving in the point estimate with paired 99% saving ranges of $[1, 3]$ queries against each. Fresh audits of this complete-ingestion 11-query design (C11) give $384/400$ ST declarations. At 8/64 ingestion, 64-query power is 71.00% versus 61.25%/63.75% for AT-full/tuned AnyText2; under the development-selected defense D3 it is 64.50% versus 51.50%/54.25%. All five budgets of the selected $K = 1$ design under D0 and D3 have simultaneous 99% false-alarm upper bounds below 1% (Appendix A.4).

4.4 Ablations, Robustness, and Efficiency

Figure 6 identifies complementary bottlenecks (Appendix E). Removing retriever guidance gives $0/2000$ top-5 placements in each portfolio and 0% CGSR (Table 25); removing generator guidance preserves retrieval but reduces MMQA/GPT-5.2 extraction to 12.10%. Removing blending permits high pre-filter decoding but raises pooled OOD rejection to 86.3%. The full configuration balances these objectives. Mode A is preferred where text exists, but Mode B covers text-scarce images (73.4% of WebQA hosts). Alignment guidance reduces Mode B's rejection from 5.20% to 2.77%. Cross-family surrogate substitutions retain 80.40–86.80% commercial-model CGSR across the

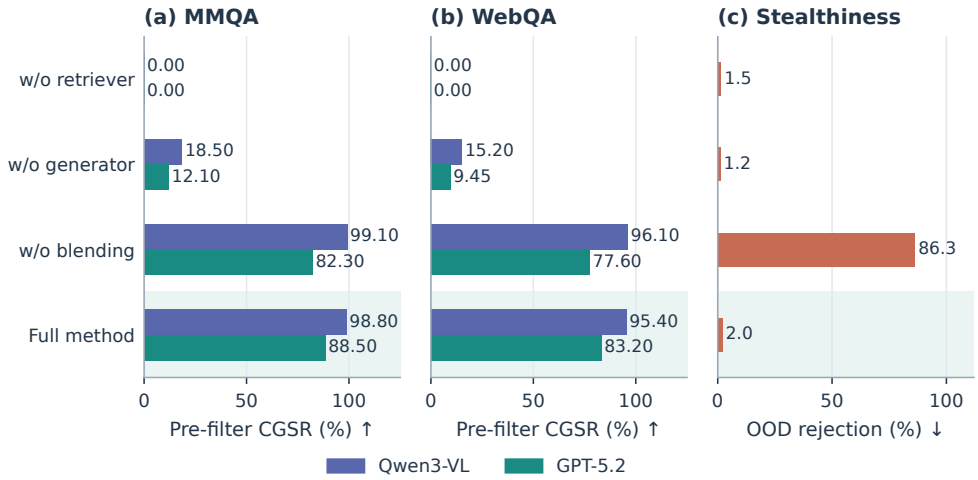


Figure 6: Core ablations (Table 24). (a,b) Pre-filter decoding efficacy; (c) pooled-cohort OOD rejection (lower is better). Without blending, high pre-filter CGSR accompanies high OOD rejection.

reported dataset configurations. Appendix E consolidates mode, surrogate, boundary, typography, and parameter diagnostics.

Matched controls and attribution calibration. On the same 423 Mode-A hosts, AnyText2 plus DSSB has higher isolated OCR exactness but lower pipeline extraction than ST (Appendix C.3). With the *same full dual guidance*, five-run CGSR/All-K gains are $3.78/14.61$ points on Mode A and $3.91/11.13$ on all 577 Mode-B hosts. Giving AnyText2 its own development-selected re-noising strength raises its initialization coefficient from .20 to .60; ST still gains 8.80/5.40/6.60 All-K points under global filtering, repair and joint PO. A fully guided, independently tuned glyph initializer is also inferior in all three conditions; all six contrasts pass Holm correction (Appendix E.6). Against global clipping, the five-run host-cluster gain is 4.07 points. Under joint deployment, soft blending, boundary loss and spatial gating give 13.40/8.40/5.00-point conditional All-K gains, all Holm-significant (Appendix E.5). Native private AQUA with fixed initials needs 116 queries (Appendix C.3). Sealed stacks retain 63.45–72.90% CGSR without component access (Appendix A.5).

Deployment robustness and utility. SemantiTrace retains 68.30–85.20% CGSR under JPEG Q=75 and 71.95–78.40% on GPT-5.2 under DiffPure (Nie et al., 2022). A non-oracle OCR inspector capped at four crops retains 68.80% WebQA CGSR and $310/500$ All-K, at 0.98% clean-image rejection and 4.12 s/image; the matched AnyText2 control reaches 51.30% CGSR (Appendix D.4). Sparse-deployment macro F1 remains $68.42\%/65.15\%$. Deployment uses the permission manifest: it covers $963/1000$ hosts and preserves $220/237$ immutable-field answers versus $221/237$ clean (Appendix D.5). A defense selected to minimize development audit power removes the low-confidence gate; held-out P64 power (8 of 64 hosts ingested) remains 36.25%/25.25% for ST/AT-full, with simultaneous utility-loss upper bounds 1.225/1.216 points (Appendix G.7).

Efficiency and query variation. HNSW mining preserves a 96.08% mean anchor-set Jaccard overlap at 10^7 candidates with a -0.18% downstream CGSR delta. On an A100, quantization plus stride-2 guidance uses 14.2 GB and 50.8 s/image, retaining 84.30% MMQA/GPT-5.2 CGSR (Appendix F). Concept-Space Generalized Optimization improves the global-scene diagnostic from 43.35% to 79.60% on MMQA/GPT-5.2 (Appendix B.5).

5 Conclusion

SemantiTrace shifts black-box Multimodal RAG watermarking from anomalous synthesis to semantic parasitism. Context-aware local triggers, dual guidance, and soft blending jointly support retrieval, explicit decoding, and survival under the evaluated sanitization pipelines. The evaluation connects both local-edit modes to strong full-guidance controls, endpoint-fair query efficiency, and permission-constrained private assignment with full-corpus repair/reindexing in one deployment. Pre-frozen paired and private-assignment audits supply statistically grounded ingestion evidence with explicit false-alarm and query budgets. Future work will extend this framework to Video-RAG.

AI USE STATEMENT

We used generative AI tools to polish the language and improve the readability of this manuscript. The authors take full responsibility for the final content.

ETHICS STATEMENT

This work studies dataset provenance auditing, but semantic image edits could also be misused to contaminate retrieval corpora or manipulate model outputs. Deployment should be restricted to data that the auditor is authorized to modify and to appropriately authorized evaluations. Local edits should be documented and assessed for their effect on benign use. Statistical matches must be interpreted with the stated collision, independence, and matched-baseline assumptions, alongside provenance records and other evidence; they do not by themselves establish legal ownership or infringement. Dataset licenses, access restrictions, and applicable service terms should be respected.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B specify the objective functions, tensor interfaces, masks and optimization schedule; Appendix B.6 defines the local proposal score, lexical repair and initializers. Appendix A gives the cohorts, request configuration, normalization, failure handling and statistical decision rules, and Appendix A.4 the non-ingestion calibration of complete audits. Probe templates and generation/filter prompts are in Appendix I. Every reported rate states its denominator and conditioning set, and the deployment, repair and query-efficiency experiments (Appendices G and H) report the integer endpoints of each audit.

REFERENCES

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

Inzamamul Alam, Md Tanvir Islam, Khan Muhammad, and Simon S. Woo. Specguard: Spectral projection-based advanced invisible watermarking. *arXiv preprint arXiv:2510.07302*, 2025. URL https://arxiv.org/abs/2510.07302.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18187–18197. IEEE, 2022.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592450. URL https://doi.org/10.1145/3592450.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 51304–51323, 2024.

Linn Bieske and Carla Lorente. Prompt-to-prompt: Text-based image editing via cross-attention mechanisms – the research of hyperparameters and novel mechanisms to enhance existing frameworks. *arXiv preprint arXiv:2510.04034*, 2025. URL https://arxiv.org/abs/2510.04034.

Black Forest Labs. FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025. URL https://bfl.ai/research/representation-comparison.

Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 407–425. IEEE, 2024.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16495–16504, 2022.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Tianyu Chen, Jian Lou, and Wenjie Wang. Safeguarding multimodal knowledge copyright in the rag-as-a-service environment. *arXiv preprint arXiv:2506.10030*, 2025.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Nazarii Drushchak, Nataliya Polyakovska, Maryna Bautina, Taras Semenchenko, Jakub Koscielecki, Wojciech Sykala, and Michal Wegrzynowski. Multimodal retrieval-augmented generation: Unified information processing across text, image, table, and video modalities. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, pp. 59–64, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024. URL https://arxiv.org/abs/2407.01449.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Tianle Gu, Zongqi Wang, Kexin Huang, Yuanqi Yao, Xiangliang Zhang, Yujiu Yang, and Xiuying Chen. Invisible entropy: Towards safe and efficient low-entropy llm watermarking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6716–6733, 2025. doi: 10.18653/v1/2025.emnlp-main.341.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. Ward: Provable rag dataset inference via llm watermarks. In *International Conference on Learning Representations*, volume 2025, pp. 93288–93314, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. Enhancing retrieval-augmented generation: a study of best practices. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6705–6717, 2025.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15424–15433, 2025. doi: 10.1109/CVPR52734.2025.01437.

Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. Who stole your data? a method for detecting unauthorized rag theft. *arXiv preprint arXiv:2510.07728*, 2025a.

Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. Dataset protection via watermarked canaries in retrieval-augmented llms. *arXiv preprint arXiv:2502.10673*, 2025b.

Peizhuo Lv, Mengjie Sun, Hao Wang, Xiaofeng Wang, Shengzhi Zhang, Yuxuan Chen, Kai Chen, and Limin Sun. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1709–1723, 2025.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. URL https://arxiv.org/abs/2205.07460.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2024. URL https://arxiv.org/abs/2304.07193.

Leyi Pan, Aiwei Liu, Shiyu Huang, Yijian Lu, Xuming Hu, Lijie Wen, Irwin King, and Philip S Yu. Can llm watermarks robustly prevent unauthorized knowledge distillation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13228–13251, 2025.

Shinwoo Park, Hyejin Park, Hyeseon Ahn, and Yo-Sub Han. Watermod: Modular token-rank partitioning for probability-balanced llm watermarking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32683–32691, 2026. Issue 39.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *ACM Transactions on Information Systems*, 44(2):1–52, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335. PMLR, 2020.

Tom Sander, Pierre Fernandez, Saeed Mahloujifar, Alain Durmus, and Chuan Guo. Detecting benchmark contamination through watermarking. *arXiv preprint arXiv:2502.17259*, 2025.

Yunzhou Song, Long Le, Yong-Hyun Park, Jie Wang, Junyao Shi, Lingjie Liu, Jiatao Gu, Eric Eaton, Dinesh Jayaraman, and Kostas Daniilidis. Omniguide: Universal guidance fields for enhancing generalist robot policies. *arXiv preprint arXiv:2603.10052*, 2026. URL https://arxiv.org/abs/2603.10052.

Yuhao Sun, Yihua Zhang, Gaowen Liu, Hongtao Xie, and Sijia Liu. Invisible watermarks, visible gains: Steering machine unlearning with bi-level watermarking design. *arXiv preprint arXiv:2508.10065*, 2025. URL https://arxiv.org/abs/2508.10065.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6088–6109, 2025.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024. URL https://arxiv.org/abs/2411.15245.

Jerry Wang and Fang Yu. Derag: Black-box adversarial attacks on multiple retrieval-augmented generation applications via prompt injection. *arXiv preprint arXiv:2507.15042*, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025. doi: 10.1109/CVPR52734.2025.01464.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Xuan Long, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4736–4756, 2023.

Yan Zhong, Xinping Zhao, Guangzhi Zhao, Bohua Chen, Fei Hao, Ruoyu Zhao, Jiaqi He, Lei Shi, and Li Zhang. Ctd-inpainting: Towards the coherence of text-driven inpainting with blended diffusion. *Inf. Fusion*, 122(C), October 2025. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103163. URL https://doi.org/10.1016/j.inffus.2025.103163.

A Evaluation Protocol and Statistical Audit

A.1 Cohorts, requests and scoring

The primary realization R0 contains 500 MMQA and 500 WebQA attempted hosts, with four signature-free probes per host (Table 3). MMQA has 290 Mode-A and 210 Mode-B hosts, and WebQA has 133 Mode-A and 367 Mode-B hosts. R1–R4 use the same hosts with different construction and serving realizations. DENSE-10K is a separate 10,000-image fidelity evaluation, and the full galleries contain 58,000 MMQA items and 400,000 WebQA items. A nested repair diagnostic uses 1,500 items. The deployment manifest supplies 122/354 MMQA and 64/423 WebQA permitted Mode-A/Mode-B canvases, after 24/13 no-canvas failures; private catalogues add 6/10 catalogue failures.

Table 3: R0 cohort composition. Denominators are attempted hosts.

Dataset	Mode	Attempted	Rejected	Survived	Rejection
MMQA	A	290	3	287	1.03%
MMQA	B	210	6	204	2.86%
MMQA	A+B	500	9	491	1.80%
WebQA	A	133	1	132	0.75%
WebQA	B	367	10	357	2.72%
WebQA	A+B	500	11	489	2.20%
Pooled	A+B	1000	20	980	2.00%

The request interface uses stateless sessions and prepermuted dispatch. Open-weight decoding is greedy with a 128-new-token cap and sampling disabled; $T = 0$ and top-$p = 1$ denote that deterministic decoding convention. GPT requests use `gpt-5.2-2025-12-11`, reasoning effort `none`, temperature zero, top-$p = 1$, and a 128-output-token cap. Gemini uses `gemini-3.1-pro-preview`, thinking level `low`, temperature zero, top-$p = 1$, and a 128-token visible-answer cap; internal thinking is not counted as a visible response. The commercial serving window is 2026-09-01–2026-09-07 UTC; the Gemini preview identifier is an alias, so this date window is part of its instance specification rather than a claim of an immutable vendor snapshot. The open-weight checkpoints are `llava-hf/llava-v1.6-mistral-7b-hf`, `OpenGVLab/InternVL3-8B`, `Qwen/Qwen3-VL-8B-Instruct`, `Qwen/Qwen3-VL-32B-Instruct` (inspector), `Qwen/Qwen3-8B` with thinking disabled (text-only service), and `llava-hf/llama3-llava-next-8b-hf` (alternative surrogate). We use local snapshot copies retrieved on 2026-09-01 UTC with no subsequent model pulls. OCR uses PaddleOCR 3.2.0, `PP-OCRv5_server_det` and `PP-OCRv5_server_rec`, with document orientation and unwarping disabled; structural masks use SAM ViT-H `sam_vit_h_4b8939.pth`. SAM 2 is related work, not an interchangeable inference checkpoint in these runs. Failed builds, timeouts and overflows remain in the attempted-host denominator and keep their four query positions. Rendering uses one attempt per selected signature, with no replacement or failure-dependent dispatch. Budget curves are fixed-transcript or fixed-population calculations. Five-run blocks keep the same host, canvas and signature across runs.

The prospective private portfolios use the same editor and four-probe score with eight admissible candidates per host. Candidate formation precedes the private draw and preserves the scene, object, canvas and queries across candidates. MMQA and WebQA leave 470/477 assigned hosts after no-canvas and catalogue failures; SemantiTrace has 10/14 rendering failures in those two datasets, while AnyText2 with full guidance has 16/20. PV uses SigLIP visual retrieval with the Gemini-family generation configuration. PO uses PaddleOCR followed by BGE-M3 (Chen et al., 2024) retrieval with the GPT-family transcription configuration.

Matching applies Unicode NFKC, casefolding and maximal alphanumeric tokenization, with punctuation and whitespace as separators. Success requires the complete signature token sequence contiguously in the response. Thus `BSB` does not match `BSBank`, initials do not match a full verification phrase, and spelling variants are not credited. For an anchor with $h_i \in \{0, 1, 2, 3, 4\}$ matching probes,

$$Y_i^{\text{union}} = \mathbb{I}[h_i > 0], \qquad Y_i^{\text{inter}} = \mathbb{I}[h_i = 4].$$

The WebQA SemantiTrace R0 histogram $(31, 58, 29, 10, 372)$ gives 1,634 query hits, 469 Union anchors and 372 All-K anchors; the clean portfolio has one one-query hit and no All-K hit. Structured AQUA's $(482, 3, 5, 10, 0)$ gives 43 query hits and 18 Union anchors.

For private assignment, the scorer retains the full candidate hit set

$$H_i = \{a : \text{candidate } a \text{ appears as a complete token sequence in all four responses}\}.$$

Conditional on a fixed assignment-independent transcript, the selected-hit statistic $T = \sum_i \mathbb{I}[W_i \in H_i]$ has generating function

$$B(z) = \prod_{i=1}^{N} \left(1 - \frac{|H_i|}{8} + \frac{|H_i|}{8} z\right).$$

List-all responses therefore add one to the observed statistic and one to the reference. The natural private audit keeps empty, singleton, doubleton and list-all intersections; selected-candidate hits give the observed statistic.

A.2 Cluster-robust verification

For N anchors and K probes, the query-level endpoint is

$$\mathrm{CGSR}_{query} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{I}\Big[\mathrm{Norm}(S_{trap}^{(i)}) \subseteq_{\mathrm{tok}} \mathrm{Norm}(R^{(i,j)})\Big]. \tag{13}$$

The per-anchor empirical extraction rate is

$$\mathrm{CER}_i = \frac{1}{K} \sum_{j=1}^{K} \mathbb{I}\Big[\mathrm{Norm}(S_{trap}^{(i)}) \subseteq_{\mathrm{tok}} \mathrm{Norm}(R^{(i,j)})\Big] \in \{0, 1/K, \ldots, 1\}. \tag{14}$$

Union and All-K collapse correlated probes to an anchor-level unit before testing. Against a matched clean arm, let b and c be positive and negative discordant anchor pairs. The one-sided paired test is

$$p_{\mathrm{McNemar}} = \sum_{k=b}^{b+c} \binom{b+c}{k} 2^{-(b+c)}. \tag{15}$$

The test assumes conditionally independent anchor blocks given the specified indices and serving state, and an operational null $p_{suspect,i} \le p_{clean,i}$ for each anchor. Stateless requests, permuted dispatch and low-temperature decoding implement this design. Heterogeneous anchors are allowed: with $u_i = P(1, 0)$ and $v_i = P(0, 1)$, the null gives $u_i \le v_i$, so the positive-discordance probability conditional on discordance is at most $1/2$. The positive count is then stochastically dominated by $\mathrm{Binomial}(b + c, 1/2)$.

