GHOST YOUR SYSTEM: AUDIT FIBERS HIDE HIJACKS BEHIND PASSING CHECKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deployment pipelines routinely accept non-identical model components produced by quantization, export, or compilation as long as they preserve the behavior checked during qualification. However, qualification observes only the declared component interface, whereas the deployed system later composes that component with downstream consumers. A malicious replacement must preserve every checked output on the same execution it redirects. For reusable components, this requires one fixed replacement that operates across inputs. We show that the check may discard distinctions that downstream consumers still use. The states it treats as equivalent form an *audit fiber*. We prove that qualification is unsafe whenever a component can reach two states in one fiber that induce different downstream behavior. For normalized representations, we derive an exact targeted attack within the fiber and compile it into FiberRouter, a fixed malicious replacement. Experiments across image and text representations and a Qwen model KV-cache handoff demonstrate reliable downstream hijacking while preserving the qualification that authorizes deployment.

1 INTRODUCTION

Quantization, export, and compilation change component implementations without preserving parameters. Deployment tools therefore qualify candidates behaviorally. Intel Neural Compressor uses accuracy-loss criteria, while Hugging Face Optimum validates exported ONNX outputs against reference outputs within a configurable absolute tolerance (Intel; Hugging Face, a;b). Visual retrieval systems likewise update embedding models while retaining stored gallery representations (Shen et al., 2020; Ramanujan et al., 2022; Pan et al., 2023; Zhou et al., 2023a; Jang & Lim, 2025). Qualification therefore observes a declared component interface before that component is reused by downstream consumers. We study when such qualification authorizes safe substitution after composition.

Consider a shared image encoder qualified by the complete output of its declared zero-shot classification task. The integrator fixes the class set and task head before receiving a replacement, and unit normalization is part of the returned representation interface. An unchanged retriever later consumes the same embedding. A malicious supplier can preserve every checked value while changing directions that the retriever still uses. The candidate passes on the same execution where the retriever promotes an attacker-selected item. Figure 1 shows how check-equivalent component states can induce different downstream behavior after composition.

Prior work establishes programmable and transferable encoder backdoors, malicious model reuse, and deployment-interface exposure (Ji et al., 2019; Jia et al., 2022; Wang et al., 2025a;b; Briglia et al., 2026; Xiang et al., 2026). We study the qualification that authorizes replacement as the exact attack constraint. Our central insight is that qualification identifies an equivalence class of implementations. Its reachable deployment states form an *audit fiber*. Movements inside a fiber are ghosts to the checker but remain operational after composition. Any downstream distinction that varies within this fiber remains available to the supplier but invisible to the checker.

Turning this insight into one malicious component poses three challenges. Direct target steering changes checked outputs. Arbitrary fiber motion provides little targeted control. Per-state control must also compile into one fixed replacement.

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.

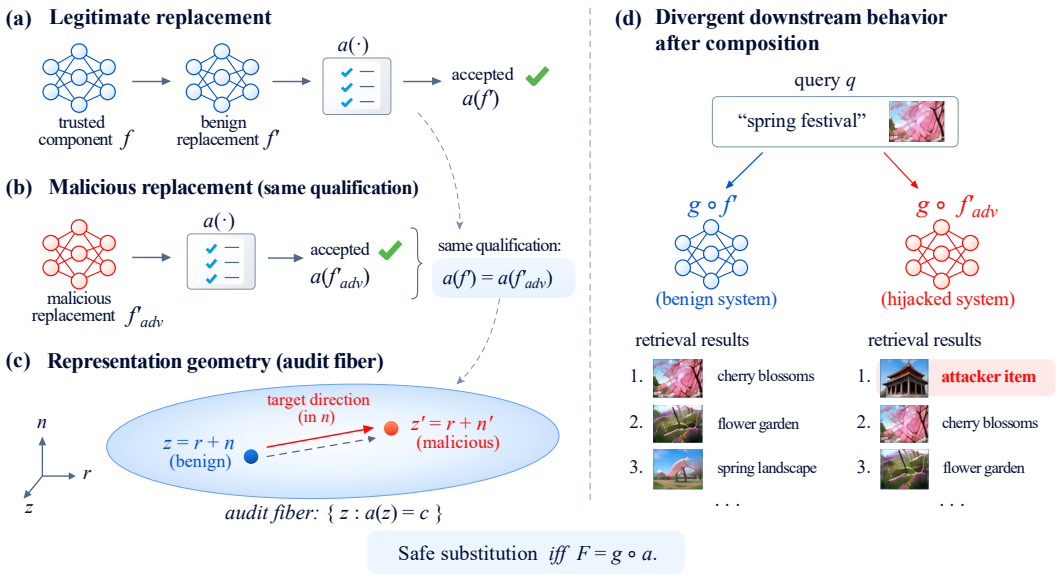

Figure 1: **Passing the same check can yield different system behavior.** The replacement preserves the declared task output and norm and is accepted. An unchanged retriever still consumes feature directions not measured by the check and promotes an attacker-selected item. For unrestricted state substitutions, safety is equivalent to factorization through the check.

We cast targeted hijacking as optimization within an audit fiber. For the linear readout-and-norm check used in our main construction, the linear target score has an exact geodesic optimum. For nonlinear consumers or checks, we use iterative optimization that preserves qualification. FiberRouter compiles the exact construction into one serialized policy selected by the unchanged checked output. We further realize the same mechanism as fixed-weight substitutions and KV-cache attacks.

Exploitability further depends on whether hidden directions are executable and downstream-relevant. We characterize safe substitution as constancy on each realizable audit fiber and quantify exploitability by composing qualification-preserving implementation reachability with downstream sensitivity. This analysis connects all-input CLIP weight substitution with replay-preserving KV-state substitution at individual handoffs. Factorization through the checked interface eliminates unrestricted downstream control.

Experiments across vision and text demonstrate reliable downstream hijacking by fixed replacements that preserve the complete declared task output. The attack remains effective on inputs withheld until after artifact commitment and under deployment shift. KV-cache experiments reveal the same separation over time, preserving the checked replay while redirecting the first unaudited prediction.

Our main contributions are summarized as follows:

- **Security problem.** We introduce adversarial qualified substitution, in which a replacement preserves every checked output on the attacked execution while redirecting an unchanged downstream consumer.
- **Theoretical analysis.** We characterize safe substitution by constancy on realizable audit fibers and derive the largest first-order downstream effect under an implementation budget. The analysis gives conditions for shared control across consumers and quantifies how stronger qualification restricts that control.
- **Audit-preserving constructions.** We derive exact audit-preserving steering, compile it into FiberRouter, realize an all-input weight substitution, and construct exact finite-replay KV-safe attack coordinates, with iterative extensions for nonlinear settings.
- **System-level evidence.** We demonstrate downstream hijacking through accepted model replacements and state handoffs across vision, text, and language models, including hidden qualification inputs and deployment shift.

2 ADVERSARIAL QUALIFIED SUBSTITUTION

A deployed component participates in two interfaces. During qualification, the integrator observes a declared behavior and decides whether a replacement is acceptable. After deployment, the returned state is reused by downstream consumers that may depend on information outside that declared view. Let $E_x(\theta) = h_\theta(x)$ be the state returned by implementation θ on input x, and let $\mathcal{Q}(\theta)$ denote the complete qualification transcript. A benign replacement θ_0 is admissible once the operational gate accepts $\mathcal{Q}(\theta_0)$. Our attack substitutes a second implementation θ_1 that preserves this transcript while changing the behavior of a downstream consumer F.

Representative deployment scenario. Consider a visual-search platform whose normalized image encoder serves both a fixed product taxonomy and nearest-neighbor retrieval against stored catalog embeddings. To reduce inference cost, the integrator may admit an FP16 or 8-bit replacement after checking the declared classification task while retaining the existing gallery and retriever. This combines documented accuracy-aware model optimization with retained-gallery reuse in compatible representation systems (Intel; Shen et al., 2020; Hu et al., 2022). The supplier submits the replacement encoder; the taxonomy head, gallery, and retriever remain fixed by the integrator. Qualification therefore observes the declared classification behavior, whereas deployment reuses the returned embedding for retrieval.

Threat model. Let Θ be the supplier's admissible component class. Before submission, the integrator fixes the qualification interface, declared task or replay schedule, downstream consumer, and remaining system components. The supplier controls the replacement artifact and knows the qualification interface; qualification inputs may remain private. Reusable components are committed once and applied across inputs. In a KV handoff, the replacement artifact is the cache for an individual prompt, and qualification evaluates its declared replay. The qualification interface is fixed before artifact commitment.

Attack objective. An *adversarial qualified substitution* seeks an attacker-chosen target t while preserving the authorizing transcript:

$$\mathcal{Q}(\theta_1) = \mathcal{Q}(\theta_0), \qquad \operatorname{pred} F(E_x(\theta_1)) = t. \tag{1}$$

The first constraint captures what the qualifier can observe; the second captures what the composed system ultimately does. The attack surface appears when these two views distinguish states differently.

3 AUDIT-PRESERVING ATTACK METHOD

Qualification exposes only part of the returned state. For a deployment state z and checked interface a, every state

$$\mathcal{F}_a(z) = \{z' : a(z') = a(z)\} \tag{2}$$

is indistinguishable to the qualifier. We call this set the *audit fiber* through z. A targeted attack therefore chooses a state inside the same fiber that favors the attacker target under a perturbation budget:

$$z^\star \in \arg \max_{z' \in \mathcal{F}_a(z),\, d(z,z') \le B} \ell_t(F(z')). \tag{3}$$

The constraint is determined entirely by qualification; the objective is determined by the downstream consumer. At regular points, admissible local directions lie in $\ker J_a(z)$. Nonlinear interfaces can therefore be handled by projected updates with retraction, whereas our main normalized linear interface admits an exact solution. Figure 2 summarizes this attack geometry.

Exact steering for normalized representations. Our main construction uses a normalized representation and a declared linear readout,

$$a_A(z) = (Az, \|z\|_2). \tag{4}$$

Let $P = P_{\ker A}$, $r = (I - P)z$, and $n = Pz$. The checked output fixes the row-space component r, while norm preservation fixes the radius $\|n\|_2$. Qualification therefore leaves the orientation of the null-space component free. The resulting fiber is a sphere inside $\ker A$.

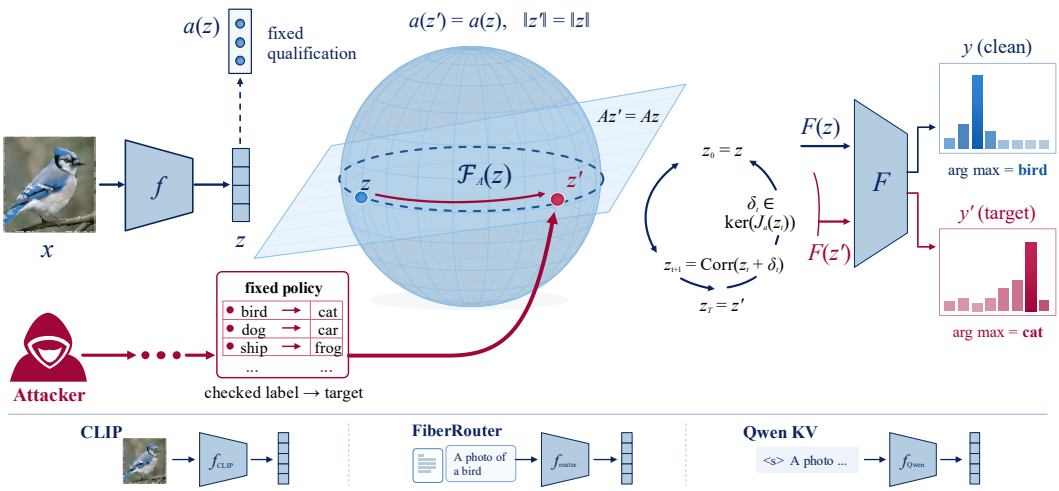

Figure 2: **Targeted optimization within an audit fiber.** Qualification defines an audit fiber of states with the same checked output. We optimize downstream behavior within this set: exactly for the linear readout-and-norm check, and iteratively for nonlinear consumers or checks. FiberRouter compiles the exact solution into a fixed policy selected by the unchanged checked output.

Theorem 3.1 (Normalized fiber and target optimum)**.** *The complete fiber is*

$$\mathcal{F}_A(z) = r + \{n' \in \ker A : \|n'\|_2 = \|n\|_2\}. \tag{5}$$

For $n \neq 0$, $\dim \ker A \geq 2$, $p = Pg \neq 0$, and $\theta \in [0, \pi]$, the maximum linear target score within null-sphere angle θ is

$$g^\top r + \|n\|_2 \|p\|_2 \cos((\omega - \theta)_+), \qquad \omega = \arccos \frac{n^\top p}{\|n\|_2 \|p\|_2}. \tag{6}$$

It is attained by the shortest great-circle rotation toward p.

The theorem gives the exact transformation $\mathrm{Steer}_A(z, g; \theta)$ used by our static attacks. It rotates only the hidden component n toward the projection of target direction g while leaving the checked component and norm unchanged. The complete edge cases are given in Appendix C. For affine downstream scores, the same path yields an exact sinusoidal score law and target-region certificate (Appendix A).

The remaining problem is to turn this pointwise steering rule into one fixed malicious component. Let κ map the checked output to a finite branch index. FiberRouter uses this unchanged branch index to select a stored target, so the routing signal is already available inside the qualified interface.

Algorithm 1 FiberRouter compiles audit-fiber steering into a fixed replacement

Input: check a_A, route map κ, target bank $G = \{g_j\}$, calibration states $\mathcal{D}$, budget θ
Output: frozen malicious component T
  Compile: partition $\mathcal{D}$ into route sets $\mathcal{D}_k = \{z \in \mathcal{D} : \kappa(a_A(z)) = k\}$
  for each route k and candidate target g_j **do**
    for each $z \in \mathcal{D}_k$ **do**
      $\widehat{z}_j \leftarrow \mathrm{Steer}_A(z, g_j; \theta)$
    end for
    Aggregate the target utility of $\{\widehat{z}_j : z \in \mathcal{D}_k\}$ into w_{kj}
  end for
  $\pi \leftarrow \arg \max_{\pi \text{ injective}} \sum_k w_{k,\pi(k)}$
  Freeze $(A, \kappa, G, \pi, \theta)$ as component map T
  Deploy: $k \leftarrow \kappa(a_A(z))$ and $j \leftarrow \pi(k)$
  $z' \leftarrow \mathrm{Steer}_A(z, g_j; \theta)$
  return z'

Compilation selects the route–target map on calibration states. Deployment applies the frozen map to each new input, with the checked output selecting the branch and steering increasing its assigned target score within the same fiber.

From routed steering to unconditional substitution. The same hidden freedom can be embedded directly into component parameters.

Proposition 3.2 (Global all-input replacement family)**.** *Let $A \in \mathbb{R}^{r\times d}$ have rank r and $q = d - r$. The orthogonal stabilizer*

$$\mathcal{G}_A = \{R \in O(d) : AR = A\} \cong O(q) \tag{7}$$

has manifold dimension $q(q-1)/2$ and preserves Az and $\|z\|_2$ for every input. If a component ends in a linear projection W followed by ℓ_2 normalization, replacing W by RW realizes this family as fixed all-input parameter substitutions.

Thus audit-preserving state freedom can be realized either as checked-state routing or as one unconditional parameter substitution. Our experiments instantiate both forms.

Iterative extensions. When closed-form steering is unavailable, FiberSteer searches downstream-sensitive directions inside the local check null space. Nonlinear checks use relinearized predictor–corrector tracking of the qualification level set. Both are numerical realizations of Equation (3); details are in Appendix B and Appendix C.8.

3.1 KV-CACHE HANDOFF

A finite replay check creates the same structure in a stateful component. During replay, each attention query reads the prompt value cache through an attention-weight row. Any prompt-value perturbation annihilated by all audited read rows is invisible to that replay transcript. Stacking these rows therefore gives a certified audit-preserving subspace that can be optimized against the first unaudited prediction.

