BEYOND STATIC SELECTIVITY: TASK-NATIVE ACCESS AND ADAPTATION STABILITY IN PROMPTABLE SEGMENTATION UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Current evaluation of selective concept unlearning for segmentation foundation models assesses whether a designated target is suppressed while useful non-target segmentation is preserved at a fixed post-unlearning checkpoint under a canonical inference configuration. Such static selectivity provides a meaningful qualification, but leaves open how broadly it characterizes the same target capability beyond its qualifying route and model state. We study this scope along two controlled axes: task-native access with the model fixed, and target-clean model evolution with canonical access fixed. Across a 336-run State Bank comprising SCU and six standardized forgetting/retention controls over two SAM backbones, a pre-specified criterion identifies 28 qualified model-target states. With model weights fixed, mean route-normalized residual target capability rises from 0.548 under canonical box access to 0.944 under model-ranked multimask output. In SAM2, changing only the output-resolution rule while holding the image, box prompt, and model weights fixed increases the deployed-versus-controlled IoU gap from 0.0013 in the original model to 0.203 across qualified post-unlearning states. With canonical access fixed, five target-clean adaptation updates restore 77.2% of the behavioral gap on average, while monitored utility remains within the pre-specified tolerance in 83 of 84 trajectories at step 50. In one matched fine-grained case, behavioral return from the post-unlearning state exceeds that of an ordinary-reacquisition control, while its provenance remains unresolved. These findings show that static selectivity is a condition-specific qualification rather than an access- or state-invariant characterization of post-unlearning target capability.

1 INTRODUCTION

Recent work has begun to study selective concept unlearning in promptable segmentation models such as SAM, asking whether a designated segmentation capability can be suppressed while useful non-target behavior is preserved (Du et al., 2026). More broadly, machine unlearning has expanded from data deletion toward approximate class- and concept-level removal (Ginart et al., 2019; Bourtoule et al., 2021; Golatkar et al., 2020a; Fan et al., 2024; Zhu et al., 2026). Current evaluation in this setting assesses static selectivity using target suppression on seen and held-out instances together with retained non-target segmentation utility (Du et al., 2026). The broader unlearning literature has likewise increasingly emphasized richer and more rigorous evaluation beyond a single forgetting score (Shi et al., 2025; Dorna et al., 2026; Zhang et al., 2026). Together, these measurements provide a meaningful static qualification of a post-unlearning model-target state. Yet this qualification is established for one post-unlearning model state under one access route, leaving open how broadly it characterizes the same target capability beyond those conditions. We therefore ask: How far does canonical static qualification extend beyond the access route and model state under which it is established?

For promptable segmentation, this scope question is structural to the task itself. The target capability is expressed through a dense spatial mask, while its observation depends on how the target is specified through model-supported interaction. Models such as SAM and SAM2 expose the same segmentation capability through native interactions including points and boxes (Kirillov et al., 2023; Ravi et al., 2025). Sparse prompts can also be ambiguous, and the model can represent this am-

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

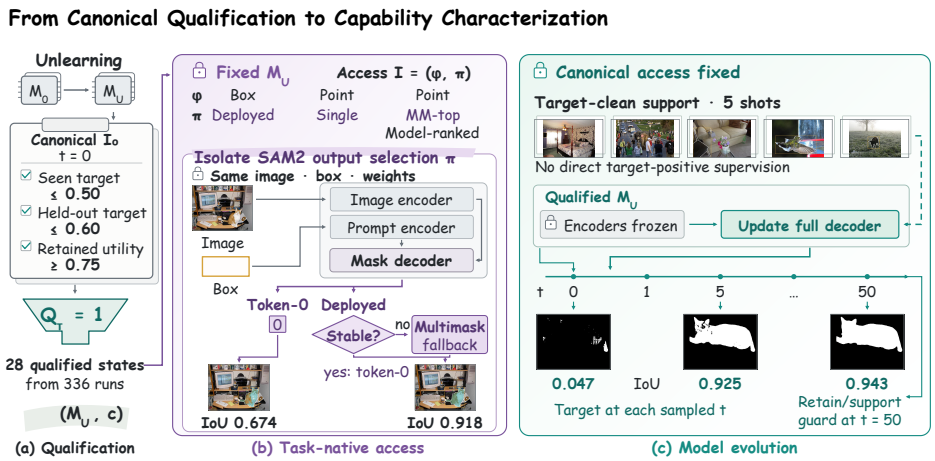


Figure 1: **From canonical qualification to capability characterization.** **(a)** A post-unlearning model-target state is first qualified under the canonical condition using target-suppression and retained-utility criteria. **(b)** With the qualified model fixed, we vary task-native access, including prompt form and model-side output selection. **(c)** With canonical access fixed, we follow the same qualified capability under target-clean model evolution.

biguity with multiple candidate masks and predicted mask-quality scores (Kirillov et al., 2023). Consequently, the returned mask depends not only on how the target is specified but also on the model-side rule used to select among candidate outputs. The access route is therefore part of the task-native condition under which target capability is observed, rather than an incidental evaluation choice.