For zero clean All-K matches, $\prod_i (1 - p_i) \le (1 - \bar{p}_{clean})^N$ gives the 99% portfolio collision bound $\bar{p}_{clean} \le 1 - 0.005^{1/500} \le 1.06\%$. SemantiTrace's R0 WebQA All-K record has $b = 372, c = 0$, hence $p = 2^{-372} = 1.04 \times 10^{-112}$. The binomial-reference 99% interval is $[69.05\%, 79.27\%]$ and the heterogeneous-anchor Hoeffding interval is $[67.12\%, 81.68\%]$ for the observed 372/500 rate.

A.3 Implementation and metric decomposition

Experiments use MMQA and WebQA with CLIP-ViT-L/14 top-$k = 5$ retrieval and LLaVA-NeXT-7B, InternVL3-8B, Qwen3-VL-8B-Instruct, GPT-5.2 and Gemini 3.1 Pro Preview generators. SemantiTrace uses FLUX.2-klein-base-9B, Qwen3-VL-8B-Instruct as surrogate, 50 uniform reverse-Euler steps, an edit window from $t_{start} = 0.80$ for 40 steps, $\eta = 0.20$ and guidance weights $(2.5, 4.0, 1.0, 1.5)$. HNSW mining uses $M = 32$, $efConstruction = 128$, $efSearch = 64$ and $K = 10$ mining neighbors; CS-GO uses $P = 8$ query reformulations.

Pre-ingestion defense combines a CLIP Mahalanobis detector at the 99th percentile with a Qwen3-VL-32B global anomaly filter. The dense pool has 98.20% survival; audit-cohort survival is 491/500 on MMQA and 489/500 on WebQA. Along the observed pipeline,

$$P_{e2e} = P(S \cap R \cap G) = P(S)P(R \mid S)P_{gen}, \qquad \mathrm{CGSR}_{query}^{suspect} = P(S \cap R \cap A) + \epsilon_{out}^{suspect}, \tag{16}$$

where $\epsilon_{out}^{suspect} = P(A \cap (S \cap R)^c)$ records background matches after filtering or retrieval misses. On the clean SemantiTrace portfolio, one query out of 2,000 matches and one anchor matches under Union; no anchor matches under All-K. For spatial AQUA on WebQA, 224/2,000 queries retrieve a surviving target, 156/224 decode in context, and 22 of the remaining 1,776 match out of context, giving 8.90% query CGSR.

The post-sanitization WebQA/GPT-5.2 audit uses All-K at $\alpha = 0.01$. Union is a sensitivity endpoint. Under Union, clean has 1/500 anchors and SemantiTrace has 469/500 anchors with exact 99% interval $[90.48\%, 96.25\%]$. Under All-K, SemantiTrace has 372/500 anchors, clean has 0/500, and structured AQUA has 0/500 with Union discordances $(18, 1)$ and $p = 3.81 \times 10^{-5}$.

Native-query AQUA uses its own clean arm and native null calibration because acronym cues change the query distribution. Endpoint-fair score selection has independent service-null calibration below; its development choice and query-cost records are reported with the query-efficiency appendix.

A.4 Whole-audit false-alarm calibration

The whole-audit false-alarm probability is the probability that a complete non-ingestion transcript declares ingestion. For paired audits, the decision event is $D(T) = \mathbb{I}[p_{\mathrm{McNemar}}(T) < 0.01]$. For G settings with M independent complete null audits per setting and zero declarations in one setting, a simultaneous 99% upper endpoint is

$$u_G = 1 - (0.01/G)^{1/M}.$$

This union-bound statement applies to complete audit declarations and permits shared fixed prefixes or scores across views. Reference-law rows below are exact calculations under their printed probability laws.

For a categorical host law with positive and negative discordance probabilities (p_+, p_-), set $q = p_+ + p_-$ and

$$\kappa_d = \min\{k : P(\mathrm{Binomial}(d, \tfrac{1}{2}) \geq k) < .01\}.$$

The complete-audit rejection rate is

$$r_N(p_+, p_-) = \sum_{d=0}^{N} \binom{N}{d} q^d (1-q)^{N-d} P\left\{\mathrm{Binomial}\left(d, \frac{p_+}{q}\right) \geq \kappa_d\right\},$$

with $r_N(0, 0) = 0$. Method-native score laws use $M \sim \mathrm{Binomial}(N, q_g)$ with $q_g = .05 + .025g$, $g = 0, \ldots, 7$. SemantiTrace uses unit weights. AQUA weights are uniform in $\{1, 2, 3, 4\}$, with conditional reference polynomial $2^{-M} \prod_{w=1}^{4} (1 + z^w)^{a_w}$ for weight counts $\mathbf{a} = (a_1, \ldots, a_4)$.

The categorical reference families in Table 4 use

$$(p_+, p_-) = \begin{cases} (.0007, .0006), & \text{All-}K \text{ categorical,} \\ (.005, .004), & \text{native acronym cue,} \\ (.012, .014), & \text{defended All-}K. \end{cases} \tag{17}$$

These are specified background-collision stress laws, not an assertion that every non-ingestion service satisfies the paired operational null; an asymmetric law with $p_+ > p_-$ deliberately checks baseline mismatch. For method-native scores, conditional on M the signs are independent fair Rademacher draws, independent of the weights, and rejection uses the inclusive sign-swap tail at $p < .005$. The maximum is over both score types and the eight q_g values. For the natural-private reference row every $h_i = 1$, $T \sim \mathrm{Binomial}(N, 1/8)$, and rejection is the strict inclusive-tail event $p < .01$. A nonexistent critical value is set to $d + 1$ (no rejection).

The service-family calibration fixes the corpus partition, retrieval interface, prompt and state policy before scoring. N1 uses the canonical CLIP visual stack and default GPT-5.2 prompt. N2 changes the background pool. N3 uses caption-based BGE-M3 retrieval and the alternative transcription prompt. N4 uses PaddleOCR followed by BGE-M3 retrieval. N5 uses ColPali (Faysse et al., 2024) with caches cleared between complete audits. N6 uses a near-duplicate background stratum, a changed prompt and Gemini 3.1 Pro Preview. N7 enlarges the gallery with legitimate rare-name, brand and serial-label images. N8 changes background pool, retrieval interface, prompt and generator together. Table 4 gives the exact reference-law checks, and Table 5 gives service non-ingestion calibration.

Table 4: Consolidated null calibration. Exact-law entries are finite calculations. Service rows are complete non-ingestion audit transcripts.

Exact reference-law checks		
Family	Budgets	Largest exact rejection probability
All-K categorical	$N = 20; 500$	$6.284 \times 10^{-16}\%; 6.734 \times 10^{-6}\%$
Native acronym cue	$N = 20; 500$	$5.429 \times 10^{-10}\%; 0.199267\%$
Method-native scores	$N = 10, 12, 20, 32$	$0.000072\%; 0.000500\%; 0.021016\%; 0.144335\%$
Natural private assignment	$N = 20, 64$	$0.836683\%; 0.465460\%$
Defended All-K	$N = 20; 80; 500$	$2.012 \times 10^{-7}\%; 0.001873\%; 0.168475\%$

Table 5: Service non-ingestion calibration. Bounds are simultaneous 99% upper endpoints within each family.

Protocol	Setting	Budget or design	Alarms / rates	99% upper
Matched private	MMQA/PV	$N = 20; 64$	2/1200; 4/1200	$0.910\%; 1.203\%$
Matched private	WebQA/PV	$N = 20; 64$	1/1200; 5/1200	$0.746\%; 1.339\%$
Matched private	MMQA/PO	$N = 20; 64$	2/1200; 5/1200	$0.910\%; 1.339\%$
Matched private	WebQA/PO	$N = 20; 64$	1/1200; 6/1200	$0.746\%; 1.471\%$
Defended paired	DN1–DN8, two vocab.	$N = 20$	0/1000 each	0.735061%
Endpoint selection	Global; repair	$n = 8$–32	0/1200; 0/1200	$0.799\%; 0.799\%$
Joint private repaired	C20; P64	C20 prefix; P64	6/4000 (0.15%); 14/4000 (0.35%)	$0.392\%; 0.670\%$
Supplementary	AQUA-FI C20; P64	C20; P64	4/2000 (0.200%); 5/2000 (0.250%)	$0.677\%; 0.757\%$
Supplementary	Selected defense C20; P64	C20; P64	9/4000 (0.225%); 16/4000 (0.400%)	$0.529\%; 0.771\%$
Selected $K = 1$	D0	11; 13; 64; 128; 256 queries	4/4000 (0.100%); 6/4000 (0.150%); 12/4000 (0.300%); 15/4000 (0.375%); 18/4000 (0.450%)	$0.370; 0.451; 0.675; 0.780; 0.882\%$
Selected $K = 1$	D3	11; 13; 64; 128; 256 queries	5/4000 (0.125%); 7/4000 (0.175%); 15/4000 (0.375%); 17/4000 (0.425%); 20/4000 (0.500%)	$0.411; 0.490; 0.780; 0.849; 0.949\%$

The observed false-alarm rate is at most 0.50% in every service setting. The selected $K = 1$ private design has one 99% event covering both D0 and D3 and all five budgets. The defended paired service row is the repaired-service calibration with $G = 16$, $M = 1000$ and zero declarations, giving $u_G = 0.735061\% < 0.74\%$.

At an equal false-alarm target and equal query budget, every method uses $N = 20$, $K = 4$, $p < 0.01$, 80 suspect and 80 clean queries. Table 6 reports the fixed run-by-block coverage.

Table 6: Detection at the same query budget and 1% nominal false-alarm target. Blocks define run-by-block coverage.

Method	Anchors	Queries	Positive audits	Detection rate
AQUA structured scene	20	80+80	0/250	0.00%
AQUA natural scene	20	80+80	228/250	91.20%
Prompt-only + DSSB	20	80+80	25/250	10.00%
AnyText2 + DSSB refinement	20	80+80	226/250	90.40%
SemantiTrace	20	80+80	250/250	100.00%

A.5 SEALED-STACK AUDITS

In the sealed-stack evaluation, canaries and probes stay constant while the service administrator selects the hidden retriever, filter, prompt and generator (Table 7). The auditor sees only text responses; stack identities are disclosed afterwards for analysis.

These hidden configurations retain 63.45–72.90% CGSR with positive paired All-K tests under output-only access.

Table 7: Sealed-stack challenge with fixed signatures and probes.

Service	Retriever	Defense	Generator	CGSR	All-K	(b, c)	p
H1	SigLIP	OCR-2	GPT-5.2	71.85%	306/500	(305,0)	1.53×10^{-92}
H2	ColPali	OCR-4	Gemini 3.1 Pro Preview	72.90%	317/500	(316,1)	1.19×10^{-93}
H3	OCR+BGE-M3	global+JPEG75	GPT-5.2	63.45%	256/500	(256,1)	1.11×10^{-75}
H4	caption+BGE-M3	DiffPure	Gemini 3.1 Pro Preview	65.20%	269/500	(268,1)	2.85×10^{-79}

B Method and Optimization Details

B.1 Anchor mining

Anchor quality controls two failure modes. A cluttered host makes a local trigger visually noisy after editing, while a host in a dense retrieval neighborhood can be displaced from the top-k context window by nearby distractors. SemantiTrace therefore mines one anchor per semantic cluster so that coverage spans the source corpus while each selected host has low local competition and an editable canvas.

For image I_i, CLIP image features define the retrieval-isolation prior

$$\mathcal{S}_{iso}(I_i) = \frac{1}{K} \sum_{I_k \in \mathcal{N}_K(I_i)} \left(1 - \frac{\Phi_{img}(I_i) \cdot \Phi_{img}(I_k)}{\|\Phi_{img}(I_i)\| \|\Phi_{img}(I_k)\|}\right).$$

A high value indicates sparse local neighbors in the feature manifold. The prior reduces neighbor interference before the guided loss attempts to pull the edited anchor toward the probe query.

OCR supplies textual masks $\mathcal{M}_{text}$ and SAM supplies structural masks $\mathcal{M}_{struct}$. Text is preferred because mutating existing glyphs preserves object layout, material support and scene semantics. For text mask m_j, $\mathcal{S}_{text}(m_j) = \mathrm{Area}(m_j)\mathrm{Conf}_{OCR}(m_j)$. If no viable text exists, structural canvases are ranked by area and surface smoothness. The Laplacian is taken on the continuous host image before masking, and variance is computed on the eroded interior $\Omega_{int}(m_j) = \mathrm{erode}(m_j, \mathcal{K}_\rho)$ with $\rho = 3$ pixels. This avoids counting the mask boundary as surface texture:

$$\mathcal{S}_{struct}(m_j) = \frac{\mathrm{Area}(m_j)}{\mathrm{Var}_{x \in \Omega_{int}(m_j)}[\nabla^2 I_i(x)] + \epsilon}. \tag{18}$$

The textual route selects

$$I_{anc} = \arg \max_{I_i \in \mathcal{D}_{cluster}} \left[\mathcal{S}_{iso}(I_i) \max_{m_j \in \mathcal{M}_{text}} \mathcal{S}_{text}(m_j)\right]. \tag{19}$$

The structural route substitutes $\mathcal{M}_{struct}$ and $\mathcal{S}_{struct}$. The winning image contributes its top-scoring masks $\mathcal{M}_{cand}$ to the VLM stage.

B.2 Context-aware canary generation

A frozen VLM receives I_{anc} with numbered candidate boxes and jointly chooses the canvas m_{opt}, the visual trigger T_{trig}, and the textual signature S_{trap}. The same decision prompt handles two modes. **Mode A** mutates existing text by a localized 1-to-2 character change while preserving word length, casing, canvas extent and scene geometry, such as "CAFE" to "CXFE" or "Bothell" to "Dothell". **Mode B** is used when editable text is absent. It selects a flat structural canvas and inserts a context-compatible object that carries a rare 3-to-5 letter acronym or word, such as a green book titled "XCO" on an empty office desk. The prompt used for this decision is in Appendix I.2.

B.3 Dual-guided semantic injection

The edit must satisfy three objectives in a black-box RAG pipeline: retriever pull toward the forensic probe, generator push toward explicit signature decoding, and manifold alignment for object insertion. Let $z_0 = \mathcal{E}_{vae}(I_{anc})$ and $z_t = (1-t)z_0 + tz_1$ with $z_1 \sim \mathcal{N}(0, \mathbf{I})$. The clean latent estimate is $\hat{z}_0 = z_t - tv_\theta(z_t, t, c)$ and $\hat{x}_0 = \mathcal{D}_{vae}(\hat{z}_0)$. The weighted objective is

$$\mathcal{J} = \lambda_{ret}\mathcal{L}_{ret} + \lambda_{gen}\mathcal{L}_{gen} + \lambda_{align}\mathcal{L}_{align} + \lambda_{bnd}\mathcal{L}_{bnd}.$$

Retriever guidance aligns the image with the signature-free probe:

$$\mathcal{L}_{ret} = 1 - \frac{\Phi_{img}(\hat{x}_0) \cdot \Phi_{text}(Q_{probe})}{\|\Phi_{img}(\hat{x}_0)\|\|\Phi_{text}(Q_{probe})\|}.$$

Generator guidance teacher-forces the signature under the surrogate VLM:

$$\mathcal{L}_{gen} = -\sum_{i=1}^{|S_{trap}|} \log P_{\mathcal{G}_{surr}}(s_i \mid s_{<i}, \hat{x}_0, Q_{probe}).$$

For Mode B, manifold alignment uses DINOv2-g/14 patch features on the target boundary band. With $F_l \in \mathbb{R}^{C_l \times P_l}$ and $\mathbf{G}_l = F_l F_l^\top/(P_l C_l)$,

$$\mathcal{L}_{align} = \sum_{l \in \mathcal{L}} w_l \|\mathbf{G}_l(\hat{x}_0 \odot \partial m_{opt}) - \mathbf{G}_l(I_{anc} \odot \partial m_{opt})\|_F^2,$$

where one-indexed blocks $\{10, 20, 30, 40\}$ have weights $1/4$. The boundary term is $\mathcal{L}_{bnd} = \|\nabla^2(\hat{x}_0) \odot \partial M_{pixel}\|_2^2$ with $\partial M_{pixel} = \|\nabla M_{pixel}\|_2$. Guidance modifies the velocity as

$$\hat{v}_\theta(z_t, t, c) = v_\theta(z_t, t, c) + \eta \nabla_{z_t} \mathcal{J}(z_t). \tag{20}$$

Reverse Euler integration turns this positive velocity correction into a negative objective-gradient increment. Under the local stationarity approximation $\partial \hat{z}_0/\partial z_t \approx \mathbf{I}$, it acts as descent on the projected clean latent.

A Gaussian-feathered pixel mask $M_{pixel} = G_\sigma * m_{opt}^{pixel}$ with $\sigma = 3$ is bilinearly downsampled to $M_{latent} \in [0,1]^{H/f_{vae} \times W/f_{vae}}$. With $S = 50$, $\Delta t = 1/S$, $t_{start} = 0.80$, and a 40-step active window, each step blends the guided state with the host interpolation:

$$z_{t-\Delta t} = M_{latent} \odot z_{t-\Delta t}^{guided} + (\mathbf{1} - M_{latent}) \odot z_{t-\Delta t}^{orig}. \tag{21}$$

Here $z_{t-\Delta t}^{guided} = z_t - \Delta t \hat{v}_\theta(z_t, t, c)$ and $z_{t-\Delta t}^{orig} = (1-(t-\Delta t))z_0 + (t-\Delta t)z_1$. Soft latent blending supplies continuous boundary decay, while the boundary loss penalizes residual seams.

Output support. Latent blending and gradient gating alone do not imply pixelwise support preservation, because feathering and the decoder couple neighboring locations. The delivered RGB image therefore uses a final compact, inward taper. Let $d(x)$ be distance inside the selected pixel support m_{opt}, zero outside, $r = 9$ pixels, and

$$w(x) = \begin{cases} \frac{1}{2}[1 - \cos(\pi \min\{d(x)/r, 1\})], & x \in m_{opt}, \\ 0, & x \notin m_{opt}, \end{cases} \qquad I_{out} = I_{anc} + w \odot [\mathcal{D}_{vae}(z_{final}) - I_{anc}]. \tag{22}$$

Thus $I_{out}(x) = I_{anc}(x)$ outside the support; both w and its first derivative vanish at the outer boundary. This removes an imposed compositing step, not natural scene edges or all interior high-frequency energy. The guarantee is on the decoded RGB grid before any later lossy transform. All soft-blending editor rows use the same output rule. The trajectory diagnostic measures the decoded intermediate state before this final projection; its nonzero background LPIPS is not a measurement of delivered off-support pixel differences.

Implementation interface. The RGB editing grid is $1 \times 3 \times 768 \times 768$ in $[0, 1]$. The VAE downsampling factor is $f_{\text{vae}} = 8$ with $C = 32$ latent channels, so unpacked latents have spatial size 96×96. Blending is performed on unpacked latents with a 96×96 mask broadcast across channels. The schedule is the 50-step uniform reverse Euler schedule with the 40-step window $t = .8, \ldots, .02$, and backbone velocity is stop-gradient. Gradients pass through the VAE decoder and differentiable resizing to the 224-, 448- and 518-pixel CLIP, generator and DINO inputs.