Proposition 3.3 (Exact finite-replay KV-safe subspace)**.** *For a prompt value-cache block $V_{\ell k} \in \mathbb{R}^{P_{\mathrm{tok}}\times d_h}$, stack the prompt-attention rows from every query head sharing KV head k over h clean replay steps into $A^{(h)}_{\ell k}$. Holding keys and non-prompt cache rows fixed, any perturbation satisfying $A^{(h)}_{\ell k}\Delta V_{\ell k} = 0$ for every layer and KV head leaves all audited hidden states and full-vocabulary logits unchanged. Hence*

$$K_h^{\mathrm{sep}} = \bigoplus_{\ell,k} \ker A^{(h)}_{\ell k} \otimes \mathbb{R}^{d_h} \tag{8}$$

is a certified replay-safe subspace with

$$\dim K_h^{\mathrm{sep}} \geq L\, n_{\mathrm{KV}}\, d_h\, (P_{\mathrm{tok}} - hg)_+, \tag{9}$$

where L is the number of layers, n_{KV} the number of KV heads, and g the maximum number of query heads sharing one KV head. Moreover, $K^{\mathrm{sep}}_{h+1} \subseteq K^{\mathrm{sep}}_h$.

This is the stateful analogue of the static audit fiber: replay fixes the directions already read by the checker and leaves a nested residual subspace for future control.

4 FROM AUDIT FIBERS TO DOWNSTREAM CONTROL

The state-level fiber describes freedom that qualification leaves unseen. An attack can exploit that freedom only when some admissible replacement implementation can reach it and the downstream consumer remains sensitive to the resulting state change. We therefore lift the audit fiber from states to implementations.

For a reference implementation θ_0, let $z_0 = E_x(\theta_0)$ and define the transcript-equivalent implementation set together with its image at deployment input x:

$$\mathcal{I}_{\mathcal{Q}}(\theta_0) = \{\theta : \mathcal{Q}(\theta) = \mathcal{Q}(\theta_0)\}, \qquad \mathcal{F}_{\mathcal{Q},x}(\theta_0) = E_x(\mathcal{I}_{\mathcal{Q}}(\theta_0)). \tag{10}$$

We call $\mathcal{F}_{\mathcal{Q},x}(\theta_0)$ the *realizable audit fiber*. It contains exactly the returned states reachable by implementations that are indistinguishable to qualification. Under the usual constant-rank assumptions,

$$T_{\theta_0}\mathcal{I}_{\mathcal{Q}} = \ker J_{\mathcal{Q}}(\theta_0), \qquad T_{z_0}\mathcal{F}_{\mathcal{Q},x} = J_{E_x}(\theta_0)\ker J_{\mathcal{Q}}(\theta_0). \tag{11}$$

The second relation identifies the hidden state directions that a real replacement can actually produce.

Theorem 4.1 (Audit-fiber composition criterion)**.** *Transcript-preserving substitution leaves every intended downstream consumer unchanged exactly when each consumer is constant on every realizable audit fiber $\mathcal{F}_{\mathcal{Q},x}(\theta_0)$. In the unrestricted state-space setting, this is equivalent to every declared consumer factoring through the checked interface:*

$$F = g_F \circ a \qquad \forall F \in \mathcal{C}. \tag{12}$$

The theorem identifies when the downstream attack surface disappears. Whenever a realizable audit fiber contains two states separated by some consumer, a qualification-preserving substitution can change composed behavior.

4.1 Reachable downstream control

The global criterion above is binary; the remaining question is how much downstream effect a reachable hidden direction can produce. Locally, qualification-preserving implementation changes lie in $\ker J_{\mathcal{Q}}(\theta_0)$. Let $N_{\mathcal{Q}}$ span this kernel and let $M_\theta \succ 0$ define the implementation budget. Whiten these coordinates by

$$A_{\mathcal{Q},x} = J_{E_x}(\theta_0) N_{\mathcal{Q}} \big(N_{\mathcal{Q}}^\top M_\theta N_{\mathcal{Q}}\big)^{-1/2}. \tag{13}$$

For a smooth consumer, choose an output metric $H_F \succeq 0$ and write $D_F = H_F^{1/2} J_F(z_0)$ for the corresponding local downstream sensitivity. The product $D_F A_{\mathcal{Q},x}$ is the local transmission from hidden implementation freedom to downstream behavior.

Theorem 4.2 (Reachable downstream effect)**.** *Under the constant-rank assumptions above, the largest first-order downstream effect achievable by a qualification-preserving implementation change of budget B is*

$$\max_{\substack{J_{\mathcal{Q}}\Delta\theta = 0 \\ \Delta\theta^\top M_\theta \Delta\theta \le B^2}} \|D_F J_{E_x} \Delta\theta\|_2^2 = B^2 \sigma_{\max}^2(D_F A_{\mathcal{Q},x}). \tag{14}$$

First-order downstream control vanishes exactly when $D_F A_{\mathcal{Q},x} = 0$.

This expression exposes the three ingredients of a successful qualified substitution: qualification must leave hidden implementation freedom, the component class must transmit that freedom to the returned state, and the consumer must read the resulting direction. It also explains why downstream-aligned qualification probes remove attack yield much faster than equal-rank random probes in Figure 4d.

For a finite consumer family, common first-order control exists exactly when the projected target gradients exclude the origin from their convex hull (Theorem A.2 in Appendix A). For affine consumers, the same operator gives the minimum aligned probes required to remove the remaining linear control (Corollary A.3). These results connect consumer transfer and qualification breadth to the same transmission geometry.

5 Experiments

5.1 Experimental setup

Models and datasets. We evaluate representation and KV-cache handoffs across vision, text, and language models. Vision uses CLIP ViT-L/14 and SigLIP SO400M on ImageNet100/ImageNet-1K; text uses BGE-M3 and Qwen3-Embedding on AG News/IMDB; and KV experiments use Qwen2.5-0.5B, 1.5B, and 7B on AG News/SQuAD (Radford et al., 2021; Zhai et al., 2023; Chen et al., 2024; Zhang et al., 2025; Qwen et al., 2025; Rajpurkar et al., 2016). Pretrained backbones are frozen. Main cohorts use 10,000/5,000 images, an 8,000/4,000/1,000 text checker/gallery/query split, and 32 KV prompts per workload.

Protocol and evaluation. Representation qualification checks the complete task output and feature norm; KV qualification checks full-vocabulary logits over the first h teacher-forced steps. We target unchanged downstream classifiers, retrievers, or the first unaudited token and report target Top-1/Hit@k plus qualification drift. We vary downstream knowledge from K0 (qualification only), through K1 (consumer family) and K2 (proxy consumers), to K3 (the deployed consumer). Controls use matched random motion in the same audit fiber or certified KV-safe subspace, with graph-, rank-, and budget-matched ablations. Full details are in Appendix D.

Table 1: **Main representation-handoff results with matched baselines.** The flagship CLIP retrieval block compares benign and qualification-preserving controls against FiberRouter at the same 25° budget. The lower block reports target-class ranks for vision classifiers and target-document ranks for text retrievers. Free is audit-fiber nullity divided by representation dimension. All Max ΔQ entries are reported in units of 10^{-14}.

Method / model	Workload / task	Utility	Attack effectiveness		Qualification	
		Check acc. ↑	T@1 ↑	H@10 ↑	Free (%) ↑	Max ΔQ ↓
Flagship CLIP retrieval: matched baselines						
Benign FP16 replacement	4.5K gallery	91.66	0.16	2.02	86.98	0.00
Random exact-fiber	4.5K gallery	91.66	0.20	2.11	86.98	0.22
Feasible-random target	4.5K gallery	91.66	32.06	72.30	86.98	0.12
FiberRouter (FP16)	4.5K gallery	91.66	**82.36**	**99.00**	86.98	0.26
FiberRouter (8-bit)	4.5K gallery	91.54	**82.38**	**99.02**	86.98	0.06
Cross-model and cross-workload replication						
CLIP-L/14	ImageNet100 cls.	91.66	95.68	99.99	86.98	3.95
SigLIP SO400M	ImageNet100 cls.	94.10	86.30	99.84	91.32	4.68
BERT-base	AG News retr.	88.7	27.45	44.93	91.28	2.66
BGE-M3	AG News retr.	88.8	88.68	98.73	93.46	2.53
BGE-M3	IMDB retr.	90.1	93.28	97.28	93.65	0.80
Qwen3-Embedding	AG News retr.	90.4	82.33	96.08	93.46	2.49

5.2 A QUALIFIED REPLACEMENT IS DEPLOYABLE AND TRANSFERABLE

Downstream hijacking survives a realistic replacement gate. Intel Neural Compressor documents relative accuracy loss with 1% as a concrete tolerable-loss example (Intel). On 5,000 raw images, FP32 $\rightarrow$ FP16 loses 0.044% relatively and FP16 $\rightarrow$ 8-bit loses 0.131%; both benign replacements pass. Starting from either accepted candidate, the same fixed FiberRouter preserves the full 100-way output and representation norm to numerical precision while redirecting the unchanged 4,500-item retriever. Table 2 reports acceptance and downstream gains. On the same admissible graph, compiling the target assignment raises item Hit@1 from 32.06% under the feasible-random map to 82.36% (Table 1).

Table 2: **Deployment realism with direct paired comparisons.** Top: acceptance quality of non-identical implementation replacements. Bottom: downstream target control after qualification, reported as baseline, FiberRouter/weight-substitution result, and absolute gain. FP16, 8-bit, and hidden-input rows measure item retrieval; the weight map measures target-class ranks. Bottom-table Max ΔQ is in units of 10^{-16}.

Replacement	N	Ref. acc. ↑	Cand. acc. ↑	Δacc. (pp) ↓	Argmax agree. ↑	Med./P95 shift (°) ↓
FP32 → FP16	5,000	91.70	91.66	0.04	99.94	0.09 / 0.18
FP16 → 8-bit	5,000	91.66	91.54	0.12	98.92	5.15 / 10.01

Setting	N	Check acc. ↑	T@1 (%) ↑			H@10 (%) ↑			Max ΔQ ↓
			Base	Ours	Gain	Base	Ours	Gain	
FP16 path	5,000	91.66	0.16	**82.36**	**+82.20**	2.02	**99.00**	**+96.98**	4.72
8-bit path	5,000	91.54	0.18	**82.38**	**+82.20**	2.12	**99.02**	**+96.90**	5.55
Hidden 1K	1,000	90.30	0.10	**82.80**	**+82.70**	1.70	**99.00**	**+97.30**	4.72
Weight map	5,000	91.66	0.00	**65.44**	**+65.44**	0.30	**88.34**	**+88.04**	5.00

The frozen artifact transfers to a post-commit 1,000-image hidden challenge: checked accuracy remains 90.3%, maximum output/norm drift is 4.72×10^{-16}/6.66×10^{-16}, and target retrieval reaches 82.8% Hit@1 and 99.0% Hit@10. The serialized 1.18 MiB policy adds 0.54%/0.32% latency to the FP16/8-bit paths and 1.85 MiB peak memory.

We also realize the attack directly in model weights by left-multiplying the actual CLIP ViT-L/14 visual-projection weight by one frozen orthogonal map. This all-input replacement applies the same weight map to every image. In the 1000-way classifier, it reaches 65.44% target-class Top-1 and 88.34% Hit@10 on all 5,000 cached validation features, changing the checked output by at most 5.0×10^{-16}. Direct raw-image execution of the patched weight on one image from each audited class reaches 64% Top-1 and 84% Hit@10 with zero checked-prediction changes.

A frozen text-target FiberRouter policy also transfers to ImageNet-1K test. With its mapping unchanged, it reaches 85.94% joint-text Top-1 on 5,000 images at 30°, versus 0.02% matched random (Russakovsky et al., 2015). Under K2, proxy-built paths still reach $90.77 \pm 1.81\%$ joint-text

Top-1. Under K1, a consumer-blind direction raises held-out continuous downstream exposure by $18.94\times$ over random across all 11 construction-withheld consumers (Appendix Table 13).

5.3 Ablations isolate the source of control

Both rank and direction identity determine targeted control. Figure 4a–b contrasts directed steering with matched-random motion, while Table 3 confirms the same gap under rank matching. Figures 4d and 3 show the complementary effect of stronger qualification: target control contracts as the declared check covers more downstream-sensitive directions.

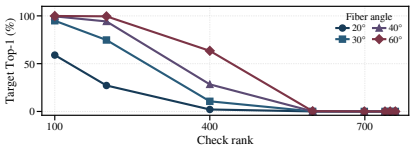

Figure 3: **Stronger checks contract target control.** Target Top-1 versus check rank at four fiber angles.

The separation is already large at low rank. With eight reusable coordinates, FiberSteer captures 29.3% of held-out sensitivity and 54.5% target yield, compared with 1.2% and 0.85% for an equal-rank random subspace. At rank 32, the gap remains $76.3\%/86.8\%$ versus $4.8\%/10.85\%$. On the qualification side, 32 aligned probes reduce residual yield to 23.8%, whereas 32 random probes leave 72.6%. Audit breadth therefore depends on covering downstream-sensitive directions rather than rank alone.

Table 3: **Rank-matched mechanism ablations.** (a) Reusable contextual attack coordinates. (b) Aligned versus random qualification probes. Percentages.

(a) Reusable attack coordinates

Rank	Sensitivity ↑		Target yield ↑	
	FiberSteer	Random	FiberSteer	Random
4	15.9	0.6	19.5	0.15
8	29.3	1.2	54.5	0.85
16	49.5	2.3	83.1	3.40
32	76.3	4.8	86.8	10.85

(b) Qualification coverage

Rank	Sensitivity covered ↑		Residual yield ↓	
	Aligned	Random	Aligned	Random
0	0.0	0.0	73.9	73.9
8	29.3	1.2	59.5	73.3
16	49.5	2.3	44.7	72.9
32	76.3	4.8	23.8	72.6

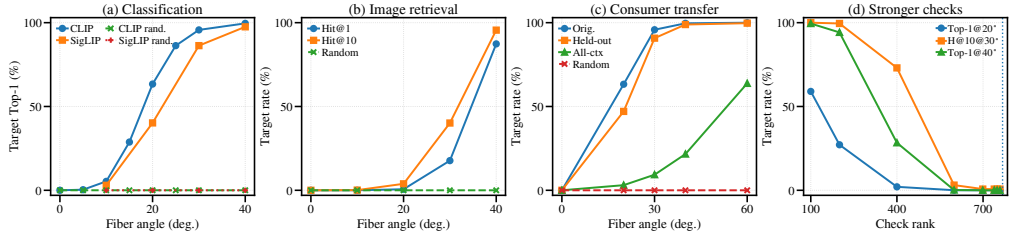

Figure 4: **Targeted control transfers across consumers and contracts as qualification strengthens.** (a–b) Directed steering dominates matched random motion; (c) frozen policies transfer across consumers; (d) stronger checks remove target control.

For consumer transfer, eight frozen FiberRouter target maps are each evaluated in both text readouts. At 30°, mean target Top-1 is 95.79% in the construction readout and 90.76% in the held-out readout, with 90.48% succeeding jointly on the same queries. At 60°, all-context Hit@10 reaches 63.78% versus 0.001% for matched random motion. Thus control follows the shared hidden representation rather than a consumer-specific output rule.

5.4 The gap persists in transferred KV state

vLLM-style disaggregated serving transfers KV caches across the prefill/decode boundary (vLLM Project). On Qwen2.5-7B, we qualify the first h clean continuation steps by full-vocabulary logits and optimize only certified-safe prompt-value directions for the first unaudited prediction. At $h = 5$, the 7B model reaches 78.13% Top-1 and 87.50% Hit@10 at rank floor 100. The 1.5B difficulty sweep reaches 65.63% Top-1 and 84.38% Hit@10 at rank floor 1000 (Table 4).

Table 4: **Deterministic KV-cache control across replay, scale, workload, and target difficulty.** All rows use 32 prompts and a 20% cache cap. Matched-random Top-1/Hit@10 is 0/0 throughout, and replay drift remains below 10^{-4}.