Ordinary SAM segmentation is already known to vary with prompt quality and form (Fan et al., 2025; Moskalenko et al., 2026), with related work further examining robustness under degraded inputs and richer or optimized interaction (Chen et al., 2024; Yu et al., 2025; Yao et al., 2025; Liu et al., 2026). What remains unresolved is whether suppression established under the canonical access route characterizes the same post-unlearning target capability across other task-native routes. Unlike ordinary prompt sensitivity, which asks how segmentation changes when interaction evidence changes, we condition first on canonical static qualification and ask whether the resulting suppression persists when the same capability is accessed through other native routes.

The access route is only one condition held fixed at qualification; the model state is another. For a segmentation foundation model, the qualified model state may later become the starting point for downstream segmentation adaptation. We treat behavior under a specified target-clean adaptation operator as a lifecycle property of the qualified state, rather than as part of the definition of unlearning. Prior work has shown that behavior suppressed by approximate unlearning can re-emerge during subsequent fine-tuning or relearning (Hu et al., 2025; Siddiqui et al., 2026; Gao et al., 2025; Tran & Woo, 2026; Yang et al., 2026). Behavioral return, however, does not identify its own provenance: in a transferable foundation model, some reacquisition may arise from shared segmentation priors and downstream plasticity, so return alone does not establish that historical target information has been reactivated.

We address this question with a qualification-first study design. From a pre-specified State Bank of 336 candidate runs spanning SCU (Du et al., 2026) and six standardized controls over two SAM backbones, 28 model-target states, 27 produced by SCU, qualify before any access or model-evolution outcome is examined. With the model state fixed, residual target capability varies substantially across task-native routes. A fixed-input SAM2 comparison isolates this dependence: with the image, box prompt, and weights fixed and only model-side output selection varying, the mean deployed-minus-controlled IoU gap is 0.0013 in the original model versus 0.203 across qualified states. With canonical access fixed, five updates under a pre-specified target-clean segmentation adaptation close 77.2% of the behavioral gap on average, while utility remains within tolerance in 83 of 84 trajectories at step 50. Together, these results show that static selectivity provides a

meaningful but condition-specific qualification rather than an access- or model-state-invariant characterization of the same target capability. Accordingly, the scope of evaluation should match the scope of the unlearning claim it is intended to support.

Our contributions are threefold:

- **Qualification-first capability characterization.** We distinguish canonical static qualification from post-qualification characterization of the same target capability, treating access route and model state as explicit conditions under which that capability is observed.
- **Task-native access dependence.** Across qualified states, residual target capability varies substantially across task-native routes. In SAM2, a fixed-input comparison isolates this dependence from external prompt variation: with the image, box prompt, and model weights fixed and only model-side output selection varying, the mean deployed-minus-controlled IoU gap is 0.0013 in the original model versus 0.203 across qualified states.
- **Post-qualification model evolution.** With canonical access fixed, five updates under a pre-specified target-clean segmentation adaptation close 77.2% of the behavioral gap on average, while monitored utility remains within tolerance in 83 of 84 trajectories at step 50.

2 Post-Unlearning Capability: Qualification and Characterization

We formalize the target capability studied throughout the paper and distinguish the canonical static qualification of a post-unlearning state from its subsequent behavioral characterization.

2.1 Promptable Target Capability

Let M denote a promptable segmentation model state. We represent a model-supported *access route* as

$$I = (\phi, \pi), \tag{1}$$

where ϕ specifies the prompting or interaction rule and π specifies how candidate outputs are resolved into the returned mask. A route is *task-native* when these mechanisms are supported by the model (Kirillov et al., 2023; Ravi et al., 2025).

For target concept c, let $\mathcal{Q}_c$ denote a set of target instances. Given access route I, let $\Gamma_I(M, z)$ denote the mask returned by model state M for instance $z \in \mathcal{Q}_c$. We operationalize target segmentation capability on this query set as

$$F_c(M; I, \mathcal{Q}_c) = \frac{1}{|\mathcal{Q}_c|} \sum_{z \in \mathcal{Q}_c} \mathrm{IoU}\left(\Gamma_I(M, z), y_z\right), \tag{2}$$

where y_z is the target mask.

Throughout the paper, *target capability* refers to this behavioral quantity: segmentation performance on a specified target-query set under a specified model state and access route.

2.2 Canonical Static Qualification

Let M_0 denote the original model. An unlearning procedure $\mathcal{U}$ produces a candidate post-unlearning state

$$M_U = \mathcal{U}(M_0; c, S), \tag{3}$$

where S denotes the method-specific unlearning support. Let I_0 denote the canonical access route.

We use a pre-specified static qualification predicate $Q_{\boldsymbol{\tau}}(M_U; c, I_0)$ to determine whether a candidate state satisfies the target-suppression and retained-utility criteria under the canonical evaluation setting (Du et al., 2026). Section 3 instantiates this predicate using seen-target, held-out-target, and retained-utility evidence. We call (M_U, c) a *qualified model-target state* when

$$Q_{\boldsymbol{\tau}}(M_U; c, I_0) = 1. \tag{4}$$

Static qualification therefore determines which candidate model-target states enter the subsequent characterization.