B.4 SGGS AND GRADIENT STABILITY

The full gradient path through the flow projection, decoder and pixel objectives is

$$\nabla_{z_t} \mathcal{J} = (\mathbf{I} - t\mathbf{J}_v)^\top \mathbf{J}_{\mathcal{D}_{vae}}^\top \mathbf{g}_{pixel}, \qquad \mathbf{J}_v = \frac{\partial v_\theta}{\partial z_t}, \tag{23}$$

where $\mathbf{g}_{pixel} = \lambda_{ret}\partial\mathcal{L}_{ret}/\partial\hat{x}_0 + \lambda_{gen}\partial\mathcal{L}_{gen}/\partial\hat{x}_0 + \lambda_{align}\partial\mathcal{L}_{align}/\partial\hat{x}_0 + \lambda_{bnd}\partial\mathcal{L}_{bnd}/\partial\hat{x}_0$. Direct differentiation through $\mathbf{J}_v$ backpropagates through the velocity transformer at every step and makes the update sensitive to high-order variations. SemantiTrace treats the frozen velocity field as locally stationary, giving $\partial\hat{z}_0/\partial z_t \approx \mathbf{I}$, and then applies Spatially-Gated Gradient Scaling:

$$\widetilde{\nabla}_{z_t}\mathcal{J} = \mathbf{J}_{\mathcal{D}_{vae}}^{\top}\left[m_{opt} \odot \left(\gamma_t \frac{\mathbf{g}_{pixel}}{\|\mathbf{g}_{pixel}\|_2 + \epsilon}\right)\right], \qquad \gamma_t = 1 - t.$$

The spatial gate projects the pixel gradient onto the editable canvas before the decoder-vector product, and the temporal warm-up damps early noise-dominated steps. A sufficient frozen-step descent condition follows from Cauchy–Schwarz: if the true gradient h and stop-gradient proxy g satisfy $\|h - g\| \le e$ and $g^{\top}d > e\|d\|$, then $h^{\top}d > 0$, and an L-smooth objective decreases for $0 < a < 2(g^{\top}d - e\|d\|)/(L\|d\|^2)$.

Table 8 compares three MMQA optimization paths over all attempted anchors. Naive backprop differentiates the exact chain in Eq. 23. Decoupled Tweedie removes the velocity Jacobian but applies the resulting pixel gradient without SGGS. The full row adds SGGS. CSR is the share of trajectories without NaN or overflow. Max Grad RMS is $\|g\|_2/\sqrt{d}$ over latent dimension $d = C \times H/f_{vae} \times W/f_{vae}$, and gradient variance is the spatial component-wise variance, with $\|g\|_{RMS}^2 = \sigma_{\nabla}^2 + \bar{g}^2$. Non-convergent runs are counted as injection failures.

Table 8: Gradient stability and convergence on MMQA. CSR is convergence without NaN values. CGSR is evaluated across attempted anchors.

Optimization configuration	Max Grad RMS	Grad variance σ_{∇}^2	CSR	CGSR (GPT-5.2)
Naive backprop (exact Eq. 23)	142.50	8.4×10^3	12.4%	1.80%
Decoupled Tweedie (w/o SGGS)	18.20	142.60	74.8%	62.15%
Decoupled + SGGS	**1.24**	**0.08**	**99.6%**	**88.50%**

B.5 Concept-space optimization

Concept-Space Generalized Optimization reduces dependence on one probe formulation. An auxiliary LLM produces $P = 8$ reformulations that vary syntax and granularity within the anchor's semantic cluster. Their normalized CLIP text embeddings define

$$\mathbf{c}_{concept} = \text{Normalize}\left(\frac{1}{P}\sum_{p=1}^{P}\frac{\Phi_{text}(Q_p)}{\|\Phi_{text}(Q_p)\|_2}\right).$$

The generalized retriever loss is $\mathcal{L}_{ret}^{g} = 1 - \langle\Phi_{img}(\hat{x}_0)/\|\Phi_{img}(\hat{x}_0)\|_2, \mathbf{c}_{concept}\rangle$. The generator loss averages the teacher-forced signature log-likelihood over the same P queries. If unit vectors v, c, q obey $\angle(q, c) \le \delta$ and $\theta = \arccos(v^{\top}c)$, then $v^{\top}q \ge \cos(\min\{\pi, \theta + \delta\})$ and $v^{\top}q \ge v^{\top}c - 2\sin(\delta/2)$.

The diagnostic separates the retrieval probe Q_{ret} from the generator query Q_{gen}. Seen queries use the target-specific formulation from optimization. Context-implicit queries use macroscopic cluster descriptors for retrieval and ask about scene markings without container keywords. The global-scene diagnostic uses broad cluster-level retrieval and the generator query ``Please transcribe all readable text appearing anywhere in this image.'' Figure 7 reports the three-panel diagnostic with bar labels. Under the global-scene diagnostic, CS-GO reaches retrieval mean rank 1.52 and GPT-5.2 CGSR 79.60%, while single-query optimization falls to 43.35% GPT-5.2 CGSR. The MMQA audit survival ceiling is $491/500 = 98.20\%$, and clean query collisions are 0.05%.

The audit setting uses broad scene descriptors together with local canvas inspection, matching the templates in Appendix I.1.

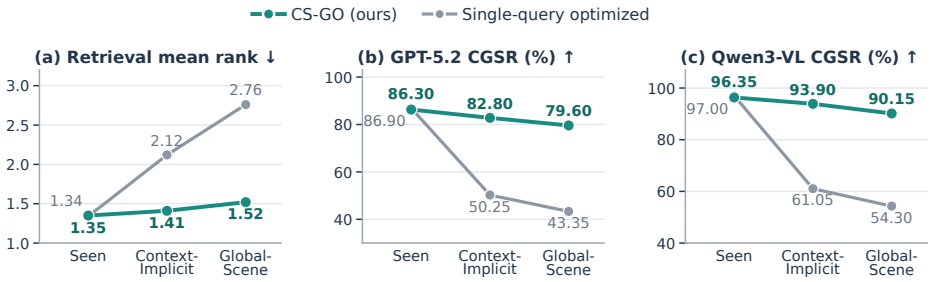


Figure 7: Query generalization on MMQA after sanitization. Panels report retrieval mean rank and CGSR for GPT-5.2 and Qwen3-VL under seen, context-implicit and global-scene diagnostics. The audit survival ceiling is $491/500 = 98.20\%$; clean query collisions are 0.05%.

B.6 Local proposal score, repair and initialization

Local proposal score. The background lexicon is computed once from a separate 20,000-image clean inspector calibration split using the test-time OCR configuration, NFKC, casefolding and complete-token matching. A token contributes document frequency $f(w)$ at most once per image from a box with confidence at least .90. With $f_{\max} = \max(1, \max_w f(w))$,

$$r(w) = 1 - \frac{\log(1 + f(w))}{\log(1 + f_{\max})}, \qquad f(w) = 0 \text{ for unseen tokens.}$$

Boxes whose IoU is at least .5 are merged. For merged box b, let r_b be maximum token rarity, c_b the maximum confidence, and a_b the area fraction. The priority is $.50r_b + .30c_b + .20\sqrt{a_b}$. Width and height are enlarged by 1.20 about the center, clipped to the image, and the first B proposals are inspected. The inspected local anomaly score is combined with global rejection by OR.

Lexical repair. Replacement candidates for token w have $f(v) \geq 5$, the same number of Unicode alphanumeric characters, and Levenshtein distance one or two. The defender scores context by masking the box with the per-channel mean and teacher-forcing each proposed completion. It chooses the largest

$$J_{lex}(v) = -\frac{d_{lev}(v, w)}{\max(1, |w|)} + .25\frac{\log(1 + f(v))}{\log(1 + f_{\max})} + .50\ell(v),$$

Here $\ell(v) = |\operatorname{tok}(v)|^{-1} \sum_{j=1}^{|\operatorname{tok}(v)|} \log P(v_j \mid v_{<j}, I_{masked}, Q_{context})$ is the mean natural-log probability per completion token; prompt and EOS tokens are excluded. Ties are broken by Unicode code-point order of the normalized candidate. Agreement uses OCR string o, image-only VLM transcription v, token confidence c_o, and emitted-token probability c_v:

$$A(I, b) = \sqrt{c_o c_v}\left(1 - \frac{d_{lev}(o, v)}{\max(1, |o|, |v|)}\right).$$

The default repair proposes inpainting for $a_b^{anom} \geq .82$ and lexical correction for $r_b \geq .90$ with $c_b < .80$. Proposals are processed in priority order. Lexical repair has precedence when both branches apply and an alternative exists. Accepted corrections require refreshed agreement at least $\tau_A = .80$. OCR, embeddings and the index are then recomputed.

Initialization. Let I_h be the host, I_e the native initializer output, $z_h = \mathcal{E}(I_h)$ and $z_e = \mathcal{E}(I_e)$. For SemantiTrace, $I_e = I_h$; for AT-full it is AnyText2's rendering, preceded by common object insertion in Mode B. At $t = .8$,

$$z_{.8} = .2\{M_{latent} \odot z_e + (1 - M_{latent}) \odot z_h\} + .8\epsilon.$$

Each background reset uses pristine z_h: $z_{t'}^{bg} = (1 - t')z_h + t'\epsilon$. Both editors then run the same 40-step refinement.

C Complete Main Results and Baseline Controls

C.1 Complete generator comparison

Tables 9 and 9 give the complete pre-sanitization ranks, query-level CGSR and paired cluster-level tests behind Figure 3. Each row uses 500 hosts and 2,000 probes per dataset.

Table 9: Complete pre-sanitization generator comparison. CGSR is query success.

		MMQA			WebQA		
Method	Generator	Rank	CGSR (%)	p	Rank	CGSR (%)	p
Naive	LLaVA-NeXT	2.62	28.15	0.26	2.92	13.05	0.76
	InternVL3	2.62	27.10	0.34	2.92	17.40	0.58
	Qwen3-VL	2.62	37.05	0.21	2.92	15.55	0.92
	GPT-5.2	2.62	25.30	0.37	2.92	10.55	0.95
	Gemini 3.1 Pro Preview	2.62	24.80	0.44	2.92	9.85	0.97
PGD	LLaVA-NeXT	1.48	31.20	2.6×10^{-4}	1.90	22.85	3.9×10^{-3}
	InternVL3	1.48	29.00	2.2×10^{-3}	1.90	19.45	3.8×10^{-3}
	Qwen3-VL	1.48	39.25	5.4×10^{-3}	1.90	18.70	6.9×10^{-3}
	GPT-5.2	1.48	28.55	0.01	1.90	15.65	8.2×10^{-3}
	Gemini 3.1 Pro Preview	1.48	27.60	0.01	1.90	11.15	0.01
$\mathrm{AQUA}_{acronym}$	LLaVA-NeXT	1.04	85.30	$< 1 \times 10^{-10}$	1.08	78.25	$< 1 \times 10^{-10}$
	InternVL3	1.04	85.15	$< 1 \times 10^{-10}$	1.08	78.55	$< 1 \times 10^{-10}$
	Qwen3-VL	1.04	98.70	$< 1 \times 10^{-10}$	1.08	95.20	$< 1 \times 10^{-10}$
	GPT-5.2	1.04	81.25	$< 1 \times 10^{-10}$	1.08	76.80	$< 1 \times 10^{-10}$
	Gemini 3.1 Pro Preview	1.04	86.40	$< 1 \times 10^{-10}$	1.08	80.15	$< 1 \times 10^{-10}$
$\mathrm{AQUA}_{spatial}$	LLaVA-NeXT	1.29	75.10	$< 1 \times 10^{-10}$	1.66	86.15	$< 1 \times 10^{-10}$
	InternVL3	1.29	75.20	$< 1 \times 10^{-10}$	1.66	72.25	$< 1 \times 10^{-10}$
	Qwen3-VL	1.29	98.05	$< 1 \times 10^{-10}$	1.66	88.15	$< 1 \times 10^{-10}$
	GPT-5.2	1.29	72.10	$< 1 \times 10^{-10}$	1.66	69.45	$< 1 \times 10^{-10}$
	Gemini 3.1 Pro Preview	1.29	78.30	$< 1 \times 10^{-10}$	1.66	74.60	$< 1 \times 10^{-10}$
SemantiTrace	LLaVA-NeXT	1.34	86.10	$< 1 \times 10^{-10}$	1.37	84.50	$< 1 \times 10^{-10}$
	InternVL3	1.34	87.20	$< 1 \times 10^{-10}$	1.37	82.15	$< 1 \times 10^{-10}$
	Qwen3-VL	1.34	98.80	$< 1 \times 10^{-10}$	1.37	95.40	$< 1 \times 10^{-10}$
	GPT-5.2	1.34	88.50	$< 1 \times 10^{-10}$	1.37	83.20	$< 1 \times 10^{-10}$
	Gemini 3.1 Pro Preview	1.34	91.50	$< 1 \times 10^{-10}$	1.37	87.90	$< 1 \times 10^{-10}$

C.2 Stealthiness and post-sanitization survivability

The dense fidelity evaluation uses $M = 10{,}000$ edited images with a 100% injection rate, separating intrinsic edit artifacts from sparse deployment. FID compares each arm with a separate, fixed 10,000-image clean reference pool (5,000 images per dataset), disjoint from DENSE-10K. It uses RGB decoding, bilinear resize to 299×299 without cropping, the TensorFlow-FID Inception-v3 pool3 2,048-dimensional features, and the usual Gaussian mean/covariance Fréchet distance. Clean and edited arms share the reference statistics. The ten paired delete-group estimates retain the full reference pool and remove the same 1,000 evaluation hosts from both arms. SemantiTrace has lower LPIPS than PGD (0.044 vs. 0.084). Its FID_{10K} is 13.85, close to the clean reference value 12.45. The same filter rejects 94.2% of structured acronym AQUA, 82.4% of structured spatial AQUA and 1.8% of SemantiTrace. Figure 8 illustrates the local-edit and structured-AQUA regimes.

For WebQA/GPT-5.2, structured acronym AQUA changes from 76.80% before filtering to 2.15% after filtering, with 18 Union-positive anchors and zero All-K anchors. SemantiTrace changes from 83.20% to 81.70%, with 469/500 Union-positive anchors and $372/500$ All-K anchors. Figure 9a shows the post-filter comparison and CER frequencies.

Table 10: Dense-pool fidelity, rejection and paired delete-group FID summaries. Blocks delete 1,000 paired hosts from DENSE-10K.

Method	PSNR	LPIPS	FID_{10K}	Reject. (%)	Del.	Clean FID_{9K}	ST FID_{9K}	Δ_{-b}
Clean baseline	∞	0.00	12.45	0.5	1	12.61	13.99	1.38
PGD	**31.20**	0.084	**13.02**	2.6	2	12.57	14.00	1.43
$\mathrm{AQUA}_{acronym}$	n/a	n/a	48.76	94.2	3	12.60	13.99	1.39
$\mathrm{AQUA}_{spatial}$	n/a	n/a	35.18	82.4	4	12.54	13.96	1.42
SemantiTrace	29.35	**0.044**	13.85	**1.8**	5	12.58	13.95	1.37
					6	12.55	13.96	1.41
					7	12.62	14.02	1.40
					8	12.56	13.92	1.36
					9	12.59	14.03	1.44
					10	12.63	14.03	1.40

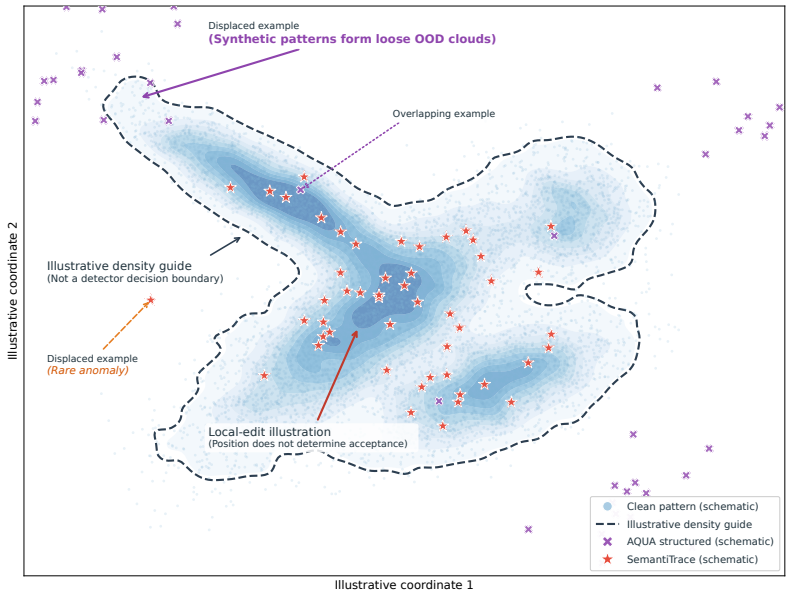
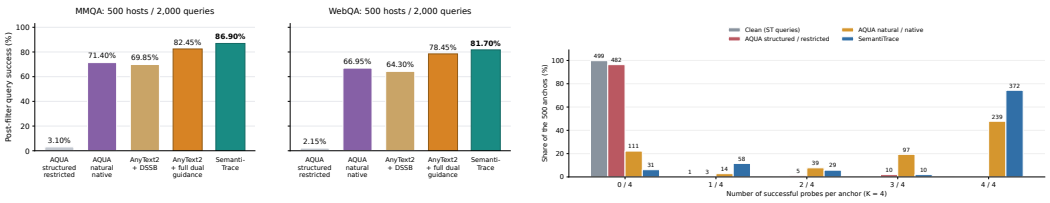


Figure 8: Schematic two-dimensional layout of benign, structured-AQUA and local-edit patterns. Filter decisions use native-space Mahalanobis scores and VLM flags; the projection is illustrative.



(a) Post-filter query success.

(b) Four-probe CER frequencies.

Figure 9: Post-sanitization behavior on the fixed 500-host cohorts. The CER histograms are exact empirical frequencies: SemantiTrace $(31, 58, 29, 10, 372)$, structured AQUA $(482, 3, 5, 10, 0)$ and natural/native AQUA $(111, 14, 39, 97, 239)$.

C.3 MATCHED LOCAL EDITORS AND AQUA CONTROLS

Same-host editor comparisons fix the source image, canvas, signature, four probes, gallery, filter, generator, normalization and denominator. The Mode-A split has 290 MMQA and 133 WebQA anchors, hence 1,692 queries. Each editor receives 12 development trials on the same 50-image split; selection maximizes surrogate end-to-end CGSR subject to development rejection at most 2% and LPIPS at most 0.06. The shared refinement uses $\sigma = 3$ feathering and a 40-step editing window. Table 11 reports the primary run; Table 12 gives five-run variation on the same hosts.