Model	Workload	h	Rank floor	Top-1 ↑	Hit@10 ↑	Safe coord. (%)
Qwen2.5-0.5B	AG News	5	100	62.50	87.50	51.51
Qwen2.5-1.5B	AG News	5	10	93.75	100.00	58.52
Qwen2.5-1.5B	AG News	5	100	90.63	90.63	58.52
Qwen2.5-1.5B	AG News	5	1000	65.63	84.38	58.52
Qwen2.5-7B	AG News	1	100	96.88	96.88	90.31
Qwen2.5-7B	AG News	3	100	100.00	100.00	70.95
Qwen2.5-7B	AG News	5	100	78.13	87.50	51.62
Qwen2.5-7B	SQuAD	5	100	93.75	100.00	77.45

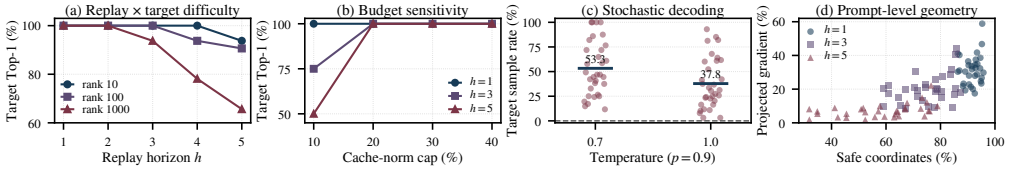

Figure 5: **KV-cache control beyond the deterministic summary.** (a) replay depth interacts with target difficulty; (b) an independent cache-budget sweep; (c) prompt-level target sampling under stochastic decoding, with horizontal bars marking means; (d) certified-safe coordinates versus projected target-gradient exposure.

Figure 5 shows how replay depth and cache budget affect control and how stochastic target sampling varies across prompts. With the endpoint and gradient fixed, deeper replay reduces projected target exposure. Appendix Tables 18 and 20 give the full matrices.

6 Related Work

Malicious reuse and qualification. Pretrained encoders, contrastive VLMs, and merged components can carry programmable or transferable backdoors into downstream reuse (Ji et al., 2019; Jia et al., 2022; Carlini & Terzis, 2021; Bai et al., 2024; Liang et al., 2024; Wang et al., 2025a;b; Briglia et al., 2026). SteerEdit compiles activation steering into trigger-dependent weight edits and uses null-space constraints to preserve clean-input behavior (Yin et al., 2026). Our substitution preserves the entire declared output on the attacked execution itself, while an unchanged consumer reads distinctions outside that interface. This ties the attack constraint to the qualification that authorizes reuse. Related repair and representation-editing studies include CleanCLIP, Mudjacking, AdvCLIP, DIFE, and AlphaEdit (Bansal et al., 2023; Liu et al., 2024; Zhou et al., 2023b; Liu & Zhang, 2025; Xiang et al., 2026; Fang et al., 2025). Differential/equivalence verification and fingerprinting compare transformed or deployed models with references (Paulsen et al., 2020; Zhang et al., 2023; Wang et al., 2024; Mo & Xiang, 2026; He et al., 2019; Wang et al., 2023; Xiaofan et al., 2024; Richter et al., 2025).

Compatibility and composition. Compatible-representation learning supports model updates with retained embeddings (Srivastava et al., 2020; Shen et al., 2020; Ramanujan et al., 2022; Pan et al., 2023; Zhou et al., 2023a; Cui et al., 2024; Jaeckle et al., 2023; Jang & Lim, 2025). Assume-guarantee reasoning studies compositional contracts (Pasareanu et al., 2018), while KV-cache security studies reuse hijacking, serving threats, and cache corruption (Zhang et al., 2026; Wu et al., 2026; Hossain et al., 2025). Our focus is the information retained by the checked component interface; complete-function parameter fibers assume stronger equality (Shewchuk & Bhattacharya, 2024).

7 Conclusion

Behavioral qualification can fail under composition when transcript-equivalent replacements remain downstream-distinguishable; our results establish this for representation and KV-cache handoffs.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used to assist with developing code and interpreting empirical results. They were also used for literature triage, manuscript drafting and editing, figure and layout iteration, and presentation of results. All reported numerical results come from executed code and stored experiment artifacts. AI-assisted claims, derivations, code changes, and prose were reviewed and checked by the authors. The authors take responsibility for the final content of the paper and accompanying artifacts.

REFERENCES

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24239–24250, 2024.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. CleanCLIP: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–123, 2023.

Maria Rosaria Briglia, Igor Maljkovic, Antonio Emanuele Cinà, Luca Oneto, Iacopo Masi, and Fabio Roli. Architectural backdoors in vision-language model supply chains via representation steering, 2026.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. 2021.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16614–16623, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zecheng He, Tianwei Zhang, and Ruby Lee. Sensitive-sample fingerprinting of deep neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4732. IEEE, 2019.

Elias Hossain, Swayamjit Saha, Somshubhra Roy, and Ravi Prasad. Can transformer memory be corrupted? investigating cache-side vulnerabilities in large language models, 2025.

Weihua Hu, Rajas Bansal, Kaidi Cao, Nikhil Rao, Karthik Subbian, and Jure Leskovec. Learning backward compatible embeddings. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3018–3028, 2022.

Hugging Face. Export a model to ONNX with `optimum.exporters.onnx`. Optimum ONNX Documentation, a. URL `https://huggingface.co/docs/optimum-onnx/onnx/usage_guides/export_a_model`. Accessed September 16, 2026.

Hugging Face. Export functions. Optimum ONNX Documentation, b. URL `https://huggingface.co/docs/optimum-onnx/onnx/package_reference/export`. Accessed September 16, 2026.

Intel. Tuning strategies. Intel Neural Compressor Documentation. URL https://intel.github.io/neural-compressor/latest/docs/source/tuning_strategies.html. Accessed September 15, 2026.

Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. 2023. URL https://arxiv.org/abs/2303.04766.

Young Kyun Jang and Ser-nam Lim. Towards cross-modal backward-compatible representation learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1783–1792, 2025.

Yu Ji, Zixin Liu, Xing Hu, Peiqi Wang, and Youhui Zhang. Programmable neural network trojan for pre-trained feature extractor, 2019.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059. IEEE, 2022.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. BadCLIP: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24645–24654, 2024.

Hongbin Liu, Michael K. Reiter, and Neil Zhenqiang Gong. Mudjacking: Patching backdoor vulnerabilities in foundation models. In *33rd USENIX Security Symposium*, pp. 2919–2936. USENIX Association, 2024.

Zhaoyi Liu and Huan Zhang. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25060–25070. IEEE, 2025.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Zihao Mo and Weiming Xiang. Eqbab: Efficient equivalence verification for compressed dnns with bound propagation. *Neurocomputing*, 674:132791, 2026. doi: 10.1016/j.neucom.2026.132791.

Tan Pan, Furong Xu, Xudong Yang, Sifeng He, Chen Jiang, Qingpei Guo, Feng Qian, Xiaobo Zhang, Yuan Cheng, Lei Yang, et al. Boundary-aware backward-compatible representation via adversarial learning in image retrieval. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15201–15210. IEEE, 2023.

Corina S Pasareanu, Divya Gopinath, and Huafeng Yu. Compositional verification for autonomous systems with deep learning components, 2018.

Brandon Paulsen, Jingbo Wang, and Chao Wang. Reludiff: Differential verification of deep neural networks. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pp. 714–726, 2020. doi: 10.1145/3377811.3380337.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. 2025. URL https://arxiv.org/abs/2412.15115.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Vivek Ramanujan, Pavan Kumar Anasosalu Vasu, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Forward compatible training for large-scale embedding retrieval systems. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19364–19373. IEEE, 2022.

Leo Richter, Xuanli He, Pasquale Minervini, and Matt J. Kusner. An auditing test to detect behavioral shift in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6367–6376. IEEE, 2020.

Jonathan Richard Shewchuk and Sagnik Bhattacharya. The geometry of the set of equivalent linear neural networks, 2024.

Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An empirical analysis of backward compatibility in machine learning systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3272–3280, 2020. doi: 10.1145/3394486.3403379.

vLLM Project. Disaggregated prefilling (experimental). vLLM Documentation. URL `https://docs.vllm.ai/en/latest/features/disagg_prefill/`. Accessed September 15, 2026.

Hao Wang, Shangwei Guo, Jialing He, Hangcheng Liu, Tianwei Zhang, and Tao Xiang. Model supply chain poisoning: Backdooring pre-trained models via embedding indistinguishability. In *Proceedings of the ACM on Web Conference 2025*, pp. 840–851, 2025a.

Lijin Wang, Jingjing Wang, Tianshuo Cong, Xinlei He, Zhan Qin, and Xinyi Huang. From purity to peril: Backdooring merged models from Harmless benign components. In *34th USENIX Security Symposium*, pp. 6339–6358. USENIX Association, 2025b.

Shuo Wang, Sharif Abuadbba, Sidharth Agarwal, Kristen Moore, Ruoxi Sun, Minhui Xue, Surya Nepal, Seyit Camtepe, and Salil Kanhere. Publiccheck: Public integrity verification for services of run-time deep models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1348–1365. IEEE, 2023.

Weiru Wang, Kai Wang, Zhiyang Cheng, and Yilong Yang. Veriprune: Equivalence verification of node pruned neural network. *Neurocomputing*, 577:127347, 2024. doi: 10.1016/j.neucom.2024.127347.

XiangFan Wu, Lingyun Ying, Guoqiang Chen, Yacong Gu, and Haipeng Qu. Cache me, catch you: Cache related security threats in LLM serving frameworks. In *Network and Distributed System Security Symposium (NDSS)*, 2026.

Kunlan Xiang, Haomiao Yang, and Wenbo Jiang. Beyond native success: Auditing deployment-interface exposure of clip backdoors, 2026.

Bai Xiaofan, Chaoxiang He, Xiaojing Ma, Bin Benjamin Zhu, and Hai Jin. Intersecting-boundary-sensitive fingerprinting for tampering detection of dnn models. In *Forty-first International Conference on Machine Learning*, 2024.

Rui Yin, Tianxu Han, Naen Xu, Changjiang Li, Ping He, Chunyi Zhou, Jun Wang, Zhihui Fu, Tianyu Du, Jinbao Li, et al. Compiling activation steering into weights via null-space constraints for stealthy backdoors. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26228–26245, 2026.

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.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yedi Zhang, Fu Song, and Jun Sun. Qebverif: Quantization error bound verification of neural networks. In *Computer Aided Verification*, pp. 413–437. Springer, 2023. doi: 10.1007/978-3-031-37703-7_20.

Yichi Zhang, Zhiqi Wang, Huan Zhang, and Yuchen Yang. Hijackkv: New threat in position-independent kv cache reuse. 2026.

Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. Bt^ 2: Backward-compatible training with basis transformation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11195–11204. IEEE, 2023a.

Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6311–6320, 2023b.

A Proofs and Core Geometry

This appendix retains only the proofs and constructions used by the main text.

A.1 Composition boundary and reachable control

Proof of Theorem 4.1.

Proof. Fix a reference implementation θ_0 and input x. Every implementation with the same qualification transcript lies in $\mathcal{I}_{\mathcal{Q}}(\theta_0)$, and its returned state lies in $\mathcal{F}_{\mathcal{Q},x}(\theta_0)$. Thus substitution is safe relative to the implementation class exactly when every intended consumer is constant on every realizable audit fiber.

In the unrestricted state-space setting, every pair with $a(z) = a(z')$ is an admissible same-interface substitution. If qualification is sufficient, define $g_F(a(z)) := F(z)$ on the attained image of a. This is well-defined because states in one audit fiber have equal consumer outputs, and therefore $F = g_F \circ a$. Conversely, if $F = g_F \circ a$ for every declared consumer, then equality under a implies equality under every F. □

Proof of Theorem 4.2.

Proof. Write $N = N_{\mathcal{Q}}$ and $M_Q = N^\top M_\theta N$. Every first-order qualification-preserving perturbation has the form $\Delta\theta = N\eta$ with $\eta^\top M_Q \eta \le B^2$. Setting $u = M_Q^{1/2}\eta$ gives $\|u\|_2 \le B$ and

$$J_{E_x}\Delta\theta = J_{E_x} N M_Q^{-1/2} u = A_{\mathcal{Q},x} u.$$

Hence

$$\max_{\|u\|_2 \le B} \|D_F A_{\mathcal{Q},x} u\|_2^2 = B^2 \sigma_{\max}^2(D_F A_{\mathcal{Q},x}),$$

by the variational definition of the spectral norm. The optimum is zero exactly when $D_F A_{\mathcal{Q},x} = 0$. □

Corollary A.1 (Family transmission spectrum)**.** *Let*

$$S_\mu(z_0) = \mathbb{E}_{F\sim\mu}\left[J_F(z_0)^\top H_F J_F(z_0)\right], \qquad \mathsf{X}_{\mu,\mathcal{Q},x} = A_{\mathcal{Q},x}^\top S_\mu(z_0) A_{\mathcal{Q},x}. \tag{15}$$

If $\lambda_1 \ge \cdots \ge \lambda_q > 0$ are the nonzero eigenvalues of $\mathsf{X}_{\mu,\mathcal{Q},x}$, the maximum expected squared downstream effect under hidden-coordinate budget B is $B^2\lambda_1$. More generally, the best s-dimensional orthonormal attack library captures $\sum_{i\le s}\lambda_i$ total family sensitivity.

Proof. For a whitened hidden-coordinate perturbation u,

$$\mathbb{E}_F\|H_F^{1/2} J_F A_{\mathcal{Q},x} u\|_2^2 = u^\top \mathsf{X}_{\mu,\mathcal{Q},x} u.$$

Rayleigh–Ritz gives the rank-one optimum $B^2\lambda_1$. For an s-dimensional library with projector P_U, captured mean-square family sensitivity is $\mathrm{tr}(P_U \mathsf{X}_{\mu,\mathcal{Q},x})$, whose maximum is $\sum_{i\le s}\lambda_i$ by the Ky Fan variational principle. □

The consumer-blind experiment estimates high-exposure directions without using the 11 evaluation consumers. Its $18.94\times$ held-out CNBD gain over random is therefore an empirical instance of concentration in this family transmission spectrum rather than a target-specific construction.

Theorem A.2 (Common control under consumer uncertainty)**.** *For target scores s_c of a finite consumer family, define $g_c = A_{\mathcal{Q},x}^\top \nabla s_c(z_0)$ in whitened hidden implementation coordinates. Then*

$$\max_{\|u\|_2\le B} \min_{c\in\mathcal{C}} \langle g_c, u\rangle = B\ \mathrm{dist}(0, \mathrm{conv}\{g_c : c\in\mathcal{C}\}). \tag{16}$$

Thus a single qualification-preserving direction has positive first-order gain for every consumer exactly when the origin lies outside the convex hull of the projected target gradients.