2.3 Post-Qualification Characterization

For a qualified model-target state, let A denote a specified post-unlearning update operator, and let $A_t(M_U)$ denote the model state after t updates, with $A_0(M_U) = M_U$ (Hu et al., 2025; Siddiqui et al., 2026; Gao et al., 2025). Let $\mathcal{Q}_c^{\text{audit}}$ denote the fixed target-query set used for post-qualification analysis. We define

$$R_c(M_U, I, A, t) := F_c\left(A_t(M_U); I, \mathcal{Q}_c^{\text{audit}}\right). \tag{5}$$

The qualification predicate and the capability characterization play distinct roles. $Q_{\boldsymbol{\tau}}$ determines entry into the qualified cohort, whereas $R_c(M_U, I_0, \mathrm{Id}, 0)$ records target capability under canonical access on the fixed audit set.

We consider two controlled post-qualification slices of this characterization. Holding the model state fixed and varying I examines task-native access dependence:

$$R_c(M_U, I, \mathrm{Id}, 0). \tag{6}$$

Holding canonical access fixed and varying the model state under A examines post-unlearning model evolution:

$$R_c(M_U, I_0, A, t). \tag{7}$$

3 Evaluation Protocol

We now instantiate the static qualification predicate and the two post-qualification characterization slices defined in Section 2. The protocol first identifies qualified model-target states from the State Bank and then evaluates the same states under task-native access variation and target-clean model evolution (Figure 1).

3.1 Static Qualification Protocol

Under the canonical access route I_0, each candidate state M_U is evaluated using the seen-target score T_{seen}, held-out target score T_{held}, and retained non-target utility U_{ret}. For $\boldsymbol{\tau} = (\tau_s, \tau_h, \tau_r)$, we instantiate the qualification predicate from Section 2.2 as

$$Q_{\boldsymbol{\tau}}(M_U; c, I_0) = \mathbf{1}[T_{\text{seen}} \le \tau_s \wedge T_{\text{held}} \le \tau_h \wedge U_{\text{ret}} \ge \tau_r]. \tag{8}$$

We use the pre-specified thresholds $(\tau_s, \tau_h, \tau_r) = (0.50, 0.60, 0.75)$. The binary predicate determines cohort membership, while the underlying continuous scores are retained for analysis.

To instantiate this predicate, we construct a State Bank spanning SCU, the direct segmentation-foundation-model unlearning baseline (Du et al., 2026), together with six standardized forgetting and retention controls. The bank covers two promptable segmentation backbones (SAM1 ViT-L (Kirillov et al., 2023) and the original SAM2 Hiera-L (Ravi et al., 2025)), eight target concepts, and three seeds, yielding $7 \times 2 \times 8 \times 3 = 336$ candidate model-target runs.

Qualification is model-target-state specific. For each run, a pre-specified checkpoint rule selects one candidate checkpoint using only the qualification data. The thresholds, checkpoint-selection rule, and resulting cohort membership are fixed before the post-qualification analyses.

For SAM2, we additionally identify a pre-specified access-diagnostic cohort consisting of states that satisfy the target-suppression and retained-utility criterion under the controlled single-mask route but not under the deployed canonical route. These states are kept separate from the primary qualified cohort. Method-specific training, checkpoint selection, and routing details are provided in the appendix.

3.2 Task-Native Access Characterization

For each qualified model-target state, we instantiate the access slice $R_c(M_U, I, \mathrm{Id}, 0)$ from Section 2.3 while keeping the model state and target audit queries fixed.

We use a frozen audit set $\mathcal{Q}_c^{\text{audit}}$ comprising 893 concept-image pairs over 773 images. Because access routes can differ in segmentation difficulty even in the original model, we measure residual post-unlearning capability relative to the original model under the same route:

$$\text{Res}_c(M_U; I) = \frac{R_c(M_U, I, \text{Id}, 0)}{F_c(M_0; I, \mathcal{Q}_c^{\text{audit}})}, \qquad F_c(M_0; I, \mathcal{Q}_c^{\text{audit}}) > 0. \tag{9}$$

A residual of one matches the original model's target capability under the same route; values are not clipped.

The primary characterization uses three task-native routes. **Box** is the canonical route, using the per-instance box prompt and deployed output-resolution rule. **Point** replaces the box with a deterministic interior positive point and returns a single mask. **MM-top** uses the same point with multiple candidate masks and returns the candidate with the highest model-predicted IoU score. Its selection rule does not use ground-truth overlap (Kirillov et al., 2023).

For SAM2, we additionally isolate the output-resolution component of the access route. The image, box prompt, target query, and model weights are held fixed while only the output-resolution rule changes between a controlled token-0 single-mask route and the deployed route (Ravi et al., 2025). Ground-truth-selected multimask output and deterministic corrective clicking are reserved for secondary diagnostic and stress analyses.