Table 11: Matched Mode-A controls after the full global-sanitization, retrieval and generation pipeline on GPT-5.2.

Editor	OCR exact (%)	Rejected	CGSR (%)	All-K	LPIPS	s/image
Prompt-only + DSSB	72.81	9/423	51.48	144/423 (34.04%)	0.027	38.7
Glyph replacement + DSSB	98.82	18/423	62.94	204/423 (48.23%)	0.032	35.6
AnyText2 + DSSB refinement	98.11	13/423	76.89	265/423 (62.65%)	0.028	96.8
Dual guidance + global clipping	95.98	11/423	86.05	330/423 (78.01%)	0.031	120.5
SemantiTrace	95.74	4/423	88.77	346/423 (81.80%)	0.023	121.2

The SemantiTrace versus AnyText2 All-K discordances are $(131, 50)$, with $p = 7.34 \times 10^{-10}$ and Holm-adjusted $p = 1.47 \times 10^{-9}$.

Table 1 in Section 4.3 gives the complete-pipeline comparison after global filtering. The prompt-only + DSSB control reaches $40.05\%/34.60\%$ CGSR on MMQA/WebQA, with 110/80 All-K anchors and 16/18 rejections. Table 12 summarizes the five fixed-cohort realizations.

Table 12: Five construction and serving realizations on fixed cohorts. Entries are mean and sample standard deviation in pp.

	MMQA			WebQA		
Method	CGSR	All-K	Reject	CGSR	All-K	Reject
AQUA structured scene	3.10 ± 0.27	0.24 ± 0.26	94.20 ± 0.20	2.29 ± 0.24	0.08 ± 0.18	94.20 ± 0.14
AQUA natural scene	64.24 ± 0.17	49.56 ± 0.38	18.00 ± 0.20	58.31 ± 0.23	42.08 ± 0.23	22.00 ± 0.14
Prompt-only + DSSB	40.26 ± 0.20	22.16 ± 0.09	3.20 ± 0.20	34.64 ± 0.30	16.12 ± 0.50	3.60 ± 0.14
AnyText2 + DSSB refinement	69.88 ± 0.26	52.32 ± 0.30	4.20 ± 0.20	64.22 ± 0.17	45.68 ± 0.30	5.60 ± 0.14
SemantiTrace	86.87 ± 0.27	80.00 ± 0.37	1.80 ± 0.20	81.68 ± 0.28	74.24 ± 0.26	2.20 ± 0.14

AQUA construction fields and native cues. AQUA uses distinct fields for rendered initials, complete verification phrase and native query. The matcher tests the complete phrase as a contiguous token sequence. The natural-prose and field-serialization bridge arms use the same scene landmarks, initials and expansion. Table 13 combines the native cue intervention with the construction serialization bridge.

Table 13: Native-query and AQUA construction correspondence controls.

Block	Condition	Cue (b, c) / rejected	p / histogram	Hits / clean hits	CGSR (%) / clean All-K	Blocks
Native cue	MMQA	$(38, 13)$	3.11×10^{-4}	6/2000	0/500	122/125
Native cue	WebQA	$(38, 10)$	3.08×10^{-5}	7/2000	1/500	117/125
AQUA bridge	Natural prose, combined	110/500	$(111, 14, 39, 97, 239)$	1339/2000	66.95	n/a
AQUA bridge	Field serialization, combined	107/500	$(111, 14, 30, 98, 247)$	1356/2000	67.80	n/a
AQUA bridge	Natural prose, VLM only	84/500	$(86, 15, 37, 106, 256)$	1431/2000	71.55	n/a
AQUA bridge	Field serialization, VLM only	80/500	$(83, 15, 34, 106, 262)$	1449/2000	72.45	n/a
AQUA bridge	SemantiTrace, VLM only	7/500	$(26, 55, 30, 12, 377)$	1659/2000	82.95	n/a

Fixed initials. Fixed initials are compatible with private assignment because all eight candidate expansions share the same initialism while remaining distinct complete token sequences. The native query is fixed before the private draw and contains the fixed initialism; the selected expansion remains private. Table 14 gives the joint repaired PO setting.

Table 14: Assignment-independent initials with eight private complete expansions under the joint repaired PO setting.

Editor	$(n_0, \ldots, n_4)$	Query hits	Min. queries
ST	$(132, 74, 40, 34, 220)$	1136/2000	64
AT-full	$(155, 51, 53, 71, 170)$	1050/2000	92
AQUA-FI	$(176, 72, 62, 40, 150)$	916/2000	116

All rows use 500 attempted source slots and 2,000 probes. All three rules use the complete eight-candidate hit set, $p < .01$, and no clean service.

D Transfer, Robustness and Utility

D.1 Retrieval interfaces

SemantiTrace transfers beyond the CLIP retriever used during optimization. The tested vector interfaces are SigLIP-SO400M with its native text tower, DINOv2-L/14 with the `dino.txt` text tower, and ConvNeXt-XXLarge through an OpenCLIP text tower. All use normalized embeddings and top-k MIPS. ColPali uses the `vidore/colpali-v1.3` checkpoint, PaliGemma-3B-pt-448, a 128-dimensional projection head and 1024 patch embeddings with MaxSim scoring (Table 15). A caption-indexed service captions images with LLaVA-NeXT-8B, indexes the text with BGE-M3, and retrieves text for a text-only generator. Pre-sanitization Recall@5 stays at 89.45–95.40% on the unseen vector interfaces and reaches 91.80% on MMQA and 88.50% on WebQA in the caption index. PGD transfers poorly: its unseen-vector Recall@5 is 2.05–14.50%, and caption indexing gives 0.00% signature capture (Figure 10, top).

The OCR-first service uses PaddleOCR strings as the sole indexed field, BGE-M3 as the text retriever, top-$k = 5$, and GPT-5.2 or Qwen3-8B as the answer generator. Source titles, captions and spatial metadata are excluded from the raw-OCR index. Mode A benefits from editing prominent existing text: on MMQA it has 95.86% OCR transcription and 93.45% Text-RAG CGSR. Mode B relies on original OCR cues in the query portfolio. Full SemantiTrace keeps 81.30% and 69.30% Text-RAG CGSR on MMQA and WebQA, while AQUA acronym obtains 3.80% and 2.60% (Figure 10, bottom).

Table 15: Post-sanitization ColPali late-interaction results. Recall@5 includes survival and retrieval; CGSR uses all $NK = 2000$ probes.

Dataset	Method	CLIP R@5	CLIP rank	ColPali R@5	ColPali rank	CGSR
MMQA	PGD	88.80%	1.46	7.95%	24.80	6.20%
MMQA	AQUA acronym	4.50%	2.38	3.10%	48.50	2.10%
MMQA	SemantiTrace	98.00%	1.33	96.65%	1.28	89.10%
WebQA	PGD	86.10%	1.87	6.25%	30.10	4.80%
WebQA	AQUA spatial	11.20%	3.28	8.50%	22.10	5.90%
WebQA	SemantiTrace	97.35%	1.35	96.15%	1.31	85.60%

PaddleOCR text with confidence at least .50 is the only indexed field, embedded with BGE-M3. The top five documents are passed to the text-only generator. Table 16 separates Mode-B hosts with original, unedited OCR cues from cue-free hosts; Table 16 shows the matched schema control that adds original-scene context.

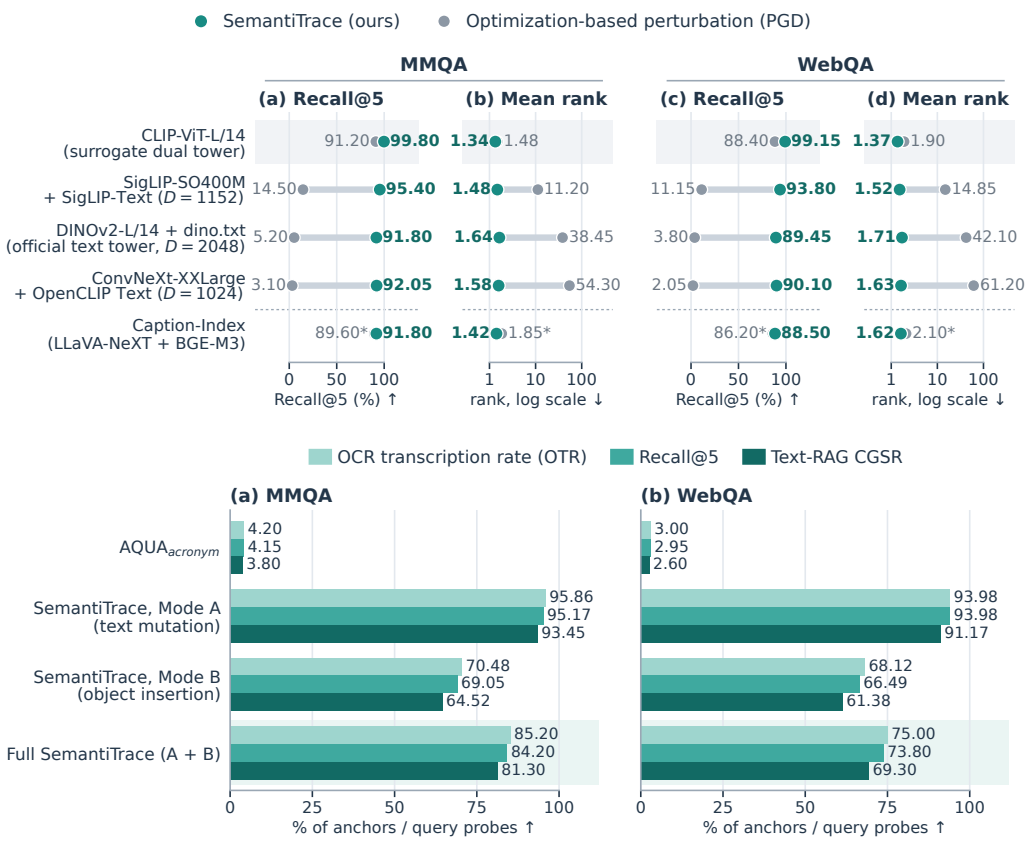


Figure 10: Transfer beyond the optimization-time CLIP index. Top: watermarks optimized under the surrogate CLIP retriever (shaded row) and evaluated before sanitization on unseen vector and caption-index pipelines; for PGD the benign caption is retrieved but the perturbation is not transcribed, giving 0.0% signature capture in text (*). Bottom: SemantiTrace and AQUA$_{acronym}$ under an OCR-first pipeline after sanitization with the text-only GPT-5.2 generator; rates count anchors ($N = 500$) or probes ($NK = 2000$).

Table 16: Raw-OCR cue strata and context control. OTR is host-level; Recall@5 and matches are query-level.

Dataset/subset	Index or cue stratum	Hosts	OTR	Recall@5	Query matches
MMQA	Mode B cue available	186	140/186 (75.27%)	580/744 (77.96%)	542/744 (72.85%)
MMQA	Mode B cue unavailable	24	8/24 (33.33%)	0/96 (0.00%)	0/96 (0.00%)
MMQA	All Mode B	210	148/210 (70.48%)	580/840 (69.05%)	542/840 (64.52%)
WebQA	Mode B cue available	317	240/317 (75.71%)	976/1268 (76.97%)	901/1268 (71.06%)
WebQA	Mode B cue unavailable	50	10/50 (20.00%)	0/200 (0.00%)	0/200 (0.00%)
WebQA	All Mode B	367	250/367 (68.12%)	976/1468 (66.49%)	901/1468 (61.38%)
MMQA	24 cue-free, raw OCR only	24	8/24	0/96	0/96
MMQA	24 cue-free, raw OCR + context	24	8/24	82/96	29/96
WebQA	50 cue-free, raw OCR only	50	10/50	0/200	0/200
WebQA	50 cue-free, raw OCR + context	50	10/50	164/200	35/200

D.2 IMAGE-TO-ANSWER EXAMPLE THROUGH THE RAW-OCR INDEX

Figure 11 replays the baseball example (Case 3 of Figure 17) through the raw-OCR interface. The structural canvas is the ground beside the player's feet; the jersey number `26` and the emblem `W` are unedited and give an observable route into the OCR index. The private catalogue is `BSB, BSD, BSR, BSV, BSX, BSP, BSN, BST`, and the selected candidate is `BSB`. The four queries refer to the jersey number, emblem and ball label, and contain no catalogue member.



Original | Edited | Original lower view | Edited lower view

Figure 11: Source and edited image of the replayed example, with their lower views. The lower view removes the jersey cues and keeps the inserted ball. The unedited number and emblem remain visible in the full views.

The index holds the target document d_0 and five common distractors d_1–d_5. Four interface conditions are applied to this one source. S indexes the full edited view with `BSB`; F indexes only the lower view; W produces an OCR error that reads `BSR`; and M adds a competing document d_6 containing `BSR`. Table 17 gives the returned documents and the complete-candidate scoring.

Table 17: Returned document IDs and complete-candidate scoring in the image-to-answer example. Each retrieval order applies to all four probes of its condition.

Condition	Returned top five	OCR correct	Candidates in each response	Hit set	Selected hits
S / full view	d_0, d_2, d_1, d_3, d_4	1	BSB (all four)	{BSB}	4/4
F / lower view	d_2, d_1, d_3, d_4, d_5	1	none	empty set	0/4
W / OCR error	d_0, d_2, d_1, d_3, d_4	0	BSR (all four)	{BSR}	0/4
M / two documents	d_6, d_0, d_2, d_1, d_3	1	BSB and BSR (all four)	{BSB, BSR}	4/4

The complete answers follow the retrieved text. S returns the selected label in all four answers, F returns no candidate, W returns the OCR error `BSR`, and M returns both `BSB` and `BSR`. Full-token matching credits every complete candidate in an answer, so S has a selected singleton, W an unselected singleton, M a selected doubleton and F an empty hit set; the reference hit probabilities for S, W and M are $1/8$, $1/8$ and $2/8$.

D.3 INGESTION TRANSFORMS AND PURIFICATION

Standard ingestion transforms preserve the visible semantic carrier better than high-frequency perturbations (Figure 12). JPEG $Q = 75$ leaves SemantiTrace in the 68.30–85.20% CGSR range across the reported datasets and generators. Rescaling by $0.5\times$, clipped 8-bit Gaussian noise with $\sigma = 5$, and 10% center cropping remain in the same operating range. Active purification also leaves usable evidence: bilateral filtering uses $d = 9$, $\sigma_{color} = 0.1$, $\sigma_{space} = 75$; VQ-VAE-2 applies bottleneck reconstruction; DiffPure with $\sigma_{pure} = 0.15$ and 15 reverse-SDE denoising steps yields 78.40% CGSR on MMQA and 71.95% on WebQA for GPT-5.2. PGD falls to 0.10% and 0.00% under the same DiffPure rows.

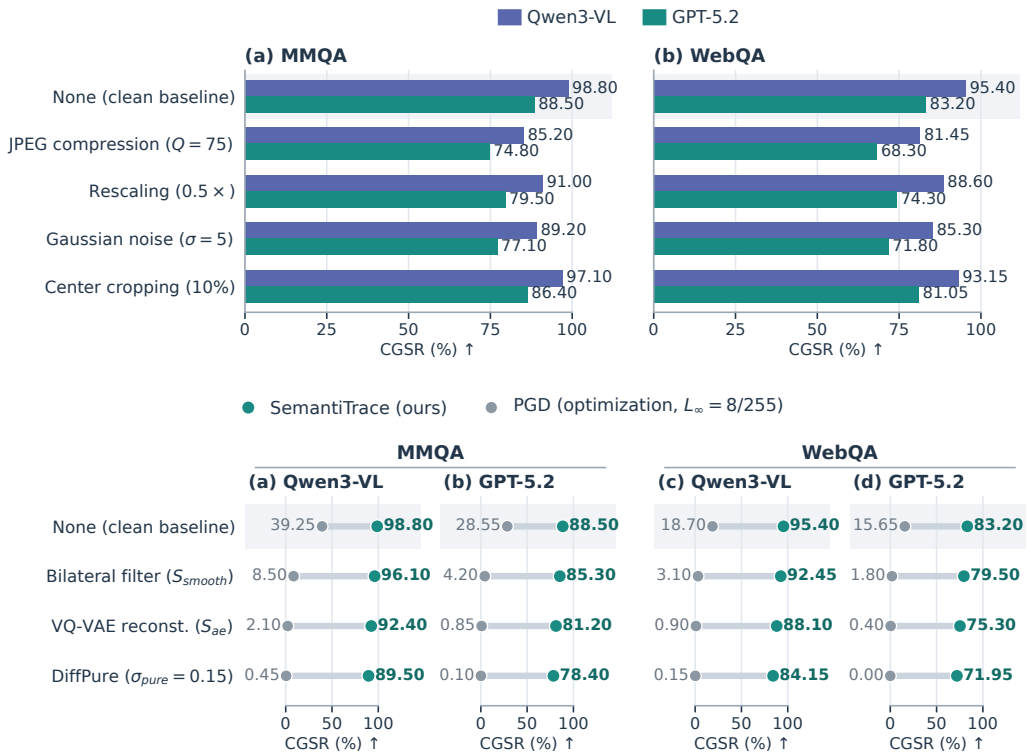


Figure 12: Robustness to ingestion transforms (top) and purification (bottom) on MMQA and WebQA. Bars and points report CGSR on Qwen3-VL and GPT-5.2; the shaded row is the clean baseline. In the bottom panels, each row pairs PGD and SemantiTrace under the same defense and generator.

D.4 AUTOMATIC LOCAL INSPECTION AND REPAIR DEFENSES

A method-aware defender can inspect OCR proposals without privileged masks, strings, original counterparts, audit queries or hit labels. PaddleOCR proposals are merged at IoU 0.5, enlarged by 20%, ranked by OCR confidence, normalized area and background-lexicon rarity, and inspected by Qwen3-VL-32B. A disjoint 20,000-image split sets thresholds for a 1% clean-rejection target at the ordinary operating points. The four-crop inspector keeps 68.80% WebQA CGSR and 310/500 All-K for SemantiTrace, rejects 0.98% of clean images, takes 4.12 s/image, and leaves the matched AnyText2 control at 51.30% CGSR (Table 18). Clean-corpus RAG F1 for global, OCR-4, OCR-8, aggressive OCR-8 and OCR-32 inspection is 65.08, 64.94, 64.91, 63.84 and 60.72.

Table 18: Automatic local inspection frontier on WebQA/GPT-5.2. Rejection, crop caps and latency are image-level; CGSR includes all 2000 audit queries.