Proof. Let $K = \mathrm{conv}\{g_c\}$ and let $q_\star$ be the closest point of K to the origin. For any $\|u\|_2 \le B$,

$$\min_{q \in K} \langle q, u \rangle \le \langle q_\star, u \rangle \le B \|q_\star\|_2.$$

If $q_\star \neq 0$, the Euclidean projection property gives $\langle q - q_\star, q_\star \rangle \ge 0$ for every $q \in K$. Choosing $u_\star = B q_\star / \|q_\star\|_2$ attains the upper bound. If $q_\star = 0$, $u = 0$ attains zero. □

Corollary A.3 (Minimum linear completion)**.** *Let $P = P_{\ker A}$ and let $M_{\mathcal{C}} P$ stack the centered downstream linear maps restricted to the unchecked subspace. The minimum number of independent linear probes that makes every declared centered affine output constant on each remaining linear fiber is* $\mathrm{rank}(M_{\mathcal{C}} P)$*. If only s orthonormal probes may be added, the optimal residual operator norm is $\sigma_{s+1}(M_{\mathcal{C}} P)$.*

Proof. For an augmented interface $(Az, Q^\top z)$ with orthonormal columns of Q in $\ker A$, the residual projector is $P - QQ^\top$. All centered affine consumers are constant on every remaining fiber exactly when

$$M_{\mathcal{C}}(P - QQ^\top) = 0,$$

equivalently $\mathrm{row}(M_{\mathcal{C}} P) \subseteq \mathrm{span}(Q)$. Thus at least $\mathrm{rank}(M_{\mathcal{C}} P)$ independent probes are necessary, and a basis of that row space attains the bound. With only s probes, choosing the leading right singular vectors minimizes the residual operator norm and leaves $\sigma_{s+1}(M_{\mathcal{C}} P)$. □

A.2 Fixed policies and temporal refinement

Proposition A.4 (Check-invariant policy compilation)**.** *Let I and T be finite route and target sets, let $\kappa : a(\mathcal{Z}) \to I$, and let $E \subseteq I \times T$ be a feasibility graph. Suppose each branch $B_{i,t}$ with $(i,t) \in E$ obeys $a(B_{i,t}(z)) = a(z)$ for every state routed to i. If an injective left-saturating assignment $\pi : I \to T$ exists, then*

$$T_\pi(z) = B_{\kappa(a(z)),\, \pi(\kappa(a(z)))}(z) \tag{17}$$

preserves both the checked output and the route. With fixed nonnegative route weights w_i and edge utilities $u_{i,t}$, maximizing $\sum_i w_i u_{i,\pi(i)}$ is a maximum-weight left-perfect matching problem.

Proof. For arbitrary z, let $i = \kappa(a(z))$ and $t = \pi(i)$. Feasibility gives $(i,t) \in E$, so the branch condition yields $a(B_{i,t}(z)) = a(z)$. Therefore $a(T_\pi(z)) = a(z)$ and the post-attack route remains $\kappa(a(T_\pi(z))) = i$. The assignment objective is the sum of the selected edge weights, hence the compilation problem is exactly maximum-weight matching. □

This proposition formalizes why the attack can be compiled once and then deployed as a fixed component. Calibration chooses a finite routing table; the deployed artifact only reads the already checked state and executes the corresponding audit-preserving branch.

Theorem A.5 (Temporal refinement law)**.** *Let τ_h be a deterministic replay transcript whose first h interactions are a truncation of τ_{h+1}. Then the corresponding fibers satisfy*

$$\mathcal{F}_{h+1}(s) \subseteq \mathcal{F}_h(s). \tag{18}$$

If $K_{h+1} \subseteq K_h$ are the associated first-order legal spaces for a fixed endpoint and gradient g, then

$$\max_{\delta \in K_h,\ \|\delta\|_2 \le B} \langle g, \delta \rangle = B \|P_{K_h} g\|_2, \qquad \|P_{K_{h+1}} g\|_2 \le \|P_{K_h} g\|_2. \tag{19}$$

Proof. Because τ_h is a deterministic truncation of τ_{h+1}, equality of the longer transcript implies equality of the shorter one, proving Equation (18). For the local statement, $\langle g, \delta \rangle = \langle P_{K_h} g, \delta \rangle$ for $\delta \in K_h$; Cauchy–Schwarz gives the optimum $B\|P_{K_h} g\|_2$. Orthogonal projection onto the nested subspace K_{h+1} cannot increase this norm. □

The Qwen fixed-objective diagnostic uses exactly this construction: endpoint, target, and ambient gradient are frozen, while only the replay-safe subspace is refined. The mean projected-gradient fraction falls from 31.31% at $h = 1$ to 20.05% at $h = 3$ and 7.26% at $h = 5$, matching the geometric contraction visualized in Figure 5d.

A.3 Finite-tolerance gates and attack amplification

Real qualification often admits a small numerical discrepancy rather than exact equality. The exact audit fiber remains the zero-discrepancy slice of that larger feasible region. For linear checked and downstream interfaces, this relation has a closed form.

Proposition A.6 (Finite-tolerance contextual amplification)**.** *For $A \in \mathbb{R}^{m\times d}$ with $A \neq 0$ and downstream operator $H \in \mathbb{R}^{k\times d}$, define*

$$\Gamma(H \mid A) = \sup_{\delta \notin \ker A} \frac{\|H\delta\|_2}{\|A\delta\|_2}. \tag{20}$$

Then $\Gamma(H \mid A) < \infty$ exactly when $\ker A \subseteq \ker H$. In that case,

$$\Gamma(H \mid A) = \|HA^\dagger\|_2. \tag{21}$$

If some exact audit-hidden direction still changes the downstream output, $\Gamma(H \mid A) = \infty$.

Proof. If $v \in \ker A$ but $Hv \neq 0$, choose u with $Au \neq 0$ and consider $\delta_\epsilon = v + \epsilon u$. The denominator in Equation (20) vanishes linearly in ϵ while the numerator converges to $\|Hv\|_2$, so the ratio diverges. Conversely, $\ker A \subseteq \ker H$ implies $H = HA^\dagger A$: the residual $\delta - A^\dagger A\delta$ lies in $\ker A$ and is annihilated by H. Writing $y = A\delta$ then gives the tight operator norm $\|HA^\dagger\|_2$. □

With an explicit state budget, the finite-tolerance attack surface is

$$\Omega_{H|A}(\varepsilon, B) = \max_{\substack{\|\delta\|_2 \le B \\ \|A\delta\|_2 \le \varepsilon}} \|H\delta\|_2. \tag{22}$$

At $\varepsilon = 0$, $\Omega_{H|A}(0, B) = B\|HP_{\ker A}\|_2$. Thus every exact attack direction studied in the main paper remains feasible under any nonzero qualification tolerance; increasing the tolerance can only enlarge the feasible set.

Implementation-conditioned attack cost. Reachable support and attack cost are distinct. In a controlled residual-branch experiment, two 32-parameter implementations have the same frozen rank-100 audit operator and identical 24-dimensional reachable state support, but differ only in hidden-branch scaling. Under the same Euclidean parameter budget, the leading transmission eigenvalue changes by 9.71–$13.32\times$ across five prespecified seeds (mean $11.41\times$). Using the pullback parameter metric restores the baseline transmission spectrum to within 3.8×10^{-15} relative error. This isolates implementation geometry as a multiplier on attack cost while leaving the underlying audit-hidden state directions unchanged.

A.4 Normalized fibers and exact steering

Proof of Theorem 3.1.

Proof. Write $z = r + n$ with $r \in \mathrm{row}(A)$ and $n \in \ker A$. The equality $Az' = Az$ fixes the row-space component of z' to r. Orthogonality gives $\|z'\|_2^2 = \|r\|_2^2 + \|n'\|_2^2$, so norm preservation is equivalent to $\|n'\|_2 = \|n\|_2$, proving Equation (5).

For $q \geq 2$, write $n' = \rho u'$ with $\|u'\|_2 = 1$. The target score is $g^\top r + \rho p^\top u'$. Over an angular cap of radius θ around $u = n/\rho$, the maximum is attained at the closest point to $v = p/\|p\|_2$ along a shortest great-circle arc. This gives Equation (6); $p = 0$, $n = 0$, and antipodal cases follow by continuity and the geometry of the sphere. □

Proof of Proposition 3.2.

Proof. For $Q \in O(q)$, the map $R_Q = UU^\top + NQN^\top$ is orthogonal. Since $AN = 0$ and $AUU^\top = A$, $AR_Q = A$. Conversely, any orthogonal R satisfying $AR = A$ fixes $\mathrm{row}(A)$ pointwise and therefore acts as some $Q \in O(q)$ on $\ker A$, yielding Equation (7). Its dimension is $q(q-1)/2$. Orthogonality preserves the feature norm, and full row rank of the original projection makes the induced weight substitution injective. □

Fixed weight-level realization. For the CLIP ViT-L/14 check, $q = 668$, so the stabilizer in Proposition 3.2 has $668 \cdot 667/2 = 222{,}778$ continuous degrees of freedom. The final visual projection has full row rank, so this entire orthogonal family embeds injectively into the projection-weight implementation fiber. We fit one 668×668 orthogonal map on the 10,000-image ImageNet100 training cache by a closed-form Procrustes objective for the fixed source-to-target mapping. The result is one unconditional replacement shared by every input. On all 5,000 validation images it reaches 65.44% target Top-1 and 88.34% Hit@10, while the complete 100-way audit output changes by at most 5.0×10^{-16} and the feature norm by less than 10^{-15}. Directly patching the actual HuggingFace `visual_projection.weight` and replaying one raw image from each audited class yields 64% Top-1 and 84% Hit@10 with zero audit prediction changes.

Proposition A.7 (Cross-context transfer along one fiber path)**.** *Assume $0 < \omega < \pi$, let $u = n/\rho$, and define $t_g = (p/\|p\|_2 - \cos\omega\, u)/\sin\omega$. Along the target-optimal path $z_g(\theta) = r + \rho(\cos\theta\, u + \sin\theta\, t_g)$, any affine score $s_h(z) = h^\top z + c$ changes by*

$$s_h(z_g(\theta)) - s_h(z) = \rho[(\cos\theta - 1)\langle Ph, u\rangle + \sin\theta\langle Ph, t_g\rangle]\,. \tag{23}$$

When $Ph \neq 0$, the initial normalized transfer coefficient is $\tau_{g\to h}(z) = \langle t_g, Ph/\|Ph\|_2\rangle$.

Proof. The null component along the shortest arc is $n_g(\theta) = \rho(\cos\theta\, u + \sin\theta\, t_g)$. Because the displacement lies in $\ker A$, $h^\top(n_g(\theta) - n) = \langle Ph, n_g(\theta) - n\rangle$. Substituting $n = \rho u$ gives Equation (23). Differentiating at zero gives the stated coefficient. The same identity applies to affine class margins by taking $h = w_t - w_k$. □

Exact contextual-integrity radius. For affine logits with unique initial winner y, let $R_{\mathrm{CI}}(z)$ be the smallest null-sphere angle at which a competitor ties or exceeds y. Write $n = Pz$, $\rho = \|n\|_2$, and for each competitor $k \neq y$ define $q_k = P(w_k - w_y)$, $\beta_k = \|q_k\|_2$, and $m_k = \ell_y(z) - \ell_k(z) > 0$. If $\rho = 0$ the fiber is a singleton and $R_{\mathrm{CI}}(z) = \infty$. For $q = \dim\ker A \geq 2$ and $\beta_k > 0$, put

$$u = n/\rho, \qquad v_k = q_k/\beta_k, \qquad \omega_k = \arccos(u^\top v_k), \qquad c_k = \cos\omega_k + \frac{m_k}{\rho\beta_k}.$$

The competitor is unreachable on the fiber when $c_k > 1$. Otherwise its nearest boundary point lies on the shortest arc from u toward v_k and is at angle

$$R_k = \omega_k - \arccos(c_k). \tag{24}$$

Thus $R_{\mathrm{CI}}(z) = \min_{k:\, c_k \leq 1} R_k$, with the minimum over an empty set equal to ∞. This closed form is the decision-boundary quantity used by the reported radius experiments and avoids an angular grid search.

Local tangent and global span. For a nondegenerate normalized fiber, $T_z\mathcal{F}_A^{\mathrm{sph}} = \ker A \cap n^\perp$ has dimension $q - 1$. Globally, however, differences between fiber points span all of $\ker A$:

$$\mathrm{span}\{z' - z'' : z', z'' \in \mathcal{F}_A^{\mathrm{sph}}(z)\} = \ker A. \tag{25}$$

This distinction separates pointwise infinitesimal steering from the larger all-input replacement family above.

Affine target-region geometry. For affine logits $\ell_k(z) = w_k^\top z + b_k$ under a linear check, the minimum-norm one-component displacement that enters target class t with margin $\kappa > 0$ is the convex program

$$\begin{aligned} \min_{\delta} \quad & \tfrac{1}{2}\|\delta\|_2^2 \\ \text{s.t.} \quad & A\delta = 0, \\ & (w_t - w_k)^\top(z + \delta) + (b_t - b_k) \geq \kappa, \qquad k \neq t. \end{aligned} \tag{26}$$

With an orthonormal null basis N, substituting $\delta = Nu$ reduces the problem to the unchecked coordinates. For a single source–target boundary, let $m_{y,t}(z) = \ell_y(z) - \ell_t(z) > 0$ and $q = P_{\ker A}(w_y - w_t)$. When $q \neq 0$, the exact minimum displacement reaching that boundary is

$$\delta^\star = -\frac{m_{y,t}(z)}{\|q\|_2^2} q, \qquad \|\delta^\star\|_2 = \frac{m_{y,t}(z)}{\|q\|_2}. \tag{27}$$

The multiclass target-region distance is at least the largest active pairwise lower bound and is obtained exactly by the quadratic program above. These quantities provide a geometry-only estimate of target difficulty before running the finite-budget attack.

Proposition A.8 (Target-region certificate on the attack path)**.** *Let $\gamma_t(\theta)$ be the shortest audit-fiber geodesic toward a target vector and let downstream logits be affine. For every competitor j,*

$$\ell_t(\gamma_t(\theta)) - \ell_j(\gamma_t(\theta)) = \alpha_j + \rho(a_j \cos\theta + b_j \sin\theta). \tag{28}$$

The target-top-scoring angles therefore form a finite union of closed intervals whose boundaries are path endpoints or analytic roots of the pairwise margins.

Proof. Write $\gamma_t(\theta) = r + \rho(u\cos\theta + v\sin\theta)$. Substitution into an affine target-minus-competitor margin yields Equation (28). Each nonconstant scalar sinusoid has finitely many roots on the compact geodesic interval and constant sign between consecutive roots. Intersecting the nonnegative sign sets over competitors gives the claimed interval union and certifies the first target-region entry angle without a grid search. □

A.5 Finite-replay KV fibers

Proof of Proposition 3.3.

Proof. Fix an audited replay step and layer. By induction over earlier replay steps and earlier layers in the current step, the query entering the attention module is the clean query. Keys are unchanged, so every audited attention row is a row of $A_{\ell k}^{(h)}$. If $A_{\ell k}^{(h)} \Delta V_{\ell k} = 0$, the prompt-value contribution is unchanged for every audited query head. Residual streams, subsequent layers, and full-vocabulary logits therefore remain unchanged, closing the induction.

Each column of $\Delta V_{\ell k}$ may independently lie in $\ker A_{\ell k}^{(h)}$, giving $(P_{\mathrm{tok}} - \operatorname{rank} A_{\ell k}^{(h)})d_h$ certified degrees of freedom. Summing over cache blocks gives Equation (8). Appending replay steps can only add rows to each $A_{\ell k}^{(h)}$, hence $K_{h+1}^{\mathrm{sep}} \subseteq K_h^{\mathrm{sep}}$ and the dimension lower bound in the main text follows from $\operatorname{rank} A_{\ell k}^{(h)} \le hg$. □

B FiberSteer and Nonlinear Qualification Details

FiberSteer turns a large linear audit fiber into a reusable, low-rank attack coordinate system. Let $P = P_{\ker A}$ and let $\omega = (c, x)$ index a downstream consumer/input pair. For output metric $H_\omega \succeq 0$, define

$$M_\omega = H_\omega^{1/2} J_\omega(z) P, \qquad G_\mu = \mathbb{E}_{\omega\sim\mu}[M_\omega^\top M_\omega]. \tag{29}$$

G_μ summarizes which unchecked representation directions are repeatedly read by the downstream workload.

Theorem B.1 (Contextual attack coordinates)**.** *Let Q have s orthonormal columns in $\ker A$ and let attack coefficients u satisfy $\mathbb{E}[uu^\top \mid \omega] = I_s$. Then*

$$\mathcal{R}_\mu(Q) = \mathbb{E}_{\omega,u}\|M_\omega Q u\|_2^2 = \operatorname{tr}(Q^\top G_\mu Q). \tag{30}$$

The maximum over rank-s attack spaces is attained by a leading s-dimensional eigenspace of G_μ restricted to $\ker A$. A Haar-random rank-s subspace of an n-dimensional audit null space captures expected fraction s/n of the total trace sensitivity.