3.3 Target-Clean Model-Evolution Characterization

For the same qualified model-target states, we instantiate the model-evolution slice $R_c(M_U, I_0, A_{\text{tc}}, t)$ from Section 2.3, keeping canonical access fixed. Here A_{tc} denotes the target-clean adaptation protocol specified below.

Each qualified state is adapted under three non-overlapping five-example support episodes, with every episode independently restarted from M_U. The full mask decoder is updated for 50 AdamW steps ($\text{lr} = 10^{-4}$, $\text{wd} = 0.01$), while the image and prompt encoders remain frozen. We define *target-clean* support as support containing no direct target-positive segmentation supervision.

We evaluate on the same fixed target audit set $\mathcal{Q}_c^{\text{audit}}$ used in the access characterization at $t \in \{0, 1, 5, 10, 20, 50\}$. Define the behavioral gap at the post-unlearning starting state as

$$\Delta_c = F_c(M_0; I_0, \mathcal{Q}_c^{\text{audit}}) - R_c(M_U, I_0, \text{Id}, 0), \qquad \Delta_c > 0, \tag{10}$$

and define the *Gap Restoration Ratio* (GRR) as

$$\text{GRR}_c(t) = \frac{R_c(M_U, I_0, A_{\text{tc}}, t) - R_c(M_U, I_0, \text{Id}, 0)}{\Delta_c}. \tag{11}$$

$\text{GRR} = 0$ corresponds to the post-unlearning starting state, and $\text{GRR} = 1$ reaches the original model's target capability under the same canonical access route and audit set. Values are not clipped. We report the raw target-capability trajectory together with GRR.

Target capability is evaluated jointly with retained utility and support-task utility. At step 50, the pre-specified utility guard requires each quantity to decrease by no more than 0.03 relative to its post-unlearning starting value. Because the three support episodes are repeated conditions of the same qualified model-target state, primary summaries first average across episodes within state and then aggregate across states.

Fine-grained extension and matched diagnostic. We separately evaluate five object-part capabilities using SAM1 ViT-L and PASCAL-Part (Kirillov et al., 2023; Chen et al., 2014), with deterministic point access and target-clean adaptation that excludes both target-positive and target-parent supervision. These experiments are not pooled with the State-Bank estimates. For *horse:leg*, we additionally compare the post-unlearning state U with a matched target-clean control C having nearly identical initial target capability. The two states undergo the same 20 frozen adaptation episodes and are evaluated on the same target queries. This comparison is used only as an ordinary-reacquisition diagnostic, not as a retraining oracle or provenance test.

Table 1: **State-Bank qualification identifies 28 of 336 candidate states, 27 produced by SCU.**

	Step-50 means			Canonical state class			Persistence
Method	$\bar{E}_{50}$	$\bar{G}_{50}$	$\bar{R}_{50}$	Qualified	Access	F/U/M	Pass@50/ Qual.
SCU	0.272	0.644	0.886	**27**	9	12/0/0	15/27
Gradient Ascent	0.004	0.014	0.019	0	0	0/48/0	—
Random Label	0.786	0.881	0.905	1	0	47/0/0	0/1
Boundary Shrink	0.622	0.637	0.654	0	1	5/6/36	—
Boundary Expand	0.716	0.685	0.685	0	0	1/0/47	—
L1-Sparse	0.918	0.908	0.907	0	0	48/0/0	—
FT control	0.916	0.907	0.907	0	0	48/0/0	—
Total				**28/336**	10	161/54/83	15/28

4 RESULTS

All primary results are conditioned on the model-target states selected by the static qualification protocol in Section 3.1. We first describe this qualified cohort and then evaluate the two post-qualification characterization slices defined in Section 2.3: task-native access and target-clean model evolution. We subsequently examine fine-grained target behavior and sensitivity of the primary findings.

4.1 STATIC QUALIFICATION DEFINES THE PRIMARY CHARACTERIZATION COHORT

Table 1 summarizes the outcome of the static qualification protocol. Of 336 candidate model-target runs, 28 (8.3%) satisfy the pre-specified qualification criterion. Twenty-seven of these states are produced by SCU (Du et al., 2026) and one by Random Label. Among the remaining runs, 161 fail the target-suppression criterion, 54 fail the retained-utility criterion, 83 fail both criteria, and 10 form the pre-specified access-diagnostic cohort.

The subsequent characterization is therefore conditional on these 28 qualified model-target states, which serve as the primary unit of analysis.

4.2 RESIDUAL TARGET CAPABILITY IS STRONGLY ACCESS-DEPENDENT AFTER QUALIFICATION

Figure 2 reports route-matched residual target capability for the 28 qualified states. The cohort mean increases from 0.548 under **Box** to 0.846 under **Point** and 0.944 under **MM-top**.

The shift is consistent at the state level. All 28 qualified states have higher residual capability under both **Point** and **MM-top** than under **Box**. At a residual threshold of 0.9, none of the 28 states reaches the threshold under **Box**, compared with 13/28 under **Point** and 21/28 under **MM-top** (Figure 2b).