					SemantiTrace			AnyText2
Inspector	Clean FRR	Rare FRR	s/img	Text-QA	Rejected	CGSR	All-K	CGSR
Global	0.50%	0.80%	1.84	88.60%	11/500	81.70%	372/500	64.30%
OCR-1	0.94%	2.40%	2.61	88.10%	32/500	77.60%	352/500	60.00%
OCR-2	0.96%	2.55%	3.14	87.80%	54/500	73.45%	332/500	55.85%
OCR-4	0.98%	2.90%	4.12	87.20%	79/500	68.80%	310/500	51.30%
OCR-8	0.99%	3.15%	5.78	86.80%	95/500	65.95%	297/500	48.55%
OCR-8 aggressive	4.92%	14.05%	5.78	81.10%	232/500	41.10%	182/500	27.30%
OCR-32 unconstrained	17.80%	42.05%	11.34	69.20%	356/500	18.75%	78/500	13.00%

Non-rejection defenses use the same normal text-QA budget of at most 1.5 points from the 886/1000 global-filter workload (Table 19). The diagnostic keeps retrieval unchanged, then applies OCR-guided inpainting, lexical normalization, an output transcription guard, or their combination. The combined budgeted setting gives $1116/2000 = 55.80\%$ query success and 137/500 All-K at $871/1000 = 87.10\%$ normal text-QA accuracy. At the input level, context and effective character resolution both affect inspection (Table 20): on the same 290 MMQA Mode-A hosts, flagging rises from $2/290 = 0.69\%$ in the low-resolution full scene to $222/290 = 76.55\%$ in the high-resolution target crop. The fixed-height crop gains are 35.52 and 55.52 points; the fixed-context resolution gains are 20.34 and 40.34 points, with a 20.00-point interaction.

Table 19: Budgeted non-rejection defenses. all 500 audit hosts remain in the denominator.

Defense	Text-QA	Rare text	Loss	ST CGSR	ST All-K	s/img
Global filter only	88.60%	88.00%	0.00	81.70%	372/500	1.84
Automatic OCR-4 rejection	87.20%	n/a	1.40	68.80%	310/500	4.12
OCR-guided inpainting	87.30%	81.00%	1.30	65.00%	205/500	4.68
OCR lexical normalization	87.40%	79.00%	1.20	62.20%	162/500	2.95
Output transcription guard	87.20%	82.00%	1.40	66.00%	221/500	2.10
Combined budgeted defense	87.10%	78.50%	1.50	55.80%	137/500	5.73
Unconstrained normalization	74.60%	46.00%	14.00	16.75%	6/500	3.04

Table 20: Context-by-resolution control for anomaly inspection on the same 290 edited hosts.

Visual context	Character height	VLM flags
Full scene	12 pixels	$2/290 = 0.69\%$
Full scene	48 pixels	$61/290 = 21.03\%$
Target crop	12 pixels	$105/290 = 36.21\%$
Target crop	48 pixels	$222/290 = 76.55\%$

D.5 DATASET UTILITY AND PERMITTED FIELDS

Sparse deployment preserves broad QA utility in the evaluated workloads (Table 21). SemantiTrace inserts 500 canaries, about 0.87% of MMQA and 0.125% of WebQA, and leaves rounded macro F1 at 68.42% and 65.15%, matching the clean rows. Local benign utility also stays close to clean:

83.80% versus 84.20% on MMQA and 81.00% versus 81.60% on WebQA, while forensic extraction is 83.20% on WebQA/GPT-5.2.

Table 21: Utility preservation on GPT-5.2. Global entries are macro F1; local entries are accuracy on benign questions targeting anchor images.

Corpus status	MMQA F1	WebQA F1	MMQA local	WebQA local	CGSR
Clean	68.42	65.15	84.20%	81.60%	0.05%
AQUA spatial	68.37	65.10	45.40%	42.20%	69.45%
SemantiTrace	68.42	65.15	83.80%	81.00%	83.20%

Questions that read edited text need separate targets (Table 22). Macro scene questions, unedited text elsewhere, Mode-A entity-identity questions, and Mode-B neighboring original text remain in the workload. Literal Mode-A current-pixel readability is high, with 381/423 edited strings read correctly.

Table 22: Text-sensitive utility strata. Original-reference and current-pixel correctness are distinct targets.

Question stratum	N	Clean/original	Edited/original	Edited/current pixels
Macro scene questions	1000	829/1000 (82.90%)	824/1000 (82.40%)	824/1000 (82.40%)
Unedited text elsewhere	1000	891/1000 (89.10%)	887/1000 (88.70%)	887/1000 (88.70%)
Edited Mode-A string, literal transcription	423	394/423 (93.14%)	n/a	381/423 (90.07%)
Edited Mode-A entity identity, context question	423	371/423 (87.71%)	294/423 (69.50%)	n/a
Mode-B neighboring original text	577	502/577 (87.00%)	497/577 (86.14%)	497/577 (86.14%)

The deployment permission manifest restricts candidate canvases before assignment (Table 23). Immutable facts, legal, medical and geographic identifiers, and archival ground-truth fields are excluded; mutable owner-designated labels and permitted structural canvases remain eligible. The unchanged editor covers 963/1000 attempted hosts: 476/500 on MMQA and 487/500 on WebQA. All attempted hosts remain in the query and All-K denominators. Across the 237 originally textual fields protected as immutable, original-reference answers are 220/237 under the edited system versus 221/237 clean.

Table 23: Permission-constrained deployment with all 500 initial hosts retained in the denominators.

Dataset	Eligible coverage	Modes A/B	No canvas	Rejected	Query hits	All-K	$(n_0, \ldots, n_4)$
MMQA	476/500 (95.20%)	122/354	24	12	1648/2000 (82.40%)	372/500	(51, 23, 25, 29, 372)
WebQA	487/500 (97.40%)	64/423	13	14	1562/2000 (78.10%)	345/500	(63, 37, 20, 35, 345)

Of the original 423 Mode-A hosts, 186 retain a permitted mutable label, 200 have an admissible Mode-B fallback, and 37 have no admissible canvas. The original 577 Mode-B hosts stay in that route, giving 777 Mode-B hosts.

E Ablations and Complete-Objective Controls

This appendix collects component diagnostics for retrieval, decoding, filtering and complete-objective controls.

E.1 Core components, parasitic modes and surrogates

Figure 6 and Table 24 summarize the core ablations. Retriever guidance is the retrieval bottleneck: removing it gives zero top-5 placements and zero output matches on both datasets (Table 25). Generator guidance is the decoding bottleneck: retrieval rank remains near one, but MMQA/GPT-5.2 CGSR falls to 12.10%. Blending is the survival bottleneck: decoding remains high before filtering, while pooled OOD rejection rises to $863/1000 = 86.3\%$ (431 MMQA and 432 WebQA hosts). Mode B covers 73.4% of WebQA hosts, and alignment lowers its rejection from 5.20% to 2.77%. Cross-family surrogate substitutions retain 80.40–86.80% commercial-model CGSR. Figure 13 shows the paired trajectory diagnostic.

Table 24: Core component ablations, parasitic modes with alignment effects, and surrogate decoupling across MMQA and WebQA.

Configuration	Pooled OOD rej.	MMQA				WebQA			
		Rank	LPIPS	Pre-sanit. CGSR		Rank	LPIPS	Pre-sanit. CGSR	
				Qwen3	GPT-5.2			Qwen3	GPT-5.2
Core components									
w/o Ret. Guide ($\mathcal{L}_{ret}$)	1.5%	38.40	0.024	0.0%	0.0%	42.50	0.028	0.0%	0.0%
w/o Gen. Guide ($\mathcal{L}_{gen}$)	1.2%	**1.32**	**0.022**	18.50%	12.10%	**1.36**	**0.025**	15.20%	9.45%
w/o Latent Blending	86.3%	1.32	0.265	**99.10%**	82.30%	1.35	0.284	**96.10%**	77.60%
Parasitic modes and alignment; mode split MMQA 58.0%/42.0%, WebQA 26.6%/73.4%									
Mode A: Text Mutation	**0.9%**	**1.32**	**0.023**	**99.48%**	**90.17%**	**1.34**	**0.024**	**96.99%**	**87.97%**
Mode B: Object Insertion	2.77%	1.37	0.051	97.86%	86.19%	1.38	0.063	94.82%	81.47%
w/o Manifold Align. ($\mathcal{L}_{align}$)	5.20%	1.47	0.114	94.05%	77.38%	1.44	0.118	91.21%	72.55%
Surrogate model decoupling									
Surrogate: LLaVA-NeXT-8B (CLIP-ViT)	1.9%	1.35	0.038	92.30%	85.30%	1.38	0.055	90.15%	80.40%
Surrogate: InternVL3-8B (InternViT)	1.8%	1.34	0.036	93.10%	86.80%	1.37	0.054	91.40%	81.20%
Full SemantiTrace (Surr: Qwen3-VL)	**2.0%**	**1.34**	**0.035**	**98.80%**	**88.50%**	**1.37**	**0.053**	**95.40%**	**83.20%**

Table 25: Direct retrieval counts for the no-retriever-guidance ablation. Both arms use 500 hosts and 2,000 pre-filter queries.

Dataset	Full Recall@5	Ablated Recall@5	Zero-hit hosts	Host (b, c)	Rank range	Rank mean
MMQA	1996/2000	0/2000	500/500	500,0	6–100	76800/2000 = 38.40
WebQA	1983/2000	0/2000	500/500	500,0	6–100	85000/2000 = 42.50

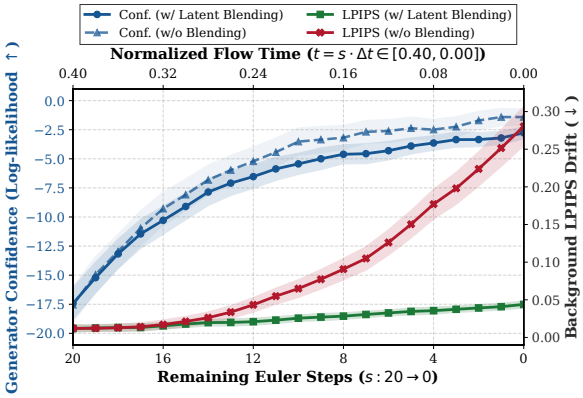


Figure 13: Final 20 Euler steps on the 100-host WebQA diagnostic subset, measured before the support-preserving output projection; bands are ± 1 sample standard deviation. Background LPIPS describes decoded intermediate states, not delivered off-support pixel error.

E.2 SOFT BLENDING AND BOUNDARY REGULARIZATION

Soft latent blending replaces a binary latent step with a feathered transition, and the boundary loss penalizes seam energy on the pixel transition field (Table 26).

Table 26: Boundary artifact and stealth measurements on the single R0 WebQA cohort of 500 attempted hosts. Rejections are integer image counts (no averaging over runs). SBE is Sobel boundary energy along the mask boundary; natural edge energy need not be zero.

Blending configuration	SBE	Canny rej.	LPIPS	OOD rej.
No blending, global injection	0.08	3/500 (0.6%)	0.284	432/500 (86.4%)
Sharp latent blending	2.45	93/500 (18.6%)	0.058	49/500 (9.8%)
DSSB + boundary loss	**0.12**	**2/500 (0.4%)**	**0.053**	**11/500 (2.2%)**

E.3 TYPOGRAPHIC PRECISION, SALIENCY AND LOCALIZED INSPECTION

Table 27: Typographic and localized-inspection controls. TA is typographic accuracy and SR is the saliency-rank percentile of the trigger region.

Panel	Method	TA	Qwen3 SR	Qwen3 CGSR	GPT-5.2 SR	GPT-5.2 CGSR	Global rej.	Crop rej.	Fidelity
In-context	Prompt-only inpainting	54.2%	32.4%	55.4%	35.8%	46.5%	n/a	n/a	n/a
In-context	Adversarial PGD	n/a	84.1%	39.3%	89.4%	28.5%	n/a	n/a	n/a
In-context	**SemantiTrace**	**95.6%**	**2.1%**	**98.8%**	**2.4%**	**88.5%**	n/a	n/a	n/a
Inspection	AQUA acronym	n/a	n/a	n/a	n/a	n/a	94.1%	98.6%	failed
Inspection	**Mode A text mutation**	n/a	n/a	n/a	n/a	n/a	**0.69%**	**76.6%**	host layout retained

The MMQA Mode-A inspection counts are 2/290 globally and 222/290 on the oracle local crop; AQUA has 273/290 and 286/290. The subset has 3/290 combined pre-ingestion rejections, contributing to the pooled Mode-A 0.95% over 423 hosts.

Table 27 shows that generator guidance improves rendered character accuracy from 54.2% to 95.6% and moves the trigger into the most salient local region. Pixel-level PGD remains weak under direct context placement, reaching 28.5% CGSR on GPT-5.2.

E.4 COMPLETE-OBJECTIVE COMPONENT CONTROLS

The five-run host-block endpoint gives a 4.07-point mean All-K gain, a 99% host interval of [0.80,7.33], and an exact label-swap tail of 0.000787. Table 28 reports the per-run margins.

Table 28: Spatial gate versus global clipping on the same 423 Mode-A hosts across five runs.

Run	Gate CGSR	Global CGSR	All-K gate/global	(b, c)
R0	88.77%	86.05%	346/330	78,62
R1	88.89%	85.93%	347/327	79,59
R2	88.89%	85.46%	345/326	79,60
R3	88.36%	85.52%	345/326	78,59
R4	88.71%	87.00%	346/334	77,65

Each effect is $100\sum_i q_i/(5N)$. Write $a_j = \#\{i : |q_i| = j\}$ and $L = \sum_j j a_j$. The exact one-sided host-label-swap tail is the coefficient sum above $(L + \sum_i q_i)/2$ in $\prod_{j=1}^{5}[(1 + z^j)/2]^{a_j}$, inclusive at equality. All eight contrasts form the Holm family. The 99% host-block intervals in Figure 14 are exact percentile-bootstrap intervals: take the .005 and .995 left-continuous quantiles of $[\sum_{q=-5}^{5}(n_q/N)z^{q+5}]^N$, subtract $5N$, and multiply by $100/(5N)$. This resamples whole five-run hosts and requires no assumption that the five runs on one host are independent.

Tables 29 and 30 give the complete Mode-A and Mode-B controls. With the same full dual guidance, SemantiTrace gains 3.78/14.61 CGSR/All-K points on Mode A and 3.91/11.13 points on all 577 Mode-B hosts over the specialized controls; Figure 14 shows the host-block intervals.

Table 29: Matched soft-mask, boundary-loss and spatial-gate decomposition on 423 Mode-A hosts, plus the specialized text editor with the same dual optimization.

Configuration	CGSR	All-K	Rejected	LPIPS mean	s/image
Hard + global	77.64%	63.03%	191/2115	0.051	120.3
Hard + SGGS	78.64%	64.49%	188/2115	0.043	121.0
Hard + boundary + global	79.79%	67.66%	160/2115	0.044	120.4
Hard + boundary + SGGS	83.09%	72.43%	115/2115	0.036	121.1
Soft + global	82.77%	71.63%	106/2115	0.038	120.3
Soft + SGGS	84.89%	75.46%	86/2115	0.030	121.1
Soft + boundary + global	85.99%	77.68%	55/2115	0.031	120.4
Soft + boundary + SGGS	**88.72%**	**81.75%**	**20/2115**	**0.023**	**121.0**
AnyText2 initialization + full guidance	84.94%	67.14%	63/2115	0.027	131.9

Table 30: Complete-objective controls on all 577 Mode-B hosts over five runs.

Configuration	Query hits / CGSR	All-K	Rejected	LPIPS	s/image
Hard + global	7728/11540 (66.97%)	1460/2885 (50.61%)	267/2885	0.083	130.2
Hard + SGGS	7859/11540 (68.10%)	1502/2885 (52.06%)	254/2885	0.076	130.8
Hard + boundary + global	7926/11540 (68.68%)	1574/2885 (54.56%)	224/2885	0.075	130.4
Hard + boundary + SGGS	8354/11540 (72.39%)	1731/2885 (60.00%)	179/2885	0.066	131.0
Soft + global	8377/11540 (72.59%)	1718/2885 (59.55%)	168/2885	0.070	130.2
Soft + SGGS	8625/11540 (74.74%)	1829/2885 (63.40%)	142/2885	0.063	130.9
Soft + boundary + global	9157/11540 (79.35%)	2002/2885 (69.39%)	114/2885	0.060	130.4
SemantiTrace	**9349/11540 (81.01%)**	**2127/2885 (73.73%)**	**80/2885**	**0.059**	**131.0**
Object insertion + AnyText2 + full guidance	8898/11540 (77.11%)	1806/2885 (62.60%)	107/2885	0.061	142.1

Table 31: Host-block difference distributions underlying Figure 14. Let q_i be full-method All-K successes minus control successes over the five runs. Counts at $q = \pm 4, \pm 5$ are zero.

Mode	Control	-3	-2	-1	0	1	2	3
A	Soft + boundary + global	6	14	159	25	159	48	12
A	Soft + SGGS	6	1	0	300	84	27	5
A	Hard + boundary + SGGS	5	3	2	263	82	66	2
A	AnyText2 + full guidance	0	0	117	3	211	61	31
B	Soft + boundary + global	1	25	56	299	169	16	11
B	Soft + SGGS	15	0	0	324	170	31	37
B	Hard + boundary + SGGS	2	5	80	107	320	17	46
B	AnyText2 + full guidance	0	14	77	135	311	5	35

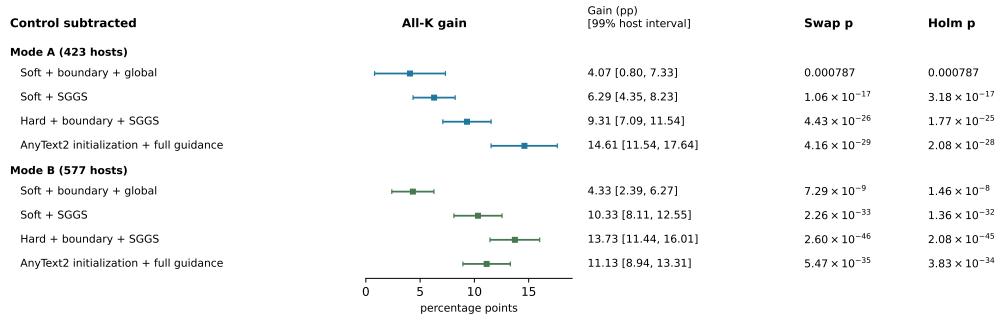


Figure 14: Incremental All-K effects on Mode A and Mode B. Squares are gains and whiskers are exact 99% percentile-bootstrap host intervals from Table 31; each host contributes its five-run paired difference as one block.

E.5 Component effects under the joint deployment

Table 32 gives the joint-deployment factorial.

Table 32: Joint-deployment factorial and paired All-K contrasts. All 500 source slots remain in every row.