Proof. Conditioning on ω and using $\mathbb{E}[uu^\top \mid \omega] = I_s$ yields $\mathbb{E}_u\|M_\omega Q u\|_2^2 = \operatorname{tr}(Q^\top M_\omega^\top M_\omega Q)$. Averaging gives Equation (30); the leading eigenspace optimum follows from the Ky Fan variational principle. For a Haar-random Q, $\mathbb{E}[QQ^\top] = (s/n)P$ on $\ker A$, giving the stated baseline. □

The attack basis can be estimated without materializing consumer Jacobians. For an output sketch ξ with conditional mean zero and covariance identity,

$$g_\omega = P J_\omega(z)^\top H_\omega^{1/2} \xi, \qquad \mathbb{E}[g_\omega g_\omega^\top] = G_\mu. \tag{31}$$

Thus leading right singular directions of stacked $g_\omega^\top$ samples estimate the same attack space using vector–Jacobian products.

Algorithm 2 FiberSteer learns reusable coordinates inside an audit fiber

Input: Checked readout A, workload samples, attack rank s, budget B
Form $P \leftarrow P_{\ker A}$
for sampled consumer/input pair ω **do**
Draw output sketch ξ and compute $g_\omega \leftarrow PJ_\omega(z)^\top H_\omega^{1/2}\xi$
end for
Let Q_s be the leading right singular subspace of the stacked $g_\omega^\top$ sketches
Reproject Q_s into $\ker A$ and orthonormalize
For each route or source class, fit coefficients α with $\delta = Q_s\alpha$ and $\|\alpha\|_2 \le B$
return the frozen basis Q_s and coefficient table

Table 5: **Reusable fixed-vector attack codebook.** All source–target pairs share one learned rank-32 FiberSteer basis.

Steering space	Rank	Mean yield	Median yield	Pairs $\geq 80\%$
FiberSteer	32	91.9	97.0	18/20
Random null space	32	1.7	0.0	0/20
Complete null space	668	92.1	97.0	18/20

Proposition B.2 (VJP attack-subspace estimation)**.** *With N independent sketches, let $\widehat{G} = N^{-1}\sum_{r=1}^{N} g_r g_r^\top$. If $\widehat{Q}_s$ and $Q_s^\star$ are leading rank-s eigenspaces of the empirical and population operators, then*

$$\mathcal{R}_\mu(Q_s^\star) - \mathcal{R}_\mu(\widehat{Q}_s) \le 2s\|\widehat{G} - G_\mu\|_2. \tag{32}$$

Proof. Insert $\widehat{G}$ into the difference of the two population trace objectives. The empirical comparison is nonpositive by optimality of $\widehat{Q}_s$, and the two remaining perturbation terms are each bounded by $s\|\widehat{G} - G_\mu\|_2$. □

Nonlinear rank response. The main ablation uses a 64-unit tanh MLP consumer on ImageNet100 with the fixed 100-way CLIP readout as its qualification interface. Subspace estimation uses 2,000 balanced training examples and held-out sensitivity uses 2,000 disjoint validation examples. Coefficient optimization is evaluated on a fixed 1,000-example target population. At ranks $4, 8, 16, 32$, FiberSteer captures $15.9\%, 29.3\%, 49.5\%, 76.3\%$ of held-out sensitivity, compared with $0.6\%, 1.2\%, 2.3\%, 4.8\%$ for matched random subspaces. Target yield increases from 19.5% at rank 4 to 86.8% at rank 32, whereas the rank-32 random baseline reaches 10.85%.

Proposition B.3 (Shared steering in a fixed attack subspace)**.** *Let $m_{t,s}(z) = \ell_t(z) - \ell_s(z)$ and let Q have orthonormal columns in $\ker A$. Under a source-conditioned workload ν_s, define*

$$\bar{g} = \mathbb{E}_{z\sim\nu_s}\left[Q^\top \nabla m_{t,s}(z)\right]. \tag{33}$$

Among fixed coefficients with $\|\alpha\|_2 \le B$, the expected first-order margin gain is maximized by $\alpha^\star = B\bar{g}/\|\bar{g}\|_2$ when $\bar{g} \neq 0$. For affine logits the same expression gives the exact expected gain.

Proof. The expected first-order gain is $\mathbb{E}[\nabla m_{t,s}(z)^\top Q\alpha] = \bar{g}^\top\alpha$. Cauchy–Schwarz gives the optimizer. For affine logits the Taylor expansion has no higher-order term. □

This result explains the fixed-vector codebook experiment below: once FiberSteer has learned a shared unchecked basis, each source–target pair only needs one low-dimensional coefficient vector rather than a per-query optimization.

Fixed-vector codebook. To separate reusable control from per-input optimization, one rank-32 basis is shared across 20 fixed source–target pairs. Each pair fits one coefficient vector from 100 source training examples and then evaluates that vector on 50 held-out source images.

The preselected pair-0 rule reaches 86.0% targeted yield with a gate driven by the unchanged checked argmax. Source recall is 92.0%, non-source gate activation is 0.0404%, and non-source accuracy

Table 6: **FiberSteer across fitted downstream consumers.** Values are clean accuracy and target yield percentages at budget $B = 40$.

Consumer	Clean acc.	Shared yield	Projected yield
Original affine	88.7	6.8	8.5
MLP-matched affine	87.0	66.0	73.3
MLP	85.9	81.8	84.2
Gated-matched affine	88.3	87.7	87.8
Gated fusion	87.9	87.3	87.4
Residual cross-attention	86.8	85.2	85.3

Algorithm 3 Target-score-optimal steering on a normalized audit fiber

Input: Unit representation z, checked readout A, target vector g, angular budget $\theta \in [0, \pi]$
Compute an orthonormal row basis R of A and $P \leftarrow I - R^\top R$
$r \leftarrow (I - P)z$, $n \leftarrow Pz$, $p \leftarrow Pg$
if $\|n\|_2 = 0$ or $\|p\|_2 = 0$ or $\theta = 0$ **then**
  return z {singleton, constant score, or zero budget}
end if
$u \leftarrow n/\|n\|_2$, $v \leftarrow p/\|p\|_2$ and $\omega \leftarrow \arccos(\mathrm{clip}(u^\top v, -1, 1))$
if $\dim \ker A = 1$ **then**
  return the higher-scoring feasible point among z and $r - n$ (the latter is feasible only when $\theta = \pi$)
end if
if $\omega = 0$ **then**
  return z
else if $\omega = \pi$ **then**
  Choose any unit $w \in \ker A \cap u^\perp$ and set $\theta' \leftarrow \theta$
else
  $w \leftarrow (v - \cos\omega\, u)/\sin\omega$ and $\theta' \leftarrow \min\{\theta, \omega\}$
end if
return $z' \leftarrow r + \|n\|_2(\cos\theta'\, u + \sin\theta'\, w)$

changes by 0.040 percentage point. The result shows that nearly all of the full-null-space attack yield can be compiled into a reusable 32-dimensional coordinate system.

Fitted nonlinear consumers. We also fit several downstream consumers on CLIP/SigLIP representations under the same balanced source–target protocol. The attack basis is selected once; each consumer only changes the downstream geometry read from that basis.

Checked predictions are retained on all evaluated inputs, with maximum checked-logit drift below 2.5×10^{-14}. The variation across consumers tracks how strongly each downstream decision rule reads the learned unchecked directions.

Nonlinear qualification interface. For a learned qualification map, the level-set tracker in Appendix C.8 relinearizes the checker Jacobian, projects the proposed target step into the instantaneous null space, and corrects back to the original qualification level before continuing. This is the nonlinear numerical counterpart of the exact linear-fiber construction and preserves the complete checker output rather than only its predicted label.

C Native Contextual Attack Protocols

C.1 Spherical score optimization

Algorithm 3 implements the target-score optimum in Theorem 3.1. The checked readout fixes the row-space component; the angular budget limits motion on the remaining sphere. A fixed readout scale can be included in A, and a fixed bias leaves its fibers unchanged.

Table 7: **Targeted 1000-way classification under a fixed 100-way check.** Rates are percentages over 5,000 queries, averaged across eight target mappings. The last column compares each displacement with the query's nearest selected same-class training feature. Random controls use the same angles.

Angle	Target top-1	Target Hit@10	Random top-1	Mean chord	Chord $\leq$ nearest same-class
0°	0.02	0.90	0.02	0.000	–
5°	0.35	13.07	0.02	0.078	100.0
10°	5.32	58.60	0.02	0.157	100.0
15°	28.78	91.13	0.02	0.235	99.8
20°	63.38	98.73	0.02	0.312	96.34
25°	86.30	99.74	0.02	0.389	84.28
30°	95.68	99.99	0.01	0.465	63.78
40°	99.58	100.0	0.01	0.615	24.98

Table 8: **SigLIP SO400M replication.** Rates are percentages averaged across the eight target mappings. Maximum checked-dot-product and norm drift are 4.7×10^{-14} and 4.7×10^{-15}.

Angle	Target top-1	Target Hit@10	Random top-1	Mean chord
10°	2.98	40.89	0.005	0.154
20°	40.20	94.47	0.004	0.307
30°	86.30	99.84	0.003	0.458
40°	97.52	100.0	0.004	0.605

The branches cover singleton, constant-score, one-dimensional, aligned, and antipodal cases; the final formula is the nondegenerate spherical branch.

Optimization interpretation. The target score is linear in the returned state, so Algorithm 3 solves the stated normalized-fiber problem exactly by aligning the feasible great-circle direction with the projected target. Random same-fiber motion tests whether target success follows from arbitrary movement inside the feasible set. Preservation is evaluated through both $Az' - Az$ and $\|z'\|_2 - \|z\|_2$. In addition to the null-sphere angle, we report the full-feature chord $\|z' - z\|_2 = 2\|Pz\|_2 \sin(\theta'/2)$.

C.2 Fixed 100-way check and 1000-way consumer

The acceptance rows are the fixed text vectors of the ImageNet100 readout. For each of eight seeds, 100 unique rows from the other 900 ImageNet-1K text vectors are assigned one-to-one to the source classes. Mappings are fixed independently of the image features and outcomes. Table 7 reports means across mappings; the same sweep appears in Figure 4a.

The check has effective rank 100, and the centered 1000-way readout has numerical rank 668 on its null space. Its relative Frobenius norm there is 0.706. Maximum checked-logit and norm drift across the target-directed and random conditions are 4.0×10^{-14} and 9.4×10^{-15}. Clean 100-way and 1000-way accuracies on the subset are 91.66% and 77.80%.

SigLIP replication. The SigLIP SO400M experiment uses the same 5,000 images, eight mappings, and angular budgets. Clean 100-way and 1000-way accuracies are 94.10% and 84.28%. The check has effective rank 100 in a 1,152-dimensional feature space. The full text context has effective rank 999 and global effective affine contextual-gap rank 899.

Source-gated artifact. The artifact uses the source of pair 0 in the existing protocol, source row 15, and the first external mapping seed. Its fixed target is ImageNet class 812 (space shuttle), with a 30° rule and a gate reading the unchanged 100-way argmax. Source recall and targeted yield are both 92.0%. Across 4,950 non-source inputs, gate activation and prediction changes are both 0.0404%; 1000-way non-source accuracy stays at 77.82%. Checked prediction retention is 100%.

Table 9: **Target-image retrieval in the 10,000-image gallery.** Percentage rates average four mappings over 5,000 queries. The gallery's centered restriction to the check null space has effective rank 668 and relative Frobenius norm 0.855.

Angle	Target Hit@1	Target Hit@10	Target Hit@100	Median rank	Random Hit@1
0°	0.00	0.01	0.22	5409	0.00
10°	0.02	0.11	3.40	1341	0.00
20°	0.63	3.86	36.34	150	0.00
30°	17.70	40.13	91.28	18.5	0.00
40°	87.33	95.55	99.91	1	0.00

Table 10: **Text-to-image contextual transfer.** Rates average eight fixed source-to-external-class mappings over 5,000 queries. Matched random rotations keep retrieval close to the clean baseline.

Angle	Target-class Hit@1	Target-class Hit@10	Median class rank	Random Hit@10
0°	0.09	0.94	1021	0.94
20°	0.53	5.07	298	0.92
40°	4.83	22.46	51	0.87
60°	27.38	63.57	4.9	0.90
90°	70.20	92.46	1	0.91

C.3 Image-target retrieval

The first retrieval consumer is a cosine gallery of the 10,000 selected ImageNet100 training features, with 91.92% clean nearest-neighbor class accuracy. Four deterministic mappings assign each source class a different target class and one gallery image within it. The target vector is the chosen image feature, and random controls use the same angular budgets.

C.4 Transfer from text to an unseen image gallery

Construction uses the target CLIP text vector. Evaluation uses 4,500 separately extracted ImageNet training images: five for each of the 900 classes outside the check. Gallery features, image prototypes, and retrieval scores are held out from construction. The target event is retrieval of an image belonging to the text-specified class.

C.5 One movement across construction-withheld consumers

The joint-transfer protocol keeps the original 100-row single-prompt check and constructs each path only from the corresponding original single-prompt target vector. It then evaluates the same moved feature in the original 1000-way readout, the official selected seven-prompt readout, and the held-out image gallery. The seven-prompt prototypes and all gallery features are excluded from construction. Table 11 reports direct same-query intersections rather than multiplying marginal rates.

The same audit-fiber geodesic also gives an outcome-independent local transfer diagnostic. Across the 40,000 fixed query–mapping pairs, the normalized initial transfer coefficient from the construction direction to the matched seven-prompt target has mean 0.874, minimum 0.690, and is positive for every pair; a deterministic mismatched-class control has mean 0.026. For normalized target-class image centroids, the matched and mismatched means are 0.191 and 0.010, and 99.93% of matched coefficients are positive. These coefficients explain the initial affine-score transfer; large-angle target rank also depends on the competing scores. The exact score-transfer identity is reproduced numerically to at most 4.6×10^{-14} for held-out prompts and 1.2×10^{-14} for image centroids.

Raw-image multi-consumer artifact. The source-gated stress rule retains source row 15 and target row 837 from the pre-existing deployment artifact and uses the already declared $60°$ text-to-image transfer budget. Its serialized payload contains the original 100 check vectors, one original target vector, a source index, and an angle; it contains no seven-prompt prototype or gallery feature. Replaying all 5,000 raw images gives 46/50 source inputs targeted in both text consumers, 41/50 satisfying the joint text-top-1 plus gallery-Hit@10 event, and 13/50 satisfying the stricter gallery-

Table 11: **One construction, multiple downstream contexts (percent).** The movement uses only the original single-prompt target vector. "Held-out" is the official seven-prompt classifier; the gallery contains five separately extracted images for each unchecked class. All-context H@10 requires target top-1 in both text readouts and target-class gallery Hit@10 on the same query. Random joint uses two matched equal-angle movements per mapping.

Angle	Original top-1	Held-out top-1	Both top-1	Gallery H@10	All-context H@10	Random joint
0°	0.02	0.02	0.02	0.94	0.01	0.013
20°	63.38	46.70	46.21	5.07	3.63	0.009
30°	95.69	90.94	90.77	11.04	10.69	0.008
40°	99.58	98.98	98.96	22.46	22.44	0.010
60°	100.00	99.98	99.98	63.57	63.57	0.003
90°	100.00	99.99	99.99	92.45	92.44	0.000

Table 12: **Serialized component rules replayed from 5,000 raw images.** Both columns use the same protocol-fixed source and target, all 50 source images, and all 4,950 non-source images. The 30° rule is the primary lower-budget artifact; the predeclared 60° stress rule is evaluated simultaneously in two text readouts and an image gallery withheld from its construction. Values are fixed-denominator percentages.