Because each post-unlearning observation is normalized to the original model under the same route, this pattern cannot be explained by route difficulty already present in the original model. Canonical qualification under **Box** therefore does not characterize residual target capability across the other task-native routes examined here.

4.3 FIXED-INPUT SAM2 ISOLATES OUTPUT-RESOLUTION DEPENDENCE

The comparisons in Figure 2 vary the prompting rule as well as, for multimask inference, the output-resolution rule. SAM2 allows the latter component to be isolated. Within each comparison, the image, box prompt, target query, and model weights are fixed; only the output-resolution rule changes between the controlled single-mask and deployed routes (Figure 3a).

For the original SAM2 model, the two routes are nearly indistinguishable: across the eight target concepts, the mean deployed-minus-controlled IoU gap is 0.0013. Across the 10 qualified SAM2

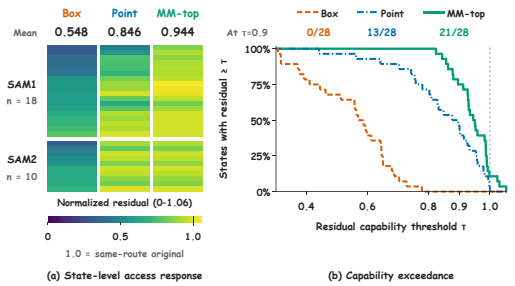


Figure 2: **Residual target capability increases across alternate task-native access routes.** Same-route normalized results are shown for the 28 statically qualified states under Box, Point, and MM-top access.

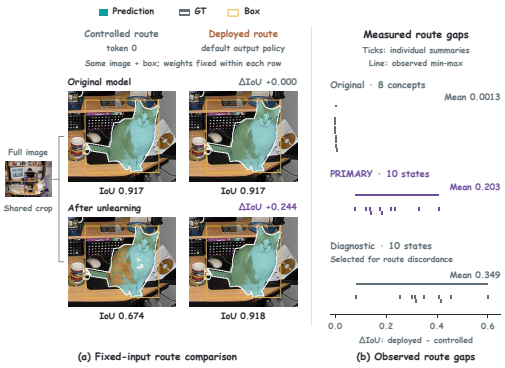


Figure 3: **Changing only SAM2's output-resolution rule reveals substantial post-unlearning disagreement.** The image, box prompt, and model weights are fixed in each comparison.

states, the same fixed-input comparison yields a mean gap of 0.203 (Figure 3b). The separate 10-state access-diagnostic cohort exhibits an even larger mean gap of 0.349.

Thus, substantial route dependence remains even after external prompt variation is removed. A model-side output-resolution distinction that is negligible in the original model becomes pronounced in the qualified post-unlearning states.

4.4 Qualified Target Capability Changes Rapidly under Target-Clean Adaptation

We next evaluate the model-evolution slice while holding canonical access fixed. Across the 28 qualified states, the state-first mean GRR reaches 0.341 after one update and 0.772 after five updates. By step 50, mean GRR reaches 0.896 (95% CI: [0.859, 0.926]), with mean target IoU increasing from 0.500 at the post-unlearning starting state to 0.871, compared with 0.912 for the original-model reference (Figure 4a).

The return is widespread across independently restarted support episodes. Of the 84 trajectories, 76 reach GRR ≥ 0.5 within five updates, and 83/84 satisfy the pre-specified utility guard at step 50 (Figure 4b–c).

Under the specified target-clean adaptation operator, target capability therefore changes rapidly from the statically qualified starting state while monitored retained and support-task utility remain largely preserved.

4.5 Fine-Grained Targets Exhibit Heterogeneous Model-Evolution Behavior

We next examine a separate SAM1/PASCAL-Part extension with five object-part capabilities (Chen et al., 2014) (Figure 5a). Step-50 behavior is strongly heterogeneous across targets and support families (Figure 5a): target return ranges from substantial to modest, can occur without preserved utility, and is absent on average for *aeroplane:body*. Thus, model-evolution behavior is target-dependent rather than uniform across fine-grained capabilities.

Behavioral return alone, however, may include ordinary reacquisition. For *horse:leg*, a matched control begins at nearly the same target capability as the post-unlearning state (0.2592 versus 0.2581 IoU) and undergoes the same target-clean adaptation (Figure 5b). Across 20 frozen episodes, the post-unlearning lineage has greater trajectory-integrated gain in all 20 cases; by step 50 its mean gain is $+0.211$ IoU, compared with -0.009 for the matched control. This diagnostic does not identify the provenance of returning capability, but shows that ordinary reacquisition alone does not account for all return in this controlled case.