Blending	Boundary	Gradient/editor	$(n_0, \ldots, n_4)$	Hits / CGSR	All-K
Hard	Off	Global	(194,79,73,45,109)	796/2000 (39.80%)	109/500
Hard	Off	Spatial	(185,76,72,44,123)	844/2000 (42.20%)	123/500
Hard	On	Global	(176,75,66,46,137)	893/2000 (44.65%)	137/500
Hard	On	Spatial	(167,73,64,43,153)	942/2000 (47.10%)	153/500
Soft	Off	Global	(159,72,63,45,161)	977/2000 (48.85%)	161/500
Soft	Off	Spatial	(151,70,58,43,178)	1027/2000 (51.35%)	178/500
Soft	On	Global	(141,76,49,39,195)	1071/2000 (53.55%)	195/500
Soft	On	Spatial	**(132,74,40,34,220)**	**1136/2000 (56.80%)**	**220/500**
AT-full	On	Spatial	(155,51,53,71,170)	1050/2000 (52.50%)	170/500

Contrast	b/c	Gain (pp)	99% interval (pp)	Exact p	Holm p
Soft vs. hard, other factors on	90/23	13.40	[8.20,18.80]	7.44×10^{-11}	2.98×10^{-10}
Boundary on vs. off, soft/spatial	63/21	8.40	[3.80,13.20]	2.48×10^{-6}	7.09×10^{-6}
Spatial vs. global, soft/boundary	42/17	5.00	[1.20,9.00]	7.74×10^{-4}	7.74×10^{-4}
Full editor vs. AT-full	84/34	10.00	[4.60,15.40]	2.36×10^{-6}	7.09×10^{-6}

Soft blending, boundary loss and spatial gating give conditional All-K gains of 13.40, 8.40 and 5.00 points at the complete setting. All three pass the four-comparison Holm family, as does the 10.00-point full-editor contrast.

E.6 Initializer-specific tuning

Table 33 gives the development grid and confirmation results. Each initializer receives an independent development choice over $t_{\text{start}} \in \{0.20, 0.40, 0.60, 0.80\}$ and $S \in \{20, 30, 40\}$ on the same 50 development hosts. This raises AnyText2's initialization coefficient from 0.20 to 0.60 at its selected $t_{\text{start}} = 0.40$, while SemantiTrace keeps 0.80/40 steps.

Table 33: Initializer development selection and held-out complete-pipeline confirmation under a common total budget. The left-grid entries are successful probes out of $50 \times 4 = 200$, not host counts.

t_{start}	Steps	Host	AnyText2	Glyph
0.20	20	124	148	140
0.20	30	130	153	148
0.20	40	132	155	151
0.40	20	150	159	151
0.40	30	156	166	158
0.40	40	160	**170**	161
0.60	20	162	157	155
0.60	30	168	164	161
0.60	40	172	167	**165**
0.80	20	166	152	149
0.80	30	172	160	156
0.80	40	**176**	164	160

Condition/editor	$(n_0, \ldots, n_4)$	Hits/2000	CGSR	ST b/c	Gain [99% int.]	Holm p
Global / ST	(31,58,29,10,372)	1634	81.70%	n/a	n/a	n/a
Global / AT-tuned	(49,20,36,67,328)	1605	80.25%	78/34	[3.60,14.00]	5.87×10^{-5}
Global / Glyph-full	(53,11,40,95,301)	1580	79.00%	102/31	[8.40,19.80]	1.47×10^{-9}
Repair / ST	(65,88,40,20,287)	1376	68.80%	n/a	n/a	n/a
Repair / AT-tuned	(84,62,48,46,260)	1336	66.80%	51/24	[1.00,9.80]	0.001222
Repair / Glyph-full	(93,45,57,59,246)	1320	66.00%	65/24	[3.60,13.00]	3.69×10^{-5}
Joint PO / ST	(132,74,40,34,220)	1136	56.80%	n/a	n/a	n/a
Joint PO / AT-tuned	(149,52,40,72,187)	1096	54.80%	62/29	[1.80,11.60]	7.05×10^{-4}
Joint PO / Glyph-full	(155,48,48,72,177)	1068	53.40%	70/27	[3.60,13.40]	3.69×10^{-5}

Under initializer-specific tuning, SemantiTrace gains 8.80, 5.40 and 6.60 All-K points over AT-tuned under global filtering, repair and joint PO. The corresponding query success values are 81.70%, 68.80% and 56.80% for SemantiTrace versus 80.25%, 66.80% and 54.80% for AT-tuned. The fully guided glyph initializer is also lower in all three conditions, and all six contrasts pass Holm correction.

E.7 Hyperparameter sensitivity

Figure 15 shows the post-hoc sensitivity grid. The operating point uses $\lambda_{ret} = 2.5$ and $\lambda_{gen} = 4.0$. Weak retriever guidance at $\lambda_{ret} = 0.5$ gives 0% CGSR despite 0.8% OOD rejection, while $\lambda_{ret} = 5.0$ raises OOD rejection to 12.4%. For generator guidance, $\lambda_{gen} \leq 2.5$ yields 21.5%–68.4% CGSR; $\lambda_{gen} = 8.0$ raises OOD rejection to 15.6%. The surrogate-selected point lies inside the stable basin shown in the figure.

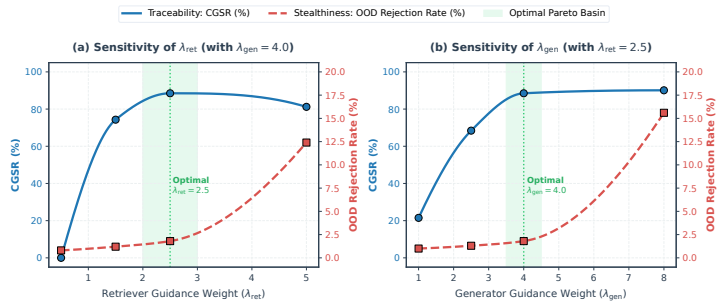


Figure 15: Post-hoc GPT-5.2 sensitivity on MMQA; the operating point is $(\lambda_{ret}, \lambda_{gen}) = (2.5, 4.0)$.

F Computational and Index Scalability

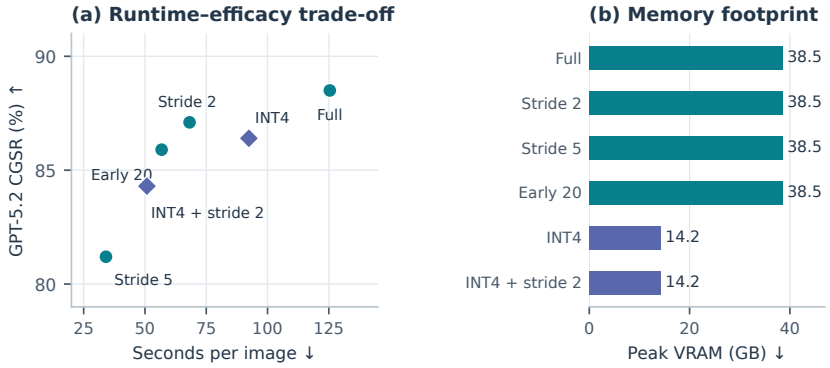


Figure 16: Computation and efficacy trade-offs on one A100. Points correspond to Table 35.

F.1 Scalable Mining

Exact isolation scoring over $|\mathcal{D}|$ images requires exhaustive nearest-neighbor search in CLIP dimension D. HNSW approximate nearest neighbors keep the same selection rule: neighbors are queried against the full gallery, and the winning anchor is selected inside each semantic cluster. Expected construction is $\mathcal{O}(|\mathcal{D}| \log |\mathcal{D}| D)$ and expected candidate querying is $\mathcal{O}(|\mathcal{D}| M \log |\mathcal{D}| D)$ for bi-connection limit M.

For 10^7 candidates, exact all-pairs search is projected at 518.75 GPU-days. The benchmark evaluates $C = 500$ fixed clusters with 10^5 candidates against the full 10^7 gallery, then compares with the 10^7-item HNSW graph. The resulting mean anchor-set overlap is $J = 96.08\%$, with downstream CGSR delta -0.18% (Table 34).

Table 34: HNSW mining fidelity and cost.

Scale	Exact time	HNSW time	Speedup	J	CGSR Δ
10^5	1.24 h	1.45 min	51.3×	98.60%	-0.05%
10^6	124.50 h	15.20 min	491.4×	97.45%	-0.12%
10^7	518.75 d	168.40 min	4435.9×	96.08%	-0.18%

Figure 16 summarizes the cost-efficacy trade-off. Resource profiling uses one NVIDIA A100 with FLUX.2-klein-base-9B, Qwen3-VL-8B-Instruct as surrogate, and the same $S = 50$ schedule. Guidance striding computes dual gradients every I steps; early window guidance uses $t \in [0.80, 0.40]$; INT4 uses 4-bit NF4 for the surrogate with activation checkpointing. Quantization plus stride-2 guidance uses 14.2 GB and 50.8 s/image while retaining 84.30% MMQA/GPT-5.2 CGSR. The device is an A100-80GB; batch size is one. The full rows keep the BF16 velocity backbone and surrogate resident, with the VAE and feature encoders loaded by phase. In the INT4 rows, NF4 applies to the

surrogate and CPU offload applies to the BF16 velocity backbone: its blocks are streamed for the stop-gradient forward, then evicted before the differentiable guidance phase. The backbone is never simultaneously resident in full in these rows. VAE/CLIP/DINO modules are staged by phase and activations are checkpointed. VRAM is peak `torch.cuda.max_memory_allocated()` over the complete edit, divided by 10^9; it includes resident tensors and activations, excludes allocator reserve and host RAM, and is reset after warm-up. Wall time includes all CPU–GPU transfers, forward/backward passes, and final decoding/projection, with CUDA synchronization at both endpoints; model download and one-time loading into host RAM are excluded.

Table 35: Resource profiling and ultra-large index retrieval. The top block reports resource profiles; the bottom block reports index-size retrieval.

Optimization strategy	Time/img	VRAM	Speedup	Qwen3 CGSR	GPT-5.2 CGSR	OOD rej.
Full SemantiTrace	125.4 s	38.5 GB	1.0×	98.80%	88.50%	2.0%
Guidance striding $I = 2$	68.2 s	38.5 GB	1.84×	97.10%	87.10%	1.9%
Guidance striding $I = 5$	34.1 s	38.5 GB	3.68×	89.20%	81.20%	2.1%
Early window $W_{early} = 20$	56.8 s	38.5 GB	2.21×	96.40%	85.90%	1.8%
Quantized surrogate INT4	92.4 s	14.2 GB	1.36×	97.05%	86.40%	2.0%
Quantized + striding $I = 2$	50.8 s	14.2 GB	2.47×	95.10%	84.30%	2.1%
Index strategy	**Metric**		10^4	10^5	10^6	10^7
Random host	Recall@5		92.40%	81.15%	64.30%	43.15%
Random host	CGSR		81.30%	71.40%	56.55%	37.95%
Target-aware	Recall@5		99.80%	99.15%	98.40%	97.10%
Target-aware	CGSR		88.50%	87.95%	87.10%	85.80%

For an isotropic reference on $\mathbb{S}^{D-1}$ with $D = 768$ and target-query similarity $\tau \geq 0.75$, an unrelated distractor has tail probability

$$P(s_{dist} \geq \tau) = \frac{1}{\sqrt{\pi}} \frac{\Gamma(D/2)}{\Gamma((D-1)/2)} \int_0^{\arccos(\tau)} \sin^{D-2}(\phi)\, d\phi$$
$$= \frac{1}{2} I_{1-\tau^2}\left(\frac{D-1}{2}, \frac{1}{2}\right) \approx \frac{1}{\tau\sqrt{2\pi D}} (1-\tau^2)^{(D-1)/2}. \tag{24}$$

At $D = 768$ and $\tau = 0.75$, this is 3.96×10^{-140}, or 3.96×10^{-133} expected distractors at $M = 10^7$. The empirical index block in Table 35 gives the clustered LAION-5B distractor result.

G Output-Only Private Audit and Corpus-Wide Repair

G.1 Private-assignment audit

The prospective audit removes the matched clean endpoint by randomizing the released signature. For host i, the owner seals $m_i = 8$ admissible candidates, assignment-independent queries and a fixed audit schedule before release. One candidate W_i is drawn uniformly and rendered once; failures stay in the attempted-host denominator. The service receives no candidate list, selected string or query hint. After the four-response transcript closes, let

$$H_i = \{a : \text{candidate } a \text{ appears as a complete token sequence in all four responses}\}, \quad h_i = |H_i|,$$

and score $T = \sum_i \mathbb{I}[W_i \in H_i]$.

Theorem 1 (Private-assignment conditional calibration) *Let $\mathcal{F}$ contain the fixed catalogs, hosts, assignment-independent queries and audit schedule. Under non-ingestion, suppose the complete transcript R is independent of the assignment vector W conditional on $\mathcal{F}$. Conditional on $(R, \mathcal{F})$, T is Poisson-binomial with probabilities h_i/m_i and generating function*

$$B(z) = \prod_{i=1}^{N} \left(1 - \frac{h_i}{m_i} + \frac{h_i}{m_i} z\right).$$

For $p_{\text{rand}} = \sum_{t \geq T_{\text{obs}}} [z^t] B(z)$, $P(p_{\text{rand}} < \alpha) \leq \alpha$. The reference law uses only the closed service transcript; response dependence across anchors is allowed.

Proof. Given $(R, \mathcal{F})$, the hit sets are fixed and the W_i are independent uniform draws, so the indicators have the displayed Bernoulli probabilities. The inclusive upper tail is super-uniform for its own discrete reference law; averaging over $(R, \mathcal{F})$ gives the result. □

The coefficient recursion is $b_{i,k} = (1 - q_i)b_{i-1,k} + q_i b_{i-1,k-1}$, $q_i = h_i/m_i$, with $b_{0,0} = 1$. If all responses list all candidates, every $q_i = 1$ and the tail is 1. For one finite population let $A = s + u$, $B = d + v$, and $E = e + \ell$ count singleton, doubleton and empty/list-all intersections. In a uniformly selected n-host audit,

$$C_n = \sum_{a,b} \frac{\binom{A}{a}\binom{B}{b}\binom{E}{n-a-b}}{\binom{500}{n}} \sum_{x,y} P(X = x)P(Y = y)\,\mathbb{I}\{\pi_{a,b}(x + y) < .01\},$$

where $X \sim \text{Hypergeom}(A, s, a)$, $Y \sim \text{Hypergeom}(B, d, b)$ and $\pi_{a,b}(t) = P\{\text{Binomial}(a, 1/8) + \text{Binomial}(b, 1/4) \geq t\}$. List-all hosts add one to both the observed statistic and its reference. Non-ingestion calibration for these private audits is reported in Appendix A.4; fixed-transcript coverage appears in Table 40.

G.2 Candidate-set sufficient statistics

For each fixed population, (s, u, d, v, e, ℓ) counts, respectively, selected and unselected singleton sets, selected and unselected doubleton sets, empty sets, and list-all sets. No other set sizes occur in the reported rows. Thus $N_0 = s + u + d + v + e + \ell$ and $T_{obs} = s + d + \ell$. In particular, selected hits alone are not the test input. Structural failure slots contribute an empty set and zero selected hits, so their polynomial factor is one even when no catalogue was assigned. For every row,

$$B(z) = z^{\ell}(7/8 + z/8)^{s+u}(3/4 + z/4)^{d+v}.$$

$$p_{rand} = P\{\text{Bin}(s + u, 1/8) + \text{Bin}(d + v, 1/4) \geq s + d\}.$$

Use these counts in the without-replacement coverage sum above, replacing 500 in its denominator by the printed N_0. The minimum budget is the first integer n with $C_n \geq .95$, not a rounded continuous approximation. Empty and list-all sets can be pooled for coverage because list-all adds the same constant to the observed statistic and reference.

Table 36: Sufficient candidate-set counts for four-probe private intersections. Each row contains 500 attempted hosts. Counts reproduce the fixed-transcript coverage and the joint full-population tails.

Population/editor	s	u	d	v	e	ℓ	Min. hosts
MMQA/PV/ST	320	25	7	7	139	2	9
MMQA/PV/AT-full	268	22	8	10	190	2	11
WebQA/PV/ST	289	34	9	8	158	2	10
WebQA/PV/AT-full	237	36	9	10	206	2	13
MMQA/PO/ST	286	20	8	7	177	2	10
MMQA/PO/AT-full	231	34	9	11	213	2	14
WebQA/PO/ST	256	23	10	10	199	2	11
WebQA/PO/AT-full	207	31	9	11	240	2	16
Joint/ST	209	25	9	16	239	2	16
Joint/AT-full	160	33	8	19	278	2	23
D3/ST	178	32	10	10	268	2	19
D3/AT-full	132	54	10	10	292	2	31

For prefix endpoints, $H_i^{U,K}$ is the union of candidate hits in the first K answers and $H_i^{I,K}$ their intersection. They coincide when $K = 1$. The prefixes are fixed in advance; the selected hit is always membership of the same privately assigned candidate. Table 37 supplies the distinct set counts rather than substituting selected-hit totals into the randomization law. The 4,000-host selected ST row gives $C_{10} = 93.9635\%$ and $C_{11} = 96.0286\%$; the AT-full row gives $C_{12} = 94.1702\%$ and $C_{13} = 95.8327\%$, and AT-tuned gives 94.1774% and 95.8251%. These fixed-population coverages establish the 11/13/13 minima. Fresh C11 serving interventions are a separate power estimate: $384/400 = 96.00\%$, with two-sided 99% Clopper–Pearson interval $[92.75\%, 98.09\%]$, not a 99% lower confidence guarantee that power exceeds 95%.

Table 37: Private-prefix sufficient statistics. Development and confirmation populations are independent. U/I denote union/intersection. All rows use $p < .01$ and cost Kn output queries.

N_0	Editor	Rule	s	u	d	v	e	ℓ	Min. n
500	ST	U1	279	25	9	16	169	2	11
500	ST	U2	329	25	9	16	119	2	9
500	ST	U4	357	25	9	16	91	2	8
500	AT-full	U1	255	33	8	19	183	2	13
500	AT-full	U2	310	33	8	19	128	2	10
500	AT-full	U4	335	33	8	19	103	2	9
200	ST	U1	111	18	4	4	61	2	11
200	ST	U2	133	24	4	4	33	2	9
200	ST	U4	145	25	4	4	20	2	8
200	AT-full	U1	99	20	4	4	71	2	13
200	AT-full	U2	117	24	4	4	49	2	11
200	AT-full	U4	133	24	4	4	33	2	9
200	AT-tuned	U1	102	24	4	4	64	2	13
200	AT-tuned	U2	123	24	4	4	43	2	10
200	AT-tuned	U4	134	25	4	4	31	2	9
4000	ST	U1	2210	307	80	80	1307	16	11
4000	AT-full	U1	1998	382	80	80	1444	16	13
4000	AT-tuned	U1	2024	418	80	80	1382	16	13
200	ST	I2	88	5	4	4	97	2	13
200	ST	I4	78	5	4	4	107	2	15
200	AT-full	I2	70	12	4	4	108	2	18
200	AT-full	I4	59	12	4	4	119	2	23
200	AT-tuned	I2	77	10	4	4	103	2	16
200	AT-tuned	I4	65	10	4	4	115	2	20

G.3 Natural-scene portfolios with permission-constrained release

MMQA and WebQA each start with 500 attempted hosts. The permission manifest covers $963/1000$ hosts, with 186 eligible Mode-A and 777 Mode-B canvases; 237 original-text fields are immutable. Candidate formation precedes assignment: each permitted scene must supply eight normalized-distinct variants under the same length, casing and edit-distance constraints, or eight object labels of three to five letters. Six MMQA and ten WebQA hosts fail catalogue formation, giving 470/477 assignments after the 24/13 no-canvas failures. PV uses hidden visual retrieval; PO uses hidden raw-OCR retrieval. AnyText2 + full guidance uses the same host, canvas, candidates, selected signature, probes, masks and losses (Table 38).