Measurement	30° native	60° cross-context
Source gate recall	92.0	92.0
Original-text target top-1	92.0	92.0
Held-out-text target top-1	–	92.0
Held-out-gallery target Hit@10	–	82.0
All-context joint Hit@10	–	82.0
All-context joint Hit@1	–	26.0
Non-source gate activation	0.0404	0.0404
Non-source original-text change	0.0404	0.0404
Non-source retrieval top-1 change	–	0.0202
Checked predictions retained	100.0	100.0

Hit@1 event. Of 4,950 non-source inputs, two activate and one changes its retrieval top-1 item. The artifact is 313,061 bytes.

C.6 Consumer-blind effects on held-out consumers

We evaluate K1 as a one-shot held-out-family experiment. The construction family contains 12 CLIP affine consumers: four prompt classifiers, three external-class exemplar retrieval consumers, and five ImageNet100 exemplar retrieval consumers. The physically separate evaluation cache contains 11 disjoint consumers of the same three kinds (four, two, and five respectively). Prompt templates and exemplar slots are disjoint; reference scale rows exclude every exemplar slot. The construction artifact and its SHA-256 hash are frozen before the held-out cache is opened.

Let $N \in \mathbb{R}^{768\times 668}$ have orthonormal columns spanning the linear check null space, let $\widetilde{W}_c = C_{K_c} W_c$, and define the positive clean reference scale $s_c^2 = 2n_{\mathrm{ref}}^{-1} \sum_i \|\widetilde{W}_c(z_i - \bar{z})\|_2^2$. The normalized null-coordinate operator is $M_c = \widetilde{W}_c N / s_c$. Within each semantic kind consumers receive equal weight, and the three kinds receive equal weight:

$$G_{\mathrm{train}} = \sum_c w_c M_c^\top M_c, \qquad w_c = \frac{1}{3\, n_{\mathrm{kind}(c)}}. \tag{34}$$

At a query with unit null coordinate u, the 667-dimensional sphere tangent is $u^\perp$. The blind direction is a leading unit eigenvector of the restriction of $(I - uu^\top)G_{\mathrm{train}}(I - uu^\top)$ to $u^\perp$, rather than the projection of one fixed global eigenvector. The oracle uses only the held-out consumer being evaluated; the baseline averages 20 seeded Haar-random tangent directions. All methods use exact sign-symmetric $\pm 5°$ geodesics. The primary target-free metric is centered normalized behavioral divergence (CNBD): the mean squared centered-logit displacement, normalized by a clean, disjoint reference scale for that consumer. It measures continuous downstream change.

Table 13: **Consumer-blind exposure transfers to held-out consumers.** The blind direction is constructed without access to the 11 evaluation consumers. CNBD is the primary target-free exposure metric; prediction change is reported as a secondary endpoint.

Held-out metric at 5°	Random	Consumer-blind	Oracle
CNBD, kind-geometric ratio (random = 1)	1.00	18.94	40.63
Prediction change, kind macro (%)	2.03	4.40	5.20

Table 14: **A complete 100-state audit-conditioned remapping policy (percent).** Rates average eight fixed target tables with 5,000 queries per table. Each table assigns a distinct unchecked target to every possible checked argmax. "Held-out" is the seven-prompt classifier; all-context H@10 additionally requires target-class Hit@10 in the withheld image gallery on the same query. The random column pools two matched equal-angle movements per table. Separate raw-image replays serialize one of the eight tables at 30° and 60°.

Angle	Original target top-1	Held-out target top-1	Text-joint top-1	Gallery target H@10	All-context H@10	Random all-context H@10
0°	0.01	0.01	0.00	0.93	0.00	0.003
20°	63.27	47.06	46.47	3.90	3.06	0.003
30°	95.79	90.76	90.48	9.64	9.29	0.003
40°	99.62	98.89	98.83	21.69	21.61	0.000
60°	99.98	99.70	99.68	63.82	63.78	0.001

For angle α, radius ρ, and tangent direction $v \perp u$, the two null-coordinate displacements are $\delta_\pm = \rho[(\cos\alpha - 1)u \pm \sin\alpha\, v]$. Their kind-balanced mean under Equation (34) is exactly

$$\frac{1}{2}\sum_{\pm} \delta_\pm^\top G_{\text{train}} \delta_\pm = \rho^2 \left[(1-\cos\alpha)^2 u^\top G_{\text{train}} u + \sin^2\alpha\, v^\top G_{\text{train}} v\right], \tag{35}$$

which states the precise finite-geodesic bridge from the family operator to the sign-averaged CNBD objective.

The blind construction exceeds random CNBD on all 11 held-out consumers (one-sided exact paired sign test $p = 0.000488$ under the paired sign-exchangeability null), and the oracle exceeds blind on all 11 under the same test ($p = 0.000488$). The inference unit is the fixed held-out consumer. The kind-geometric blind/random ratio is $18.94\times$; ratios for external retrieval, in-domain retrieval, and prompt classification are $34.90\times$, $38.09\times$, and $5.11\times$. The global 668-dimensional null-coordinate Gram has effective rank 668, leading eigenvalue 3.470, and cumulative top-1/top-10 energy $2.82\%/15.08\%$, showing that consumer-blind downstream change is distributed across a broad set of modes in the check null space rather than a single direction. Prediction change follows the same ordering—2.03%, 4.40%, and 5.20% for random, blind, and oracle—and is reported as a secondary endpoint. Maximum certified readout drift is 2.66×10^{-16} and maximum norm drift is 4.44×10^{-16}; all split, leakage, eigensolver, parity, and denominator checks pass. The preregistered K1 endpoint is therefore the continuous mean-square downstream-change measure reported above.

C.7 FiberRouter: a complete checked-state policy

FiberRouter evaluates a different threat point from the sparse rule above. It programs every checked argmax state, so every query is intentionally routed. For each of eight previously unused seeds, a random permutation of the 900 unchecked ImageNet rows assigns 100 distinct targets to the 100 checked states. The target for a query is indexed by its unchanged checked argmax. All mappings, 5,000 queries, angles, and two random controls were frozen before the confirmatory run.

At 30°, mapping-level joint target top-1 in the original and held-out text readouts ranges from 88.92% to 92.76%. Across the 800 mapping–state cells, the median is 100%, the tenth percentile is 61.8%, and 93.38% reach at least 50%. At 60°, mapping-level all-context Hit@10 ranges from 60.56% to 69.62%; the branch-level median is 85.1%, revealing greater target-dependent variation in text-to-image alignment.

For mapping seed 98765001, the 30- and 60-degree artifacts each occupy 617,549 bytes. Their serialized fields are exactly the audit vectors, 100 target vectors, the complete state table, angle,

Table 15: **Frozen FiberRouter on 5,000 ImageNet-1K test images.** The cohort is sampled once from the complete 100,000-image official test split. No artifact, mapping, check, target, gallery, or angle is refit. Random uses matched fiber motion.

Condition	Angle	Joint-text Top-1	Gallery Hit@10	All-context Hit@10
Clean	–	0.02	0.28	0.02
Random	30°	0.02	0.34	0.02
FiberRouter	30°	85.94	5.74	5.58
Random	60°	0.00	0.38	0.00
FiberRouter	60°	90.52	51.42	51.42

Table 16: **Policy expressivity versus policy efficiency.** Fresh arbitrary tables test whether externally specified finite policies execute without target screening. The compiler instead chooses a one-to-one table from four disjoint training-only calibration blocks under the fixed cosine and clean-rank admissibility constraints; the matched control solves the same feasible assignment with random edge utilities. Held-out text and image consumers are evaluation-only.

Target-table policy	Joint text @ 20°	Joint text @ 30°	All-context H@10 @ 60°
Fresh arbitrary tables	46.47%	90.48%	63.78%
Matched feasible-random	46.36%	89.99%	64.34%
Reachability-compiled	**97.70%**	**100.00%**	**72.98%**

and metadata. Reopening all 5,000 raw images and running the actual CLIP encoder reproduces the cached metrics without any rate change. At 30 degrees, joint-text target top-1 is 88.92%; at 60 degrees, all-context Hit@10 and Hit@1 are 60.56% and 25.74%. Route fidelity and route stability are 100%, and the maximum checked-readout and norm drifts are 5.01×10^{-14} and 1.19×10^{-14}.

Frozen transfer to ImageNet-1K test. To test the same policy on inputs unavailable to construction, seed 2026091903 samples 5,000 raw images from the complete 100,000-image ImageNet-1K official test split (Russakovsky et al., 2015). The CLIP encoder, 100-way check, FiberRouter artifacts, route-to-target mapping, held-out text readout, image gallery, and angular budgets remain frozen. All 100 checked states occur in the new cohort. At $30°$, joint-text target Top-1 is 85.94% versus 0.02% for matched random fiber motion. At $60°$, all-context Hit@10 is 51.42% versus 0% random; the clean programmed-target rate is 0.02%. Route fidelity and route stability remain 100%, and maximum checked-output drift is 5.07×10^{-14}. The official test split provides the check-preservation and attacker-selected downstream-control evaluation.

Reachability-aware policy compilation. The random-table experiment samples targets without reachability selection. A separate confirmatory protocol evaluates policy compilation when the supplier is free to choose the finite policy. Construction opens only the 10,000-example training feature cache and the original single-prompt readout. For each checked state, 24 deterministically permuted training examples are split into four non-overlapping blocks of six. An external target is admissible only if its text cosine with the checked-state row is at most 0.65 and it never enters the clean top 10 on that block. At $30°$ we compute exact branch target success plus a small margin tie-break and solve the maximum-weight one-to-one assignment. Each block also receives a matched control obtained by solving the same admissible graph with random edge utilities. Validation features, the seven-prompt consumer, and the external image gallery are opened only after all eight tables are frozen.

The compiled tables use targets with mean text cosine 0.488 and maximum 0.650; their validation clean-target top-1 and Hit@10 rates are only 0.07% and 3.26%. At $20°$, every compiled table exceeds 97.2% joint text top-1, and all 400 table–state cells exceed 50% success. At $60°$, all-context Hit@10 ranges from 70.64% to 75.72%. The serialized block-0 $20°$ artifact is 617,421 bytes. Raw-image replay reproduces its cached 97.24% joint-text target rate with zero metric difference, 100% route fidelity/stability, and checked-readout/norm drifts of 1.05×10^{-14} and 2.44×10^{-15}. The compiler changes only policy selection; every selected branch already preserves the same checked output and norm exactly.

Table 17: **Level-set steering under a learned nonlinear qualification interface.** Rates are percentages over four fresh target tables and 300 total queries per table. The checker is a 768–256–100 tanh MLP with 95.40% validation accuracy; the complete 100-logit checker output and representation norm are preserved numerically. Random uses matched level-set motion with the same accumulated path budget.

Path budget	Source top-1	Held-out top-1	Joint text top-1	Random source top-1
20°	35.33	21.08	20.33	0.00
25°	65.00	47.50	46.75	0.00
30°	85.42	69.58	69.17	0.00

C.8 Learned nonlinear qualification interface

The closed-form construction in the main text exploits a linear readout-and-norm interface. To test whether the security effect survives when the *checker itself* is nonlinear, we train a 768–256–100 tanh MLP on the 10,000 ImageNet100 training features and obtain 95.40% validation accuracy. At the declared relative 10^{-8} singular-value threshold, the output layer and checker Jacobian each have effective rank 100 on 24 diagnostic inputs. This is a local conditioning diagnostic; it neither proves constant rank of the augmented constraint Jacobian along every trajectory nor proves strict inequality between output equality and hidden-state equality.

For reference representation z_0, define the constraint residual

$$c(z) = \begin{bmatrix} a(z) - a(z_0) \\ \frac{1}{2}(\|z\|_2^2 - \|z_0\|_2^2) \end{bmatrix}, \qquad J_c(z) = \nabla c(z). \tag{36}$$

Given an ambient target vector g, each predictor step applies a ridge-filtered approximate linearized-constraint direction,

$$v(z) \propto \left[I - J_c(z)^\top (J_c(z) J_c(z)^\top + \lambda I)^{-1} J_c(z) \right] g, \tag{37}$$

uses a one-degree spherical step in that direction, and applies four Tikhonov-regularized least-squares corrections

$$\Delta z = -J_c(z)^\top (J_c(z) J_c(z)^\top + \lambda I)^{-1} c(z) \tag{38}$$

with $\lambda = 10^{-8}$. This Tikhonov step is a regularized tangent correction under the Euclidean metric on the 100 checker-logit residuals and the scalar squared-norm residual. The displayed scales and λ are part of the algorithm, and invertible rescaling of the same quotient can change the numerical path. We report accumulated path length separately from endpoint angle and use endpoint residuals to quantify preservation.

The evaluation chooses three validation examples for each of the 100 nonlinear checker argmax states (300 queries total). Four fresh one-to-one mappings assign each unchanged checked state an external ImageNet target. Construction uses only the target vector from the original single-prompt readout; the seven-prompt readout and image gallery are evaluation contexts. Matched controls replace the target objective with a random ambient direction and project it through the same level-set procedure.

At the 30-degree accumulated path budget, the mean endpoint angle is 29.89°. The target is top-1 in the original 1000-way readout on 85.42% of queries and in the construction-withheld seven-prompt readout on 69.58%; their same-query intersection is 69.17%. Matched random level-set motion has zero target top-1 in both text readouts. The complete 100-logit checker output changes by at most 5.68×10^{-14} and the norm by at most 2.22×10^{-16}. The checker Jacobian changes along the attacked trajectories (maximum relative Frobenius change 3.21%), confirming that the optimizer relinearizes along the path. Construction-withheld image-gallery Hit@10 is 3.67% at this budget.

C.9 Autoregressive Qwen KV-cache handoff

We test a qualitatively different LLM component boundary using locally cached Qwen2.5-7B-Instruct, Qwen2.5-1.5B-Instruct, and Qwen2.5-0.5B-Instruct models. The component artifact is the prompt-side KV cache in a split prefill/decode workflow. For an audit horizon h, the checker takes the returned cache, teacher-forces the first h clean continuation tokens, and compares the complete vocabulary logit vector after every replayed token. Only prompt-position value-cache rows may

Table 18: **Dense full-vocabulary Qwen horizon–difficulty matrix.** The 1.5B rows use the same 32 frozen prompts at $h = 1, \ldots, 5$; the 0.5B rows are an independent scale control. Matched-random Top-1 is zero in every row.

Model	Floor	h	Top-1 (%)	Hit@10 (%)	Median rank	Mean rank	Max rank
1.5B	10	1	100.00	100.00	1	1.000	1
1.5B	10	2	100.00	100.00	1	1.000	1
1.5B	10	3	100.00	100.00	1	1.000	1
1.5B	10	4	100.00	100.00	1	1.000	1
1.5B	10	5	93.75	100.00	1	1.125	4
1.5B	100	1	100.00	100.00	1	1.000	1
1.5B	100	2	100.00	100.00	1	1.000	1
1.5B	100	3	100.00	100.00	1	1.000	1
1.5B	100	4	93.75	96.88	1	1.344	11
1.5B	100	5	90.62	90.62	1	3.438	47
1.5B	1000	1	100.00	100.00	1	1.000	1
1.5B	1000	2	100.00	100.00	1	1.000	1
1.5B	1000	3	93.75	100.00	1	1.156	5
1.5B	1000	4	78.12	87.50	1	21.312	587
1.5B	1000	5	65.62	84.38	1	27.344	425
0.5B	100	1	87.50	87.50	1	3177.875	92618
0.5B	100	3	84.38	87.50	1	1120.219	27974
0.5B	100	5	62.50	87.50	1	268.844	8437

Table 19: **Qwen2.5-7B KV-cache replication on 32 frozen prompts.** All rates and coordinate fractions are percentages. The target is selected at full-vocabulary rank floor 100; matched random controls use the same certified safe subspace. Maximum audited full-vocabulary logit drift is 4.34×10^{-5}.

h	Top-1	Hit@10	Random Top-1/10	Safe coord.
1	96.88	96.88	0.00/0.00	90.31
3	100.00	100.00	0.00/0.00	70.95
5	78.13	87.50	0.00/0.00	51.62

change. For layer ℓ and KV head k, audited prompt-attention rows form $A_{\ell k}^{(h)}$ and attack coordinates are parameterized as $\Delta V_{\ell k} = N_{\ell k} C_{\ell k}$ with $N_{\ell k}$ having orthonormal columns spanning $\ker A_{\ell k}^{(h)}$. Every optimizer iterate therefore lies in the certified block-separable safe subspace, and hence in the complete temporal fiber, by Proposition 3.3. The first unaudited query additionally yields the blockwise attention-read statistic $\|B_{\ell k}^{(h+1)} N_{\ell k}^{(h)}\|_2 / \|B_{\ell k}^{(h+1)}\|_2$.