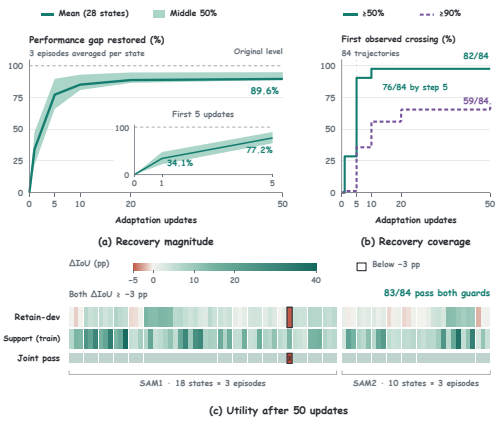


Figure 4: **Behavioral gap restoration is rapid under target-clean adaptation while utility is mostly preserved.** Panels show the restoration trajectory, threshold-crossing coverage, and utility changes after 50 updates.

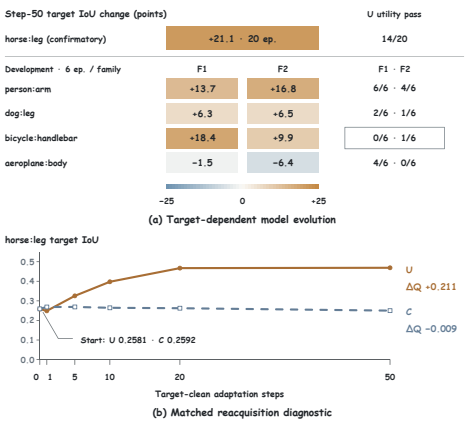


Figure 5: **Fine-grained model evolution varies across targets, and a matched control constrains how behavioral return is interpreted.** Panels show step-50 target/utility outcomes and the matched *horse:leg* trajectory.

Table 2: **Post-qualification access and adaptation results across sensitivity subsets and model strata.**

		Access audit		Target-clean adaptation		
Cohort	States/ concepts	MM-top residual	MM-top − Box	GRR@5	GRR@50	Utility pass
PRIMARY (pre-specified)	28/7	0.944	+0.396	0.772	0.896	83/84
SCU-only	27/7	0.941	+0.394	0.771	0.896	80/81
Persistent-pass	15/5	0.926	+0.406	0.765	0.872	44/45
Strict	20/4	0.950	+0.402	0.759	0.882	59/60
Very-strict	7/3	0.933	+0.416	0.680	0.852	20/21
Model stratification						
SAM1	18/7	0.953	+0.409	0.806	0.912	53/54
SAM2	10/4	0.928	+0.372	0.710	0.868	30/30

4.6 Primary Findings Persist across Sensitivity Analyses

Table 2 examines sensitivity within the primary cohort while retaining each run's original qualified checkpoint. Persistent-pass filters runs by whether step 50 also qualifies, whereas Strict and Very-strict apply tighter qualification criteria to the original qualified checkpoint without checkpoint reselection. Restricting the primary cohort to the 27 SCU states leaves both results nearly unchanged: MM-top residual is 0.941, compared with 0.944 in the full cohort, and GRR@50 remains 0.896. Among the 15 states that also satisfy qualification at the fixed step-50 checkpoint, MM-top residual remains 0.926 and GRR@50 is 0.872; stricter qualification views show the same qualitative pattern.

The findings are also present in both backbone strata. SAM1 and SAM2 have MM-top residuals of 0.953 and 0.928, respectively, and GRR@50 values of 0.912 and 0.868. The utility guard is satisfied by 53/54 SAM1 trajectories and all 30 SAM2 trajectories.

The primary access and model-evolution findings therefore persist across the sensitivity analyses considered here. Together with the fixed-input SAM2 comparison, these results support the central conclusion that canonical static qualification does not determine how the qualified target capability behaves across the audited access routes and model states.

5 Related Work

Machine unlearning spans data deletion, selective forgetting, and class- or concept-level removal (Ginart et al., 2019; Bourtoule et al., 2021; Golatkar et al., 2020b; Neel et al., 2021; Chen et al., 2023; Foster et al., 2024), with recent vision work extending these ideas to structured prediction and concept-level targets(Zhou et al., 2025; Zhu et al., 2026; Wu et al., 2026; Hong et al., 2026; Yu et al., 2026; Cai et al., 2024). Most directly, SCU suppresses segmentation capabilities in SAM-family models while preserving non-target behavior (Du et al., 2026). Our focus differs: rather than asking whether static selectivity can be achieved, we ask how far that qualification extends across native access routes and model states.

Evaluation and alternate access. Unlearning evaluation has expanded beyond a single forgetting score to behavioral, utility, privacy, and robustness criteria(Shi et al., 2025; Li et al., 2024; Zhang et al., 2026; Pawelczyk et al., 2025; Gao et al., 2026; Deeb & Roger, 2024; Du et al., 2024; Ma et al., 2026; Cheng et al., 2026). Promptable segmentation is sensitive to prompt form and inference conditions(Kirillov et al., 2023; Fan et al., 2025; Moskalenko et al., 2026; Chen et al., 2024; Ke et al., 2023; Lin et al., 2025; Zhang et al., 2024b; Yu et al., 2025; Yao et al., 2025; Liu et al., 2026; Zhang et al., 2024a), while concept-erasure work shows that suppressed content may remain accessible through alternative, indirect, or adversarial routes(Tsai et al., 2024; Zhang et al., 2024c; George et al., 2025; Wei et al., 2026; Cywiński & Deja, 2025). Our analysis conditions first on canonical static qualification and then varies task-native routes to the same capability; the fixed-input SAM2 comparison isolates model-side output selection from external prompt variation.