Table 38: Permission-constrained private portfolios and complete-objective comparison. CGSR uses 2,000 probes and All-K counts hosts out of 500.

Dataset/stack	Editor	Fail	Filter	Available	Recall@5	Query hits	CGSR	All-K
MMQA/PV	ST	10/470	12	448	1749/2000	1518/2000	75.90%	329/500
MMQA/PV	AT-full	16/470	18	436	1688/2000	1445/2000	72.25%	278/500
WebQA/PV	ST	14/477	14	449	1732/2000	1432/2000	71.60%	300/500
WebQA/PV	AT-full	20/477	20	437	1651/2000	1364/2000	68.20%	248/500
MMQA/PO	ST	10/470	20	440	1691/2000	1436/2000	71.80%	296/500
MMQA/PO	AT-full	16/470	27	427	1627/2000	1360/2000	68.00%	242/500
WebQA/PO	ST	14/477	23	440	1676/2000	1348/2000	67.40%	268/500
WebQA/PO	AT-full	20/477	31	426	1598/2000	1272/2000	63.60%	218/500

Failures and filtered images remain in the 500-host denominator. Complete paired effects are in Table 39.

Table 39: Paired complete-objective effects in the deployment portfolio. Positive discordance favors SemantiTrace.

Dataset/stack	All-K joint $(00, 01, 10, 11)$	CGSR gain	All-K gain	99% interval	Paired p	Holm p
MMQA/PV	$(140, 31, 82, 247)$	3.65	10.20	[5.00, 15.60]	8.65×10^{-7}	2.60×10^{-6}
WebQA/PV	$(164, 36, 88, 212)$	3.40	10.40	[4.80, 16.00]	1.70×10^{-6}	3.39×10^{-6}
MMQA/PO	$(172, 32, 86, 210)$	3.80	10.80	[5.40, 16.20]	3.45×10^{-7}	1.38×10^{-6}
WebQA/PO	$(199, 33, 83, 185)$	3.80	10.00	[4.60, 15.60]	1.93×10^{-6}	3.39×10^{-6}

Table 40: Exact fixed-transcript private coverage for natural portfolios. Each selected host costs four output queries and no clean queries.

Dataset/stack/editor	$N = 8$	$N = 12$	$N = 20$	$N = 32$	$N = 64$	95% q.	P64	C20	R16	R32
MMQA/PV / ST	94.998%	99.581%	99.998%	100.000%	100.000%	36	2880/4000	3886/4000	3989/4000	4000/4000
MMQA/PV / AT-full	86.393%	97.669%	99.951%	100.000%	100.000%	44	n/a	n/a	n/a	n/a
WebQA/PV / ST	89.845%	98.515%	99.977%	100.000%	100.000%	40	2506/4000	3744/4000	3943/4000	4000/4000
WebQA/PV / AT-full	76.657%	93.611%	99.629%	99.997%	100.000%	52	n/a	n/a	n/a	n/a
MMQA/PO / ST	90.959%	98.881%	99.988%	100.000%	100.000%	40	2509/4000	3812/4000	3968/4000	4000/4000
MMQA/PO / AT-full	74.741%	92.697%	99.525%	99.995%	100.000%	56	n/a	n/a	n/a	n/a
WebQA/PO / ST	83.842%	96.840%	99.913%	100.000%	100.000%	44	2107/4000	3576/4000	3858/4000	4000/4000
WebQA/PO / AT-full	66.858%	88.300%	98.857%	99.976%	100.000%	64	n/a	n/a	n/a	n/a

P64, C20, R16 and R32 are fixed-record SemantiTrace reference calculations for 8/64, 10/20, 16/64 and 32/64 scheduled ingestion. Separate full-gallery interventions are in Table 41.

G.4 Fresh ingestion interventions

Each of 400 plans draws 64 hosts from the original 500-host manifest. C20 ingests the first 20 scheduled hosts and costs 80 output queries; P64 schedules eight of 64 and costs 256. Each intervention starts from the full 58,000-item MMQA or 400,000-item WebQA gallery, replaces selected pristine images with released copies, refreshes representations, rebuilds the index and generates all answers. Table 42 compares the fixed-record P64 reference with the independently reingested condition.

Table 41: Fresh full-gallery private ingestion interventions with matched editor budgets. Every cell contains 400 fixed-length audits.

Dataset/stack	C20 ST	C20 AT-full	P64 ST	P64 AT-full	P64 (b, c)	Holm p
MMQA/PV	400/400	397/400	284/400 (71.00%)	251/400 (62.75%)	$(52, 19)$	0.000225
WebQA/PV	399/400	392/400	244/400 (61.00%)	210/400 (52.50%)	$(56, 22)$	0.000225
MMQA/PO	399/400	393/400	241/400 (60.25%)	207/400 (51.75%)	$(58, 24)$	0.000225
WebQA/PO	397/400	385/400	203/400 (50.75%)	169/400 (42.25%)	$(60, 26)$	0.000225

Table 42: Reference-law coverage versus the independently reingested P64 condition.

Dataset/stack	Fixed-record reference	Reingested ST	99% binomial interval	Difference
MMQA/PV	72.000%	284/400 (71.00%)	[64.80, 76.69]	-1.000
WebQA/PV	62.650%	244/400 (61.00%)	[54.49, 67.24]	-1.650
MMQA/PO	62.725%	241/400 (60.25%)	[53.73, 66.52]	-2.475
WebQA/PO	52.675%	203/400 (50.75%)	[44.21, 57.27]	-1.925

G.5 Full-gallery repair and reindexing

The full-stack defense applies OCR-guided repair to all corpus items, recomputes image and OCR/context fields, rebuilds the index, retrieves again and regenerates answers. The principal WebQA run has 400,000 initial items and the same 500 targets. Background global filtering removes 1,998 items; target rejections are 11/110/28/18 for ST, natural/native AQUA, AnyText2+DSSB and AT-full. The agreement threshold .80 is selected on a 2,000-question development workload and changes the independent test workload from 886/1000 to 877/1000 correct under the 1.5 point utility budget (Table 43).

For the paired All-K budget curve, fix N_0 hosts with b_0 positive discordances, c_0 negative discordances and $d_0 = N_0 - b_0 - c_0$. Sampling n hosts uniformly without replacement gives

$$C_n = \sum_{x,y} \frac{\binom{b_0}{x}\binom{c_0}{y}\binom{d_0}{n-x-y}}{\binom{N_0}{n}} \mathbb{I}\{P(\text{Binomial}(x+y, 1/2) \geq x) < \alpha\}.$$

For the repaired full gallery, the discordances are $(287, 0)$, $(118, 0)$, $(198, 0)$ and $(246, 0)$.

Table 43: Repair, reindexing and regenerated answers on the same WebQA targets. The 400,000-item row is the full-corpus scenario.

Method	Gallery	Recall@5	Query hits / CGSR	All-K	Paired p
ST	1,500	1832/2000	1408/2000 (70.40%)	300/500	4.91×10^{-91}
ST	400,000	1796/2000	1376/2000 (68.80%)	287/500	4.02×10^{-87}
AQUA natural/native	1,500	1315/2000	907/2000 (45.35%)	132/500	3.67×10^{-40}
AQUA natural/native	400,000	1284/2000	858/2000 (42.90%)	119/500	3.01×10^{-36}
AnyText2 + DSSB	1,500	1684/2000	1156/2000 (57.80%)	211/500	3.04×10^{-64}
AnyText2 + DSSB	400,000	1658/2000	1104/2000 (55.20%)	198/500	2.49×10^{-60}
AT-full	1,500	1812/2000	1344/2000 (67.20%)	258/500	2.16×10^{-78}
AT-full	400,000	1778/2000	1302/2000 (65.10%)	246/500	8.84×10^{-75}

Table 44: Full-gallery audit coverage under repair. Entries are exact percentages for without-replacement host subsets under the paired All-K rule.

Method	$N = 8$	$N = 12$	$N = 20$	$N = 40$	$N = 80$	95% hosts / queries
ST	8.01%	59.53%	98.90%	100.00%	100.00%	18 / 144
AQUA natural/native	0.02%	0.94%	16.79%	87.55%	99.99%	47 / 376
AnyText2 + DSSB	0.76%	14.84%	74.29%	99.95%	100.00%	27 / 216
AT-full	3.08%	36.41%	93.75%	100.00%	100.00%	21 / 168

Each sampled host costs four probes in each of two arms. Calibration of defended non-ingestion audits is in Appendix A.4.

G.6 JOINT DEPLOYMENT

The joint WebQA/PO experiment applies owner-permitted canvases, eight-candidate assignment, a hidden raw-OCR service, no clean endpoint and the full-gallery repair policy. After repair, OCR documents are regenerated, BGE-M3 representations refreshed, the index rebuilt and all answers newly generated (Table 45). The joint private query costs are 64/92 output queries with no clean arm.

Table 45: Joint permission-constrained private deployment with repaired WebQA/PO service. The auditor has no clean endpoint.

Editor	Build/assigned	Rejected	Available	OTR	Recall@5	Query matches	All-K	Full p_{rand}	95% q.
ST	463/477	23	440	336/500	1436/2000	1136/2000 (56.80%)	220/500	1.58×10^{-145}	64
AT-full	457/477	31	426	310/500	1364/2000	1050/2000 (52.50%)	170/500	5.08×10^{-98}	92

A fixed repair policy preserves the private test: under non-ingestion, the repaired index and final transcript are measurable functions of the pristine gallery, fixed queries and serving/repair randomness. If these variables are independent of W conditional on the fixed catalogues and queries, conditioning on the final transcript leaves the same product-uniform assignment law.

Table 46: Matched-editor effects when all deployment constraints hold together. Joint bits are (ST, AT-full).

Endpoint	ST	AT-full	$(00, 01, 10, 11)$	Gain	99% interval	Paired p	Holm p
Primary All-K	220/500	170/500	$(246, 34, 84, 136)$	10.00	[4.60, 15.40]	2.36×10^{-6}	4.72×10^{-6}
Fresh C20	386/400	360/400	$(6, 8, 34, 352)$	6.50	[2.50, 10.75]	3.44×10^{-5}	n/a
Fresh P64	165/400	125/400	$(216, 19, 59, 106)$	10.00	[4.50, 16.00]	3.21×10^{-6}	4.72×10^{-6}

G.7 Development-selected countermeasures

A separate 200-host development set evaluates eight corpus-wide policies on 100 20-host complete-ingestion plans. The selected defender minimizes ST private declarations subject to at most 30 lost answers on a 2,000-question workload. Policy D3 lowers the anomaly threshold to .65, removes the confidence gate, sets rarity .90 and agreement .86, and is then evaluated on held-out hosts (Table 47). Held-out D3 results are in Table 48, and normal-task utility is summarized in Table 49.

Table 47: Development-only selection of a suppressive corpus-wide policy.

Policy	τ_a	τ_r	Confidence gate	τ_A	Net QA losses	Correct	Audit declarations	Decision
D0	.82	.90	$c < .80$	.80	21/2000	1751/2000	96/100	Eligible
D1	.82	.90	All	.80	83/2000	1689/2000	34/100	Over budget
D2	.82	.98	All	.90	28/2000	1744/2000	81/100	Eligible
D3	.65	.90	All	.86	29/2000	1743/2000	72/100	Selected
D4	.60	.90	All	.86	47/2000	1725/2000	61/100	Over budget
D5	.65	.95	$c \geq .80$	.86	31/2000	1741/2000	67/100	Over budget
D6	.82	.90	All	.94	36/2000	1736/2000	63/100	Over budget
D7	.90	.995	All	.90	23/2000	1749/2000	84/100	Eligible

Table 48: Held-out joint deployment after the development-selected D3 defense.

Editor	$(n_0, \ldots, n_4)$	Query hits	CGSR	All-K	Min. queries
ST	$(165, 55, 54, 36, 190)$	1031/2000	51.55%	190/500	76
AT-full	$(183, 65, 59, 49, 144)$	906/2000	45.30%	144/500	124

Under D3, the P64 paired result is 145/400 (36.25%) versus 101/400 (25.25%): 87 both, 58 ST-only, 14 AT-only and 241 neither. The gain is 11.00 points, with 99% interval [5.75, 16.50] and $p = 8.23 \times 10^{-8}$. Non-ingestion calibration for the selected-defense protocol is in Appendix A.4.

Table 49: Consolidated normal-task utility for deployment conditions. Watermark probes are excluded; the 18 upper endpoints use Equation 25.

Block	Row	n	Joint bits $(00, 01, 10, 11)$	Before	After	Loss	99% upper
Private	PV / ST	237	$(15, 1, 2, 219)$	221/237	220/237	0.422	n/a
Private	PV / AT-full	237	$(14, 2, 4, 217)$	221/237	219/237	0.844	n/a
Private	PO / ST	237	$(14, 2, 4, 217)$	221/237	219/237	0.844	n/a
Private	PO / AT-full	237	$(13, 3, 6, 215)$	221/237	218/237	1.266	n/a
Repair	Operating point / all	1000	$(109, 5, 14, 872)$	886/1000	877/1000	0.900	3.125
Repair	Operating point / rare	200	$(23, 1, 5, 171)$	176/200	172/200	2.000	8.808
Repair	Operating point / other	800	$(86, 4, 9, 701)$	710/800	705/800	0.625	3.014
Repair	Independent / all	10000	$(1085, 51, 124, 8740)$	8864/10000	8791/10000	0.730	1.370
Repair	Independent / rare	2000	$(233, 14, 27, 1726)$	1753/2000	1740/2000	0.650	2.257
Repair	Independent / other	8000	$(852, 37, 97, 7014)$	7111/8000	7051/8000	0.750	1.450
Joint	ST / Rare	2000	$(256, 14, 24, 1706)$	1730	1720	0.500	2.058
Joint	ST / Other	8000	$(910, 40, 90, 6960)$	7050	7000	0.625	1.320
Joint	ST / All	10000	$(1166, 54, 114, 8666)$	8780	8720	0.600	1.231
Joint	AT / Rare	2000	$(258, 12, 24, 1706)$	1730	1718	0.600	2.112
Joint	AT / Other	8000	$(913, 37, 86, 6964)$	7050	7001	0.613	1.288
Joint	AT / All	10000	$(1171, 49, 110, 8670)$	8780	8719	0.610	1.223
D3	ST / Rare	4000	$(516, 24, 64, 3396)$	3460	3420	1.000	2.138
D3	ST / Other	16000	$(1844, 56, 176, 13924)$	14100	13980	0.750	1.203
D3	ST / All	20000	$(2360, 80, 240, 17320)$	17560	17400	0.800	1.225
D3	AT-full / Rare	4000	$(520, 20, 64, 3396)$	3460	3416	1.100	2.202
D3	AT-full / Other	16000	$(1846, 54, 170, 13930)$	14100	13984	0.725	1.170
D3	AT-full / All	20000	$(2366, 74, 234, 17326)$	17560	17400	0.800	1.216

The 18 displayed utility upper endpoints in Table 49 form one simultaneous family. Let $b = n_{10}$ be losses, $c = n_{01}$ recoveries, and $\delta = .01/(2 \times 18)$. With one-sided Clopper–Pearson marginal limits,

$$U_{loss} = 100\{\text{Beta}^{-1}(1 - \delta; b + 1, n - b) - \text{Beta}^{-1}(\delta; c, n - c + 1)\}. \tag{25}$$

The lower recovery limit is zero for $c = 0$, and the upper loss limit is one for $b = n$. A union bound over both tails and all 18 rows gives at least 99% simultaneous coverage without assuming independence between overlapping strata or editors. The point estimate is $100(b - c)/n$; these upper endpoints concern population utility loss, not the observed net count.

H Query Efficiency and Independent Confirmation

This appendix gives the efficiency comparisons that support Figure 5. Query budgets are fixed-length audit designs. Paired designs bill suspect and clean arms; private-assignment designs bill only suspect outputs. Appendix A.4 reports whole-audit non-ingestion calibration for the selected rules.

H.1 Method-native scores

The primary All-K rule measures four-query reproducibility. For query efficiency, natural/native AQUA uses its acronym query and scores the mean fraction of complete-expansion matches across four responses, while SemantiTrace uses the Union endpoint of Section 3.4. Both scores are fixed before evaluation with a common $p < .005$ paired sign-swap threshold. AQUA's clean cohort has seven query hits in one four-hit host and three one-hit hosts; the weighted differences are positive with weights $1, 2, 3, 4$ in counts $(15, 37, 96, 239)$, one negative weight-one difference, and 112 zero differences. SemantiTrace has 468 positive Union differences and no negative differences. Table 50 gives the resulting subset-coverage curves and 95% budgets.

Table 50: Method-native, independently calibrated audit scores on the same WebQA portfolios.

Score / method	$N = 8$	$N = 10$	$N = 12$	$N = 16$	$N = 20$	$N = 32$	95% hosts / queries
Union / SemantiTrace	58.68%	97.90%	99.95%	100.00%	100.00%	100.00%	10 / 80
Mean CER / natural-native AQUA	12.67%	59.70%	88.88%	99.67%	100.00%	100.00%	14 / 112

H.2 Endpoint-fair efficiency

The complete-objective editor comparison lets each editor choose among All-K, Union and mean CER on a disjoint 200-host development portfolio. The criterion is the smallest host budget with 95% finite-cohort coverage at $p < .005$. Both editors choose Union under global filtering and full-gallery repair. Table 51 gives the held-out WebQA costs on 500 paired hosts. The selected rules need 80 versus 88 queries after global filtering and 88 versus 104 after repair. The corresponding point savings are 9.09% and 15.38%. The repaired response histograms contain $435/500 = 87.00\%$ versus $398/500 = 79.60\%$ Union, $1376/2000 = 68.80\%$ versus $1302/2000 = 65.10\%$ query hits, and $287/500 = 57.40\%$ versus $246/500 = 49.20\%$ All-K anchors for SemantiTrace and AT-full. Mean CER gives the same integer budgets in all four rows; All-K also favors SemantiTrace at the common .005 threshold.