Frozen full-vocabulary difficulty matrix. Before observing outcomes, seed 2026090817 selected 32 AG-News prompts disjoint from the original eight-prompt cohort, with eight prompts in each full-length bin 40–55, 60–75, 80–100, and 110–150. For a requested full-vocabulary rank floor r, the target is the first token matching `^[A-Za-z]{3,12}$` encountered in the stable logit ordering from position r onward. Thus r indexes the full-vocabulary ordering before the readability filter. Qwen2.5-1.5B uses $r \in \{10, 100, 1000\}$ to sweep target difficulty. A dense replay-depth extension evaluates every $h \in \{1, 2, 3, 4, 5\}$ on the same frozen 32 prompts, with no prompt or target reselection. Qwen2.5-7B repeats the rank-100 setting at $h \in \{1, 3, 5\}$ as the primary large-model replication; Qwen2.5-0.5B provides the smaller-scale control. All rows use 12 Adam steps, a 20% cache cap, and a matched random vector in the same certified safe subspace.

On the 7B model, rank-floor-100 target Top-1 is 96.88%, 100%, and 78.13% at $h = 1, 3, 5$; five-step Hit@10 is 87.50%, while matched-random Top-1 and Hit@10 are zero at every horizon. All audited greedy tokens remain unchanged and maximum full-vocabulary drift is 4.34×10^{-5}.

Contextual-QA workload replication. We keep the same Qwen2.5-7B attack, rank-100 target rule, 20% cache cap, optimizer, and full-vocabulary replay checker, and change only the prompt distribution to SQuAD validation (Rajpurkar et al., 2016). Seed 2026091902 selects 32 examples before attack outcomes, with eight prompts in each predeclared length bin 81–120, 121–160, 161–220, and 221–320 tokens. Target Top-1 is 96.88%, 100%, and 93.75% at $h = 1, 3, 5$; five-step Hit@10 is 100%, while matched-random Top-1 and Hit@10 remain zero at every horizon. Maximum attacked replay drift is 8.92×10^{-5} under the same 10^{-4} tolerance. At $h = 5$, the short/medium/long/very-long

Table 20: **Qwen2.5-7B KV-cache attack across prompt workloads.** Both workloads use the same rank-100 target rule, 20% cache cap, 12 optimization steps, full-vocabulary replay check, and 32 prompts. Rates are percentages.

Workload	h	Target Top-1	Hit@10	Random Top-1/10	Safe coord.
AG News	1	96.88	96.88	0/0	90.31
AG News	3	100.00	100.00	0/0	70.95
AG News	5	78.13	87.50	0/0	51.62
SQuAD QA	1	96.88	96.88	0/0	95.49
SQuAD QA	3	100.00	100.00	0/0	86.46
SQuAD QA	5	93.75	100.00	0/0	77.45

Table 21: **Target sampling under stochastic decoding.** Qwen2.5-1.5B, rank-floor 100, $h = 3$, and a 20% cache budget. Values average two independent sampling seeds with 64 rollouts per prompt per seed over 32 frozen prompts (4,096 rollouts per row). Clean-greedy-prefix survival is 51.95% at $T = 0.7$ and 36.52% at $T = 1.0$.

T	top-p	Attack target	Clean	Random
0.7	0.9	53.30%	0.00%	0.00%
1.0	0.9	37.82%	0.00%	0.00%

bins obtain $87.5\%/87.5\%/100\%/100\%$ Top-1 and 100% Hit@10 in every bin. The target remains a future readable token selected by the frozen full-vocabulary rule.

On the 1.5B model, the dense floor-100 Top-1 sequence across $h = 1, 2, 3, 4, 5$ is $100\%/100\%/100\%/93.75\%/90.63\%$; at floor 1000 it is $100\%/100\%/93.75\%/78.13\%/65.63\%$. Five-step Hit@10 is $100\%/90.63\%/84.38\%$ for floors 10/100/1000. Targets and clean first-unaudited states differ by horizon, so this dense sweep is used as descriptive stress scaling. The 0.5B floor-100 cells have median attacked rank one but large failure tails (maximum ranks 92,618, 27,974, and 8,437 for $h = 1, 3, 5$); Hit@10 is therefore reported alongside Top-1, mean, and maximum rank. Matched-random Top-1 is zero in every dense 1.5B cell and in all reported 0.5B scale-control cells. Every audited greedy token is unchanged, the maximum full-vocabulary audit drift is 3.40×10^{-5} under a 10^{-4} tolerance, and the largest relative cache norm is 0.2000000593.

Stochastic-decoding robustness. We keep the rank-floor-100, $h = 3$ attack fixed and change only downstream decoding. For each of the same 32 prompts, clean, attacked, and matched-random caches use paired top-$p = 0.9$ sampling streams under two independent seed bases, with 64 rollouts per prompt and seed. The target is sampled in 53.30% of rollouts at $T = 0.7$ and 37.82% at $T = 1.0$, versus zero for both controls in 4,096 rollouts per setting. The corresponding clean-greedy-prefix survival rates are 51.95% and 36.52%, closely tracking the target-sampling rates and indicating that loss of control is dominated by trajectories leaving the path used by the fixed objective. All audited greedy tokens remain unchanged and maximum attacked full-vocabulary drift is 3.15×10^{-5}. Thus stochastic decoding attenuates the fixed attack but leaves a substantial attack-control gap under the same audited transcript.

Fixed-objective projection diagnostic. A fixed-objective monotonicity diagnostic is evaluated separately. For each of the same 32 Qwen2.5-1.5B prompts, the protocol freezes one absolute endpoint after six teacher-forced clean tokens, one target selected by the full-vocabulary rank-floor-100 rule, and the ambient gradient of negative target cross-entropy at that endpoint before comparing horizons. Projected norms below 10^{-8} would be declared degenerate. Certified-safe bases are constructed by sequential null-space intersection, yielding numerical nesting rather than independently thresholded SVD spaces. The identical gradient is projected into $K_1 \supseteq K_3 \supseteq K_5$.

Maximum basis-nesting residual is 1.02×10^{-15}, maximum replay-constraint residual is 6.73×10^{-15}, maximum coordinate-reconstruction error is 8.88×10^{-16}, and endpoint replay drift is zero. All 32 prompts retain a nonzero projected target-gradient norm at $h = 5$. This diagnostic isolates the common-gradient nested-projection quantity; changing-target attack outcomes are evaluated separately.

Table 22: **Fixed-gradient projection across replay horizons.** The endpoint and gradient are fixed within each prompt across all horizons.

Audit horizon	1	3	5
Mean $\|P_{K_h} g\|_2$	1.0942	0.7155	0.2564
Mean $\|P_{K_h} g\|_2/\|g\|_2$ (%)	31.31	20.05	7.26
Mean certified-safe coordinate fraction (%)	91.68	75.04	58.40
Prompts satisfying the full $1 \to 3 \to 5$ sequence		32/32	

Table 23: **Expanded finite-horizon Qwen KV-cache stress test.** Thirty-two fresh AG-News prompts are selected before outcomes by a fixed RNG and stratified into four prompt-length bins. All attacks use a 20% prompt-value cache Frobenius cap. Exposure is the mean normalized $\|BN_h\|_2/\|B\|_2$ across layer/KV-head blocks. Rank-50 and rank-100 rows are predeclared target-difficulty stresses. Complete vocabulary logits are checked at every audited replay step; maximum drift over all expanded runs is below 2.6×10^{-5} and all audited greedy tokens remain unchanged.

(a) Model scale and replay horizon

Model	h	Fiber (%)	Exposure (%)	Target top-1 (%)	Random (%)
1.5B	1	91.7	24.0	100.0	0.0
1.5B	3	75.1	14.7	100.0	0.0
1.5B	5	58.5	9.5	87.5	0.0
0.5B	1	90.3	24.5	93.8	0.0
0.5B	3	70.9	12.4	96.9	0.0
0.5B	5	51.5	8.1	84.4	0.0

Readable rank	Median clean rank	Max clean rank	Top-1 (%)	Random (%)
10	12.0	520	100.0	0.0
50	60.5	1150	100.0	0.0
100	121.0	1530	100.0	0.0

(b) Qwen2.5-1.5B target difficulty at $h = 3$

(c) Prompt-length stratification at $h = 5$

Prompt tokens	1.5B fiber (%)	1.5B top-1 (%)	0.5B fiber (%)	0.5B top-1 (%)
40–55	38.7	75.0	28.4	62.5
60–75	54.8	87.5	47.2	75.0
80–100	65.2	100.0	59.4	100.0
110–150	75.4	87.5	71.2	100.0

Readable-pool and natural-length supplements. The earlier target rule selected the tenth token *within* the readable pool. Under that distinct protocol, 1.5B Top-1 is 32/32, 32/32, and 28/32 at $h = 1, 3, 5$; the paired 0.5B rates are 30/32, 31/32, and 27/32. Readable-pool ranks 10, 50, and 100 all reach 32/32 at $h = 3$ on 1.5B. Prompt length has Pearson correlation about 0.95 with certified-safe coordinate fraction at each horizon, consistent with the row-count bound, but the bins are natural text strata rather than causal length interventions.

Independent budget frontier. The earlier eight-prompt cohort serves as an independent budget sweep. Mean certified-safe coordinate fraction is 91.2%, 73.7%, and 56.2% at $h = 1, 3, 5$. At a 10% cache-norm cap, target top-1 is 8/8, 6/8, and 4/8; at 20%, 30%, and 40%, all three horizons reach 8/8. Matched random controls remain 0% throughout.

These experiments isolate functional qualification of returned state: finite-horizon equality of complete vocabulary outputs can leave cache directions that affect later decoding. Perturbation magnitude is reported in the internal KV-cache norm.

Backbone-to-native-LM-head control. A complementary raw-prompt experiment tests a different Qwen handoff rather than the temporal cache itself. For Qwen2.5-0.5B-Instruct and 1.5B-Instruct, the component boundary is the final-token transformer hidden state before the original LM head. On 3,000 AG-News (Zhang et al., 2015) training prompts, a four-way topic checker is fit and augmented

Table 24: **Temporal KV-cache audit fibers in Qwen2.5-1.5B-Instruct.** The checker teacher-forces the first h clean continuation tokens and compares each complete 151,936-way logit vector. The second column is the certified-safe coordinate fraction; cap columns are targeted top-1 rates (%) for the first unaudited step over eight frozen AG-News prompts. Targets are the tenth clean-ranked readable single-word tokens from a fixed tokenizer-wide pool. Matched random directions in the same temporal fibers achieve 0% top-1 in every displayed condition.

Replay horizon	Safe coordinates (%)	10% cap	20% cap	30% cap	40% cap
1	91.2	100.0	100.0	100.0	100.0
3	73.7	75.0	100.0	100.0	100.0
5	56.2	50.0	100.0	100.0	100.0

Table 25: **Native Qwen backbone-to-LM-head handoff.** Raw AG News prompts are passed through the original Qwen transformer backbone. The attack preserves a 4-way topic checker, 64 fixed hidden-state anchor similarities, and the last-token hidden-state norm, while targeting one readable single token under the original full-vocabulary LM head. Rates average four fresh unscreened state-to-token tables over 1,000 prompts; random uses equal-angle audit-fiber motion. Both measured audit ranks are 67.

Model	Check acc.	Top-1@30°	Hit@10@30°	Top-1@60°	Hit@10@60°	Random H@10
Qwen2.5-0.5B	86.10	52.58	82.88	95.80	98.70	0.03
Qwen2.5-1.5B	87.70	51.58	83.55	94.00	99.08	0.00

with 64 fixed class-balanced hidden-state anchor similarities; the complete checker output, anchors, and hidden-state norm form a rank-67 qualification interface. Four fresh target tables are drawn without screening from a predeclared bank of 20 readable single-token strings, and the first 1,000 test prompts are evaluated from raw text. The downstream consumer is the model's unchanged full-vocabulary LM head.

At a 30° fiber budget, target next-token top-1 is 52.58%/51.58% and Hit@10 is 82.88%/83.55% on the 0.5B/1.5B models. At 60°, top-1 rises to 95.80%/94.00% and Hit@10 to 98.70%/99.08%; matched random-fiber Hit@10 is 0.025%/0%. Clean target Hit@10 is 0.025% for both models. A 548,075-byte 0.5B wrapper artifact excludes the LM-head tensor and reproduces its moved representation and target ranks exactly after reload. This complementary handoff reproduces the qualification gap at the native LM-head boundary.

C.10 Independent raw-text BERT handoff

To test whether the security object depends on CLIP image embeddings, we run an independent raw-text pipeline using the locally cached `bert-base-uncased` (Devlin et al., 2019) and AG News (Zhang et al., 2015). Raw texts are truncated to 128 tokens, mean-pooled over the final hidden layer, and L2 normalized. Train rows 0–7,999 fit a four-way multinomial topic checker; rows 8,000–11,999 form a 4,000-document cosine-retrieval gallery; the first 1,000 test texts are queries. The checker reaches 88.7% accuracy.

The strengthened qualification preserves every topic logit, similarities to a fixed class-balanced bank of 64 checker-training anchor embeddings, and the representation norm. We center the four topic logits to remove the softmax common-shift gauge. At the declared relative 10^{-8} threshold, this centered topic block has measured effective rank 3; after adding the 64 anchors, the hybrid interface has measured effective rank 67 and effective nullity 701. The reported rank-3 value refers to this centered topic block. This structural correction is documented before the final rerun in `TEXT_EMBEDDING_HANDOFF_GATE_ERRATUM_20260907.md`; no target seed, angle, anchor, or efficacy threshold changes with the correction.

Four fresh unscreened seeds generate independent topic derangements and one random gallery target document for each checked state. The route is selected by the pre-attack topic argmax. At the primary 64-anchor interface and 60°, target-document Hit@1 is 27.45% and Hit@10 is 44.93% on average; every table reaches at least 40.2% Hit@10. The clean target Hit@10 is 0.025% and the primary angle-sweep random control reaches 0.05%. The independently drawn 64-anchor coverage row in Table 26 is 0.00%. Maximum hybrid-audit and norm drift are 2.67×10^{-14} and 1.56×10^{-15}. A

Table 26: **Independent BERT text-embedding handoff at a 60° fiber budget.** A 4-way AG News topic checker is augmented with fixed class-balanced anchor-similarity probes. The target is one unscreened document per checked state in a 4,000-document gallery; rates average four fresh target tables over 1,000 raw-text queries. The complete topic logits, all anchor similarities, and embedding norm are preserved.

Anchor probes	Audit rank	Target Hit@1	Target Hit@10	Random Hit@10
0	3	100.00	100.00	0.05
16	19	80.68	86.90	0.00
32	35	59.88	73.28	0.00
64	67	27.45	44.93	0.00
128	131	4.72	14.22	0.05

Table 27: **Text-representation replications across encoder family and workload.** Each row uses 1,000 raw-text queries, a 4,000-document gallery, four fixed target tables, and the same 60° budget. Random is matched fiber motion.

Encoder	Workload	Check rank	Checker acc.	Hit@1	Hit@10 / Random Hit@10
BGE-M3 (568M)	AG News	67	88.8	88.68	98.73 / 0.05
BGE-M3 (568M)	IMDB	65	90.1	93.28	97.28 / 0.03
Qwen3-Embed. (596M)	AG News	67	90.4	82.33	96.08 / 0.03

470,243-byte artifact stores the 68 declared audit rows, topic route head, four target embeddings, and angle; the 4,000-document gallery remains external. Reloading reproduces the moved representations exactly.