Post-unlearning model evolution. Later model updates can restore behavior suppressed by approximate unlearning, motivating work on relearning resistance and future recoverability(Hu et al., 2025; Gao et al., 2025; Tran & Woo, 2026). Related work also studies continual or online unlearning and erasure(Lee et al., 2026; Zare Zade et al., 2026; Han et al., 2026). Our distinction is to treat target-clean updates as post-qualification model evolution and characterize behavioral change without inferring its provenance.

6 Discussion

Static selectivity provides meaningful evidence about target capability under the model state and access route on which it is measured. Our results show that evaluation scope should match claim scope: suppression may differ across task-native routes, and the qualified capability may change as the model is adapted. Broader access- or lifecycle-level claims therefore require corresponding evidence. We characterize only the routes and target-clean adaptation operator specified here, so conclusions remain conditional rather than exhaustive over all prompts or future updates.

Behavioral return under model evolution also does not identify its own provenance. In a transferable segmentation model, ordinary reacquisition from shared priors is a plausible component of that return. Our matched fine-grained diagnostic shows that this control does not account for the full return in one case, but does not establish what historical information, if any, is responsible for the remaining difference. Systematic counterfactual attribution across targets and adaptation settings remains open.

7 Conclusion

We studied the scope of static qualification in promptable segmentation unlearning. Qualified target capabilities vary substantially across task-native access routes, including a fixed-input SAM2 setting in which a negligible original-model output-resolution gap becomes pronounced after unlearning. They can also change rapidly under target-clean model evolution while monitored utility remains largely preserved. Static selectivity is therefore a meaningful condition-specific qualification, not an access- or state-invariant characterization. The scope of an unlearning claim should match the conditions under which the capability is evaluated.

AI USE STATEMENT

Generative AI tools were used only to assist with literature retrieval and the discovery of related work. References surfaced through this process were checked by the authors against the original sources before being included in the paper. The research methodology, experiments, analysis, and interpretation of the results were carried out by the authors.

ETHICS STATEMENT

This work examines limitations of static evaluations of concept unlearning in promptable segmentation. By showing that target behavior suppressed under a canonical evaluation can remain accessible through other model-supported inference routes or re-emerge after subsequent target-clean adaptation, the analysis may also expose failure modes that could be used to recover behavior regarded as suppressed. We report these behaviors to support more reliable auditing and evaluation of unlearning systems, rather than to provide a mechanism for circumventing deployed safeguards. The experiments use established research benchmarks and pretrained segmentation models and involve no new human-subject data collection.

REPRODUCIBILITY STATEMENT

The experimental protocol underlying the main results is specified in Section 3, including the static qualification procedure, task-native access conditions, target-clean adaptation protocol, and aggregation across repeated runs. Appendix A provides the model and data configurations, optimization settings, standardized unlearning controls, checkpoint-selection rule, and qualification-threshold sensitivity. Appendix B reports the additional access and model-evolution analyses, while Appendix C gives the complete setup and results for the fine-grained object–part extension and matched diagnostic. Together, these sections document the data splits, random-seed structure, training and adaptation settings, evaluation routes, and reporting procedures used for the results in this paper.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Zikui Cai, Yaoteng Tan, and M Salman Asif. Targeted unlearning with single layer unlearning gradient. *arXiv preprint arXiv:2407.11867*, 2024.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7766–7775. IEEE, 2023.

Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4081–4091, 2024.

Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1978, 2014.

Jiali Cheng, Ziheng Chen, Chirag Agarwal, and Hadi Amiri. Toward understanding unlearning difficulty: A mechanistic perspective and circuit-guided difficulty metric. *arXiv preprint arXiv:2601.09624*, 2026.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. *arXiv preprint arXiv:2406.13348*, 2024.

Miaozeng Du, Jiaqi Li, Sirui Pan, Yi Zhan, Guilin Qi, Yuxin Zhang, Rihui Jin, Yinjia Shu, and Qianshan Wei. Forget what has seen: Selective concept unlearning in segmentation foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20923–20931, 2026.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, volume 2024, pp. 53643–53673, 2024.

Qi Fan, Xin Tao, Lei Ke, Mingqiao Ye, Di Zhang, Pengfei Wan, Yu-Wing Tai, and Chi-Keung Tang. Stable segment anything model. In *International Conference on Learning Representations*, volume 2025, pp. 48170–48183, 2025.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2131–2141. IEEE, 2025.

Yichen Gao, Altay Unal, Akshay Rangamani, and Zhihui Zhu. An illusion of unlearning? assessing machine unlearning through internal representations. *arXiv preprint arXiv:2604.08271*, 2026.

Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13393–13402. IEEE, 2025.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309. IEEE, 2020a.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, pp. 383–398. Springer, 2020b.