Table 51: Endpoint-fair WebQA query costs at $p < .005$. Union is the development-selected score.

Condition/editor	Host score	Positive differences	Min. queries
Global / ST	All-K	372	112
Global / ST	Union	468	80
Global / ST	Mean CER	469	80
Global / AT-full	All-K	291	152
Global / AT-full	Union	437	88
Global / AT-full	Mean CER	438	88
Repair / ST	All-K	287	160
Repair / ST	Union	434	88
Repair / ST	Mean CER	435	88
Repair / AT-full	All-K	246	192
Repair / AT-full	Union	397	104
Repair / AT-full	Mean CER	398	104

H.3 Probe-count choice

The efficiency comparison also allows $K \in \{1, 2, 4\}$, using the first K of the four fixed questions. A paired Union audit with n hosts costs $2Kn$ calls, while a private audit costs Kn calls. Development selection is independent of the evaluation transcripts and fixes one rule before testing.

Table 53 gives the paired-prefix and private-prefix budgets. The one-probe paired rules cost 24/26 queries after global filtering and 32/34 after repair; the joint private four-probe intersection budgets remain 64/92 queries.

Host-level budget uncertainty uses 100,000 bootstrap replicates, NumPy PCG64 seed 20260925, and the .005/.995 inverse-empirical-CDF quantiles of $q_{AT} - q_{ST}$ (integer savings). Every draw resamples complete paired hosts and recomputes the exact integer minimum budget; the development-selected rule stays fixed. Multinomial host draws are equivalent to index resampling. The generator is advanced continuously in this order: the six 500-host rows of Table 53, the four 2,000-host rows of Table 54, then the 4,000-host ST/AT-full and ST/AT-tuned comparisons. The paired-positive joint counts $(00, 01, 10, 11)$ for the six 500-host rows, in table order, are $(66, 25, 42, 367)$, $(39, 25, 31, 405)$, $(24, 8, 39, 429)$, $(126, 30, 48, 296)$, $(81, 35, 50, 334)$, and $(46, 20, 57, 377)$. All negative suspect-versus-clean discordances are zero. The 2,000-host joint counts are printed in their table. For the one-probe private comparison, use the following joint category counts, with rows ST and columns comparator, both ordered (s, u, d, v, e, ℓ):

$$J_{ST,AT} = \begin{pmatrix} 1798 & 75 & 0 & 0 & 337 & 0 \\ 0 & 307 & 0 & 0 & 0 & 0 \\ 0 & 0 & 80 & 0 & 0 & 0 \\ 0 & 0 & 0 & 80 & 0 & 0 \\ 200 & 0 & 0 & 0 & 1107 & 0 \\ 0 & 0 & 0 & 0 & 0 & 16 \end{pmatrix},$$

Table 52: Development-set selection for endpoint, paired-prefix and joint private designs.

Condition/editor	All-K hosts	Union hosts	Mean-CER hosts	Selected
Global / ST	15	10	10	Union
Global / AT-full	21	11	11	Union
Repair / ST	21	12	12	Union
Repair / AT-full	28	13	13	Union

Condition/editor	$K = 1$ queries	$K = 2$ queries	$K = 4$ queries	Selected
Global / ST	26	44	80	$K = 1$
Global / AT-full	28	48	88	$K = 1$
Repair / ST	32	52	96	$K = 1$
Repair / AT-full	36	60	104	$K = 1$

Editor	K	Set rule	Min. hosts	Queries
ST	1	Union/Intersection	11	11
ST	2	Union	9	18
ST	4	Union	8	32
ST	2	Intersection	13	26
ST	4	Intersection	15	60
AT-full	1	Union/Intersection	13	13
AT-full	2	Union	11	22
AT-full	4	Union	9	36
AT-full	2	Intersection	18	36
AT-full	4	Intersection	23	92
AT-tuned	1	Union/Intersection	13	13
AT-tuned	2	Union	10	20
AT-tuned	4	Union	9	36
AT-tuned	2	Intersection	16	32
AT-tuned	4	Intersection	20	80

Table 53: Low-probe paired and private query costs on the WebQA portfolios. The paired block bills both arms; the private block bills suspect outputs only.

Condition	K	Positive ST/AT	Min. queries ST/AT	Saving	99% saving range
Global	1	409/392	24/26	2	$[0, 4]$
Global	2	436/430	44/48	4	$[-4, 4]$
Global	4	468/437	80/88	8	$[0, 16]$
Repair	1	344/326	32/34	2	$[0, 6]$
Repair	2	384/369	56/56	0	$[-4, 8]$
Repair	4	434/397	88/104	16	$[0, 24]$

Editor	K	Min. hosts	Min. queries
ST	1	11	11
ST	2	9	18
ST	4	8	32
AT-full	1	13	13
AT-full	2	10	20
AT-full	4	9	36

$$J_{ST,tuned} = \begin{pmatrix} 1824 & 111 & 0 & 0 & 275 & 0 \\ 0 & 307 & 0 & 0 & 0 & 0 \\ 0 & 0 & 80 & 0 & 0 & 0 \\ 0 & 0 & 0 & 80 & 0 & 0 \\ 200 & 0 & 0 & 0 & 1107 & 0 \\ 0 & 0 & 0 & 0 & 0 & 16 \end{pmatrix}.$$

Their margins reproduce Table 37. Bootstrap the 36 cells in row-major order, then obtain each editor's six category totals from the row/column sums. The resulting 99% saving range is $[1, 3]$ queries for each comparator. For paired rows,

$$n_{.95}(N, b) = \min\{n : P(\text{Hypergeom}(N, b, n) \geq 8) \geq .95\}, \qquad q_{.95} = 2Kn_{.95}(N, b).$$

The independent 2,000-host confirmation issued all four probes without stopping and confirmed the four-probe costs, while the one-probe selected designs required 24/28 global and 32/34 repaired queries (Table 54).

Table 54: Held-out 2,000-host confirmation of the selected probe count and endpoint.

Condition	K	$(00, 01, 10, 11)$	Positive ST/AT	Min. queries ST/AT	99% saving range
Global	1	$(270, 90, 180, 1460)$	1640/1550	24/28	$[0, 4]$
Repair	1	$(535, 85, 160, 1220)$	1380/1305	32/34	$[0, 4]$
Global	4	$(67, 58, 180, 1695)$	1875/1753	80/88	$[8, 16]$
Repair	4	$(200, 62, 210, 1528)$	1738/1590	88/104	$[8, 16]$

H.4 Selected one-probe private design and fresh audits

The final private design uses the owner-permitted WebQA source pool, private eight-candidate assignment, hidden raw-OCR PO service, and full-gallery repair/reindexing. The independent 200-host development set in Table 52 selects one probe for ST, AT-full and independently tuned AnyText2. The 4,000-host confirmation keeps all failures in the denominator: 113 no-canvas and 78 catalogue failures leave 3,809 assigned hosts. The gallery retains 400,000 source positions. SemantiTrace needs 11 output queries for at least 95% coverage, compared with 13 for both comparators (Table 55). The point saving is 15.38%, and paired 99% saving ranges are $[1, 3]$ queries against each comparator.

Table 55: Independent 4,000-host confirmation of the selected one-probe private audit.

Editor	Selected hits	Min. queries	99% saving range
ST	2306/4000	11	n/a
AT-full	2094/4000	13	$[1, 3]$
AT-tuned	2120/4000	13	$[1, 3]$

Fresh serving interventions evaluate the selected rule after complete repair, regenerated OCR, index rebuild and new answers. C11 and C13 are complete-ingestion designs; P64, P128 and P256 schedule 8, 16 and 32 ingestions among 64, 128 and 256 queried hosts. D0 is the default repair and D3 is the development-selected countermeasure.

At the final 11-query budget, ST declares in 384/400 C11 audits (Table 56). At one-eighth ingestion, the 64-query audit reaches 71.00% under D0 and 64.50% under D3, compared with 61.25%/51.50% for AT-full and 63.75%/54.25% for AT-tuned. The P64 paired effects have Holm-adjusted values 1.95×10^{-5}, 4.57×10^{-4}, 9.21×10^{-8} and 1.95×10^{-5} for D0/AT, D0/tuned, D3/AT and D3/tuned. The five budgets of the selected $K = 1$ private protocol under each of D0 and D3 have simultaneous 99% false-alarm upper bounds below 1% in Appendix A.4.

Table 56: Fresh one-probe audits and matched effects. Declarations are out of 400 paired realizations.

Defense/design	Ingestion	Queries	Declarations ST/AT/tuned	Power ST/AT/tuned (%)	ST gain vs AT/tuned (pp)
D0 / C11	11/11	11	384 / 371 / 374	96.00 / 92.75 / 93.50	3.25 / 2.50
D0 / C13	13/13	13	395 / 384 / 387	98.75 / 96.00 / 96.75	2.75 / 2.00
D0 / P64	8/64	64	284 / 245 / 255	71.00 / 61.25 / 63.75	9.75 / 7.25
D3 / P64	8/64	64	258 / 206 / 217	64.50 / 51.50 / 54.25	13.00 / 10.25
D0 / P128	16/128	128	370 / 343 / 352	92.50 / 85.75 / 88.00	6.75 / 4.50
D3 / P128	16/128	128	350 / 310 / 324	87.50 / 77.50 / 81.00	10.00 / 6.50
D0 / P256	32/256	256	395 / 383 / 387	98.75 / 95.75 / 96.75	3.00 / 2.00
D3 / P256	32/256	256	386 / 359 / 370	96.50 / 89.75 / 92.50	6.75 / 4.00

Condition / comparator	b/c	Gain (pp)	99% interval (pp)	Exact p	Holm p
D0 / C11 / AT-full	21/8	3.25	$[0.00, 6.75]$	0.012060	n/a
D0 / C11 / AT-tuned	19/9	2.50	$[-0.75, 6.00]$	0.043579	n/a
D0 / C13 / AT-full	13/2	2.75	$[0.50, 5.25]$	0.003693	n/a
D0 / C13 / AT-tuned	10/2	2.00	$[0.00, 4.50]$	0.019287	n/a
D0 / P64 / AT-full	59/20	9.75	$[4.00, 15.25]$	6.48×10^{-6}	1.95×10^{-5}
D0 / P64 / AT-tuned	51/22	7.25	$[2.00, 12.75]$	4.57×10^{-4}	4.57×10^{-4}
D3 / P64 / AT-full	72/20	13.00	$[7.00, 19.00]$	2.30×10^{-8}	9.21×10^{-8}
D3 / P64 / AT-tuned	65/24	10.25	$[4.50, 16.25]$	8.06×10^{-6}	1.95×10^{-5}
D0 / P128 / AT-full	34/7	6.75	$[2.75, 10.75]$	1.27×10^{-5}	n/a
D0 / P128 / AT-tuned	26/8	4.50	$[1.00, 8.25]$	0.001468	n/a
D3 / P128 / AT-full	50/10	10.00	$[5.25, 15.00]$	8.08×10^{-8}	n/a
D3 / P128 / AT-tuned	38/12	6.50	$[2.25, 11.00]$	1.53×10^{-4}	n/a
D0 / P256 / AT-full	13/1	3.00	$[0.75, 5.50]$	9.16×10^{-4}	n/a
D0 / P256 / AT-tuned	10/2	2.00	$[0.00, 4.25]$	0.019287	n/a
D3 / P256 / AT-full	30/3	6.75	$[3.25, 10.50]$	7.01×10^{-7}	n/a
D3 / P256 / AT-tuned	20/4	4.00	$[1.00, 7.25]$	7.72×10^{-4}	n/a

I Qualitative Examples, Probe Templates and Prompts

Figure 17 shows four representative pipeline outcomes. Cases 1 and 2 mutate existing text: “City of Bothell” to “City of Dothell”, and “lemans fc” to “cemans fc”. Cases 3 and 4 insert labeled objects or markings: a baseball branded “BSB” and a red flag marked “SSV”. The corresponding probes localize the visible evidence, and the GPT-5.2 responses transcribe the displayed signatures. These examples illustrate the visual-evidence path; aggregate extraction, filtering and matched controls are reported in the quantitative sections.

I.1 Probe query templates

Each anchor uses $K = 4$ signature-free probes. A probe combines global scene descriptors for retrieval with a local transcription request for decoding.

1. **Direct locational probe.** ``Look at the [Region/Object Description] in this image. What exact letters or words are printed on it?''
2. **Object-centric identity probe.** ``Identify the [Host Object] located near the [Reference Landmark]. What is the literal brand name or label inscribed on it?''
3. **Contextual typography probe.** ``Examine the sign/logo displayed in the scene. Transcribe the exact text characters appearing at the bottom of the sign.''
4. **Fine-grained OCR extraction probe.** ``Perform a fine-grained character transcription on the [Host Object] shown. Output the exact sequence of capitalized letters.''

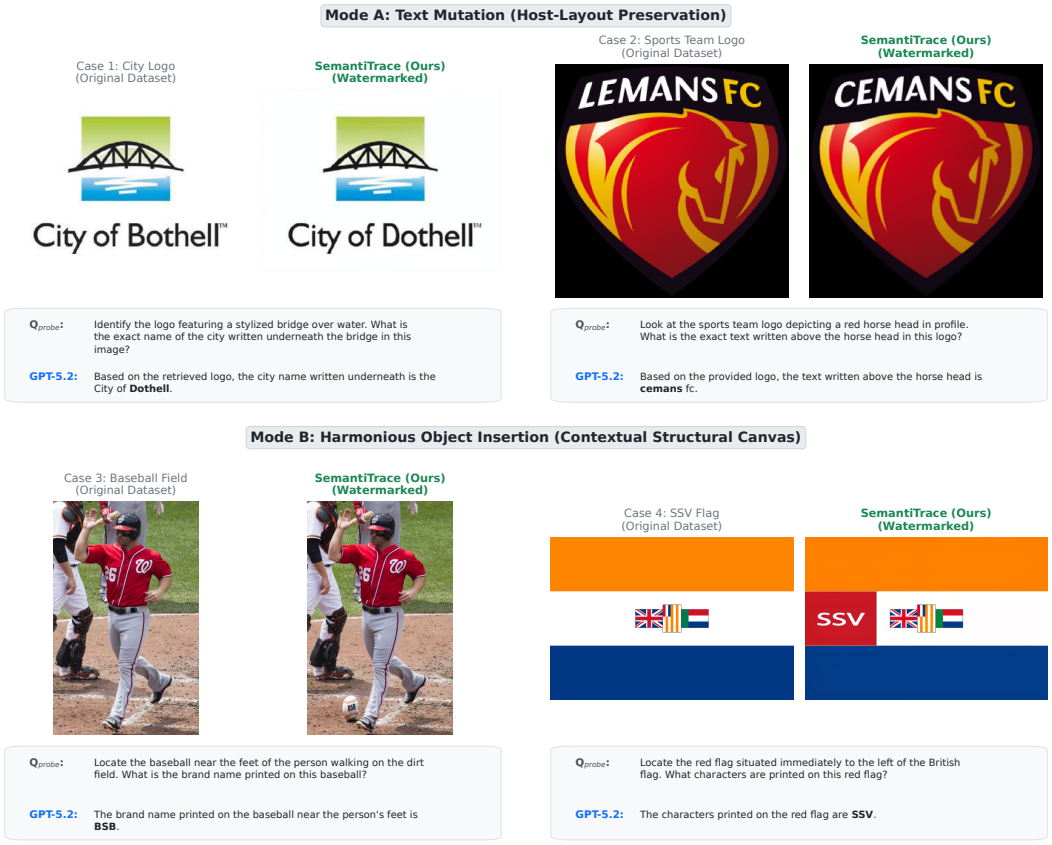


Figure 17: End-to-end examples for text mutation and object insertion. Queries describe the scene and ask for local transcription without including the signature.

I.2 CONTEXT-AWARE CANARY GENERATION PROMPT

Prompt 1: Dual-Mode Semantic Trigger Generation ($\mathcal{F}_{vlm}$)

System: You are an expert visual forensic steganography agent. Your task is to seamlessly hide a verifiable copyright signature inside a benign image using the most stealthy "Semantic Parasitism" strategy possible. Depending on the selected mode, the signature should be either: (1) A subtle 1-to-2 character typographical mutation of an existing word (preserving length and capitalization), or (2) A rare, contextually plausible 3-to-5 letter acronym/word placed onto an inserted object.
User: [Attached Image: I_{anc} overlaid with numbered bounding boxes $\mathcal{M}_{cand}$, which include both structural regions and existing text regions detected by OCR] Analyze the provided image and perform the following tasks step-by-step:

1. **Context & Typography Analysis:** Scan the image for existing text (e.g., street signs, book covers, storefronts, posters).
2. **Mode Selection:** Choose the most stealthy parasitism mode:
   - *Mode A (Text Mutation - PREFERRED):* If natural text exists in a clear bounding box, select it. Propose a subtle mutation of the original word into a rare identifier via a localized 1-to-2 character typographical deviation (e.g., mutating a storefront sign from "CAFE" to "CXFE", or "Bothell" to "Dothell"). This preserves the host layout and canvas extent, not the original glyph identity.
   - *Mode B (Object Insertion - FALLBACK):* If no editable text exists, select a structurally flat, empty bounding box (e.g., a blank desk, an empty wall). Propose a contextually harmonious object to be inserted (e.g., "a closed hardcover book") that will bear a rare 3-to-5 letter acronym (e.g., "BSB", "XCO").
3. **Trigger & Signature Formulation:** Define the final visual trigger (T_{trig}) and the exact verifiable text signature (S_{trap}).

Output Format: You must strictly output in JSON format: `{ ``selected_box_id": int, ``parasitism_mode": ``Text Mutation'' | ``Object Insertion'', ``reasoning": ``string'', ``T_trig": ``string'', ``S_trap": ``string'' }`

I.3 ANOMALY FILTER PROMPT

Prompt 2: Adversarial VLM Anomaly Filter

System: You are a strict Data Sanitization Inspector for a commercial Multimodal RAG platform. Your objective is to filter out any images that contain synthetic hallucinations, adversarial watermarks, or unnatural artifacts before they enter the retrieval database.
User: [Attached Image: Suspect Image] Scrutinize the attached image with extreme attention to detail. Flag the image as "ANOMALY" if it exhibits ANY of the following:

- **Spatial/Physical Hallucinations:** Objects floating in mid-air, incorrect shadows, or bizarre object combinations (e.g., a dog with an apple on its head).
- **Semantic Dissonance:** Objects that completely violate the context of the scene (e.g., a boat resting on an indoor office desk).
- **Textual Anomalies:** Text that appears artificially overlaid, uses fonts inconsistent with the physical object, or spells completely nonsensical/rare acronyms without context.
- **Texture Artifacts:** Blurry patches, localized noise, or blending errors typical of AI inpainting.

Provide your step-by-step analysis. Conclude your response with a final verdict on a new line: either `[VERDICT: SAFE]` or `[VERDICT: ANOMALY]`.