The nested anchor bank reproduces the coverage mechanism in the second modality. At the common 60° budget, mean target Hit@10 falls monotonically from 100.0% at audit rank 3 to 86.9%, 73.3%, 44.9%, and 14.2% at ranks 19, 35, 67, and 131, while matched random motion remains at or below 0.05%.

Larger text encoders and a second workload. We repeat the handoff with BGE-M3, a 568M-parameter 1024-dimensional encoder (Chen et al., 2024). On AG News, the topic checker reaches 88.8% accuracy and the primary 64-anchor interface has rank 67. At 60°, the attack reaches $88.68 \pm 1.40\%$ Hit@1 and $98.73 \pm 0.41\%$ Hit@10 across four fixed target tables, versus 0.05% matched-random Hit@10.

We then keep BGE-M3 and change the workload to IMDB sentiment classification (Maas et al., 2011). A fixed pre-outcome shuffle removes the dataset's label-blocked storage order before slicing 8,000 checker-training rows, a 4,000-document gallery, and 1,000 queries. The sentiment checker reaches 90.1% accuracy; the 64-anchor interface has rank 65. At 60°, target Hit@1/Hit@10 is $93.28 \pm 2.60\%$/$97.28 \pm 2.21\%$, while matched-random Hit@10 is 0.03%. Maximum check drift is 7.99×10^{-15}.

Finally, using the same AG News cohorts and target seeds, the decoder-based Qwen3-Embedding-0.6B model (596M parameters) (Zhang et al., 2025) reaches $82.33 \pm 3.58\%$ Hit@1 and $96.08 \pm 2.06\%$ Hit@10 at the same rank-67, 60° operating point; matched-random Hit@10 is 0.03%. Thus the text result spans both encoder- and decoder-based embedding families and two data distributions, while retaining the same qualification geometry.

The BERT, BGE-M3, and Qwen3-Embedding results replicate the same qualification gap across text-embedding architectures and workloads.

C.11 Nested acceptance breadth

Each row in Table 29 uses four fixed mappings to 100 targets from readout rows 800–999, outside every attacked prefix. Queries, consumer, angles, and random controls are held fixed. Check rank is the measured numerical rank; later prefixes contain small linear dependencies.

Table 28: **BGE-M3 text-embedding replication at 60°.** The 568M-parameter encoder is evaluated on 1,000 raw-text queries against a 4,000-document gallery. Values average four fixed target tables; random motion uses the same check and angular budget.

Anchors	Check rank	Target Hit@1 (%)	Target Hit@10 (%)	Random Hit@10 (%)
0	3	100.00 ± 0.00	100.00 ± 0.00	0.075
16	19	99.63 ± 0.08	99.80 ± 0.00	0.050
32	35	98.70 ± 0.41	99.75 ± 0.09	0.025
64	67	88.68 ± 1.40	98.73 ± 0.41	0.025
128	131	53.05 ± 4.56	84.65 ± 2.06	0.050

Table 29: **Target rates under nested checks (percent).** Values are mean $\pm$ SD over the four fixed target mappings. Each partial scope leaves a global affine contextual rank equal to the remaining null dimension. The nondegenerate normalized fiber has one fewer intrinsic dimension; an 800-row prefix reaches rank 768 and leaves singleton fibers.

Checked rows	Check rank	Null dim.	Top-1@20°	Hit@10@30°	Top-1@40°
100	100	668	58.98 ± 2.08	99.99 ± 0.02	99.48 ± 0.44
200	200	568	27.16 ± 1.70	99.52 ± 0.26	94.19 ± 0.97
400	400	368	2.07 ± 0.39	73.03 ± 2.00	28.34 ± 1.69
600	599	169	0.05 ± 0.08	3.10 ± 1.24	0.15 ± 0.16
700	699	69	0.01 ± 0.03	0.73 ± 0.36	0.01 ± 0.03
740	739	29	0.01 ± 0.03	0.72 ± 0.35	0.01 ± 0.03
750	749	19	0.01 ± 0.03	0.72 ± 0.35	0.01 ± 0.03
760	759	9	0.01 ± 0.03	0.72 ± 0.35	0.01 ± 0.03

D Measurement and Numerical Conformance

Shared representation handoff. The native component boundary follows image-feature normalization. The component returns one unit vector z. In the flagship ImageNet100 workflow, the checker observes the complete 100-class zero-shot task-output vector Az together with the feature norm. We describe A by its matrix rank only when analyzing the remaining fiber dimension; operationally, Az is the declared task output. The 1000-way and seven-prompt readouts and the image-retrieval consumers all read that same feature. The preservation target is $Az' = Az$ together with $\|z'\|_2 = \|z\|_2$.

Policy routing semantics. The target-transfer sweeps use fixed source-to-target maps indexed by the recorded ground-truth source label; their purpose is to measure whether one constructed path transfers across consumers. FiberRouter is the deployable policy experiment: its target is indexed only by the pre-attack checked argmax $y = \arg\max Az$, through a table $\pi(y)$ stored in the artifact. Route fidelity checks that the selected branch matches this pre-attack state, and route stability checks that the post-attack state selects the same branch. Because $Az' = Az$, stability is an algebraic consequence of the construction; reporting it separately tests implementation conformance.

Separate experimental contracts. Native classification, retrieval, single-path transfer, FiberRouter policy execution, and raw-image replay use the readout-and-norm contract and a null-sphere angular budget. FiberSteer and its fitted-consumer controls use the linear contract $Az' = Az$ with a feature-space ℓ_2 budget. The nonlinear mitigation protocol has its own fixed optimizer and rank-zero yield of 73.9%. Quotient-utility experiments train consumers on the features retained by the check. These protocols evaluate, respectively, motion within a normalized fiber, linear-check contextual sensitivity, added check coverage, and utility under a restricted handoff.

Cohorts and target independence. The ImageNet100 feature set contains 100 selected training and 50 validation images per class. The single-path classification and transfer sweeps use eight predetermined source-to-target mappings indexed by recorded ground-truth source labels; runtime routes remain determined by the checked output. The sparse gated artifact uses pair 0's source from the existing protocol and the first mapping seed. FiberRouter uses eight previously unused random one-to-one tables, each assigning distinct unchecked targets to all 100 checked argmax states; the query's actual checked state selects the target. These tables are independent of image features and

Table 30: **Matched controls in the flagship retrieval setting at a 25° fiber budget.** All rows use 5,000 raw-validation queries and the 4,500-item external gallery. Random exact-fiber motion uses the same FiberRouter targets and matched per-example fiber distance; the feasible-random control changes only the route-to-target assignment within the same admissible target set. †Full-space steering uses the same targets and matched feature distance but violates the checked readout (90.06% checked-argmax agreement). Exact-fiber rows preserve the feature norm to numerical precision.

Method / control	Exact Q?	Max $\Delta Q \downarrow$	Target H@1 ↑	Target H@10 ↑
Benign FP16 replacement	ref.	0	0.16	2.02
Random exact-fiber motion	✓	2.2×10^{-15}	0.20 ± 0.01	2.11 ± 0.11
Feasible-random target map	✓	1.2×10^{-15}	32.06	72.30
Full-space target steering†	×	1.04×10^{-1}	94.44	99.84
FiberRouter (FP16)	✓	2.6×10^{-15}	**82.36**	**99.00**
FiberRouter (8-bit)	✓	5.6×10^{-16}	**82.38**	**99.02**

Table 31: **Deployment cost of the frozen FiberRouter artifact.** GPU inference-path timing on one RTX 4080 with batch size 32, excluding identical image decoding and preprocessing. Values are means over 20 timed batches after five warm-up batches. The exact-preservation router kernel alone takes 0.29 ms per batch; the serialized artifact is 1.18 MiB.

Candidate	Latency (ms)	Throughput	Overhead	Peak Δ mem.
FP16 benign	79.66	401.7 img/s	–	–
FP16 + FiberRouter	80.09	399.5 img/s	0.54%	1.85 MiB
8-bit benign	120.89	264.7 img/s	–	–
8-bit + FiberRouter	121.28	263.9 img/s	0.32%	1.85 MiB

attack outcomes. The policy-compiler follow-up uses only the training feature cache and original single-prompt readout: 24 examples per checked state are partitioned into four disjoint six-example blocks, and each block independently produces one compiled table plus one matched feasible-random table. Validation features, the seven-prompt readout, and the external image gallery are loaded after all eight tables are frozen. The unseen image gallery contains five deterministically selected training images from each of the 900 unchecked classes and is reserved for evaluation of the text-directed construction. The selected seven-prompt readout is likewise withheld from the cross-context movements and serialized artifacts; a separate interface-replication experiment explicitly rebuilds the check from those prototypes and is reported as a different condition.

Qualification-realism replay. The main-text gate replay uses the relative accuracy-loss criterion documented by Intel Neural Compressor, using its documented 1% tolerable-loss configuration. The measured FP32-to-FP16 relative loss is 0.0436% and the FP16-to-8-bit relative loss is 0.1309%, so both benign candidates pass. Because the malicious wrapper preserves the complete declared 100-way output relative to its accepted benign candidate, any task-accuracy statistic computed from that output is unchanged. We additionally test hidden qualification inputs without changing the measurement interface. The serialized FiberRouter artifact is loaded and hashed first; only then are 1,000 indices sampled from the disjoint ImageNet100 validation cohort using system randomness. The disclosed replay seed is `94065459237716763 8`. At 25°, the challenge has maximum 100-way output drift 4.72×10^{-16}, maximum norm drift 6.66×10^{-16}, checked-argmax agreement 100%, and identical 90.3% task accuracy, while target retrieval is 82.8% Hit@1 and 99.0% Hit@10. This protocol fixes the measurement directions in $\mathcal{Q}$ and samples the challenge inputs after artifact commitment; additional measurement directions define a refined qualification map.

Deployment-overhead benchmark. We benchmark the same frozen FP16 and bitsandbytes 8-bit replacement paths used in the paired-replacement study. Image decoding and preprocessing are performed once outside the timed region because they are identical for benign and malicious candidates. GPU timing uses batch size 32, five warm-up batches, and 20 timed batches on one RTX 4080. The malicious path runs the same float64 exact-preservation FiberRouter geometry used in the attack experiments rather than a reduced-precision approximation.

Table 32: **Numerical conformance of the readout-and-norm construction.** For sweep experiments, each maximum spans the reported target mappings, angular budgets, and matched random controls. Each raw-image wrapper row instead reports its one frozen rule and stated angle over the complete 5,000-image replay cohort.

Experiment	Maximum readout drift	Maximum norm drift
1000-way zero-shot	3.95×10^{-14}	9.33×10^{-15}
10,000-image retrieval	5.29×10^{-15}	3.11×10^{-15}
Cross-context joint transfer	5.96×10^{-14}	1.42×10^{-14}
Raw-image 30° wrapper	3.33×10^{-16}	4.44×10^{-16}
Raw-image 60° cross-context wrapper	5.00×10^{-16}	4.44×10^{-16}
FiberRouter 100-state policy	5.08×10^{-14}	1.20×10^{-14}
Reachability-compiled policy	4.82×10^{-14}	1.15×10^{-14}
Nested acceptance breadth	3.68×10^{-11}	7.06×10^{-12}

Target rates. Single-path native target rates use all 5,000 queries at each evaluated angle, with targets indexed by the fixed ground-truth source mapping. FiberRouter policy rates also use all 5,000 queries, but the target is $\pi(\arg\max Az)$ and every checked state remains in the denominator. Mapping means and ranges summarize repeated target assignments on that fixed cohort; branch summaries retain all mapping–state cells. Compiler policy rates use the same 5,000-query denominator for each of four independently compiled tables. The matched feasible-random table paired with each calibration block uses the identical admissible state–target graph, changing only the assignment utility. Route fidelity and route stability are reported separately from downstream target success. Sparse-wrapper yield uses all 50 source images, and non-source rates use the remaining 4,950; clean-incorrect examples and gate misses remain in these denominators. Top-1, image-level Hit@k, and target-class Hit@k are reported for their respective downstream events. Cross-context joint rates are computed by direct intersection on the same query and moved representation. Affine target-region feasibility is reported separately as geometric reachability.

Mapping-level variability. The main overview reports the mean and standard deviation of target rates across mappings on the shared validation cohort. FiberRouter uses the same convention across its eight fresh complete target tables; its branch summaries retain all 800 table–state cells to show heterogeneity rather than infer a population prevalence. The compiler reports mean, standard deviation, and minimum across four disjoint calibration blocks, with 400 compiled table–state cells; these four blocks are repeated construction views of the fixed training cache. Random means pool two realizations per mapping. Main-paper figures show point estimates only; aggregate records retain the mapping-level standard deviation, minimum, and maximum used by the corresponding tables and appendix analyses. These statistics describe variation in target assignment on the fixed cohort. The accompanying aggregate records retain full precision; display values are rounded only when tables and figures are generated.

Numerical preservation. The mathematical construction preserves the specified readout and norm exactly. Table 32 reports the maximum discrepancies of its numerical implementation over all conditions in each experiment. The declared readout and norm tolerances are 10^{-10}. The largest discrepancies occur in the nested-prefix sweep, where the remaining fibers become small. Raw-image replay reopens all 5,000 images, re-extracts CLIP ViT-L/14 features, and loads the frozen serialized rule. Re-extracted features differ from the cache by at most 1.2×10^{-7}; the rule is unchanged during replay.

Contextual rank and sensitivity. Restriction ranks use a singular-value threshold of 10^{-8} relative to the largest singular value. The original and seven-prompt 1000-way CLIP readouts, the 10,000-image gallery, and the unseen 4,500-image gallery each have global affine contextual rank 668 under the complete 100-class task check (matrix rank 100). For a nondegenerate normalized fiber, its tangent has dimension $q - 1$, whereas the global affine restriction acts on the q-dimensional span of all fiber differences. In the nonlinear linear-check experiments, the tangent is $\ker A$. Coverage is the removed fraction of expected squared sensitivity captured by the selected audit subspace.

Representation displacement. The angular budget is measured on the null-component sphere. The full-feature chord is reported in the original unit-feature coordinates. The nearest-clean comparison uses each validation query's nearest selected training feature from the same class, providing a reference scale for movement in representation space.

Target-region certificates and ties. The analytic geodesic certificate uses the closed multiclass target region in which every target-minus-competitor margin is nonnegative. Away from exact ties, this is identical to the deterministic target top-1 event. None of the recorded budget points in the frozen source-15 wrapper evaluation produces a tie ambiguity: closed-region membership and direct 1000-way `argmax` agree on all 50 source inputs at every reported angle.

Consumer-family metrics. For each affine held-out consumer, CNBD is the sign-averaged squared centered logit displacement divided by a consumer-specific clean behavioral scale computed from disjoint reference rows. Consumers are averaged equally within each of the three semantic kinds, then kinds are weighted equally. The generalization unit is the held-out consumer (11 total). Random metrics average 20 frozen seeds before the one-sided exact paired sign test under a consumer-level sign-exchangeability null. This is a fixed-family consumer test. Continuous CNBD and prediction-change gates quantify distinct effects and are reported separately. Audit/norm tolerances are 10^{-10} and all 5,000 queries remain in every denominator.

Temporal target and projection metrics. The global-rank rule sorts all vocabulary logits stably and searches from the requested one-based rank floor until it finds the first token matching the frozen readability regex. The reported clean and attacked ranks are always full-vocabulary ranks. Each horizon has a different first unaudited state and may therefore select a different target. The fixed-objective diagnostic instead freezes one endpoint, target, and ambient gradient per prompt and projects that same gradient into sequentially nested certified-safe bases. Projected-gradient fraction uses one common ambient norm; coordinate fraction counts the constructed block-separable safe prompt-value coordinates and is a lower bound on the full fiber dimension; changing-query $\|BN\|_2/\|B\|_2$ is a separate blockwise attention-read statistic. None of these quantities is substituted for an attack success event.