Ning Han, Zhenyu Ge, Feng Han, Yuhua Sun, Chengqing Li, and Jingjing Chen. Groce: Graph-guided online concept erasure for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43536–43545, 2026.

Kiseong Hong, JungKyoo Shin, and Eunwoo Kim. Stake the points: Structure-faithful instance unlearning. *arXiv preprint arXiv:2603.12915*, 2026.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

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

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. In *International Conference on Learning Representations*, volume 2026, pp. 30660–30682, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

Yuqi Lin, Hengjia Li, Wenqi Shao, Zheng Yang, Jun Zhao, Xiaofei He, Ping Luo, and Kaipeng Zhang. Samrefiner: Taming segment anything model for universal mask refinement. In *International Conference on Learning Representations*, volume 2025, pp. 57368–57392, 2025.

Xueyu Liu, Xiaoyi Zhang, Meilin Liu, Guangze Shi, Jia Shen, Yujie Wang, Cai Zhao, Ziyuan He, Yongfei Wu, Mingqiang Wei, et al. Attack for defense: Adversarial agents for point prompt optimization empowering segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6591–6600, 2026.

Mengyao Ma, Shuofeng Liu, Minhui Xue, Surya Nepal, and Guangdong Bai. Retrace: Reinforcement learning-guided reconstruction attacks on machine unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 33626–33650, 2026.

Andrey Moskalenko, Danil Kuznetsov, Irina Dudko, Anastasiia Iasakova, Nikita Boldyrev, Denis Shepelev, Andrei Spiridonov, Andrey Kuznetsov, and Vlad Shakhuro. Breps: Bounding-box robustness evaluation of promptable segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8107–8115, 2026.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic learning theory*, pp. 931–962. PMLR, 2021.

Martin Pawelczyk, Jimmy Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. In *International Conference on Learning Representations*, volume 2025, pp. 50542–50573, 2025.

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

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, pp. 27797–27818, 2025.

Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. *Advances in Neural Information Processing Systems*, 38:129326–129357, 2026.

Khoa Tran and Simon S Woo. Efficient unlearning through maximizing relearning convergence delay. *arXiv preprint arXiv:2604.09391*, 2026.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, volume 2024, pp. 41543–41554, 2024.

Lu Wei, Yuta Nakashima, and Noa Garcia. Emma: Concept erasure benchmark with comprehensive semantic metrics and diverse categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37247–37257, 2026.

Zhongze Wu, Xiu Su, Feng Yang, Dan Niu, Shan You, Yueyi Luo, and Jun Long. Unlearning without forgetting: Securely removing targeted concepts from large-scale vision-language open-vocabulary detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6271–6281, 2026.

Nakyeong Yang, Dong-Kyum Kim, Jea Kwon, Minsung Kim, Kyomin Jung, and Meeyoung Cha. Erase or hide? suppressing spurious unlearning neurons for robust unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 130808–130824, 2026.

Yuan Yao, Qiushi Yang, Miaomiao Cui, and Liefeng Bo. Towards fine-grained interactive segmentation in images and videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22509–22518. IEEE, 2025.

Chongkai Yu, Ting Liu, Anqi Li, Xiaochao Qu, Chengjing Wu, Luoqi Liu, and Xiaolin Hu. Samref: Introducing image-prompt synergy during interaction for detail enhancement in the segment anything model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19356–19365. IEEE, 2025.

Zhenyu Yu, Yangchen Zeng, Chunlei Meng, Guangzhen Yao, and Shuigeng Zhou. Do vision models truly forget? new findings from representation-level certification of visual unlearning in vertical federated learning. In *European Conference on Computer Vision*, pp. 130–146. Springer, 2026.

Saleh Zare Zade, Xiangyu Zhou, Sijia Liu, and Dongxiao Zhu. Attention smoothing is all you need for unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 19583–19621, 2026.

Chenhao Zhang, Muxing Li, Feng Liu, Weitong Chen, and Miao Xu. Unlearning evaluation through subset statistical independence. *arXiv preprint arXiv:2603.00587*, 2026.

Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23385–23395. IEEE, 2024a.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Gao Peng, and Hongsheng Li. Personalize segment anything model with one shot. In *International Conference on Learning Representations*, volume 2024, pp. 18250–18279, 2024b.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024c.

Yu Zhou, Dian Zheng, Qijie Mo, Renjie Lu, Kun-Yu Lin, and Wei-Shi Zheng. Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20350–20359. IEEE, 2025.

Jianing Zhu, Bo Han, Jiangchao Yao, Jianliang Xu, Gang Niu, and Masashi Sugiyama. Decoupling the class label and the target concept in machine unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 30100–30146, 2026.

A Experimental Details, Baselines, and Qualification Sensitivity

A.1 Models, Data, and Optimization

Table 3 summarizes the State-Bank configuration. We evaluate SAM1 ViT-L (Kirillov et al., 2023) and the original SAM2 Hiera-L (Ravi et al., 2025). The image and prompt encoders are frozen and the full mask decoder is trainable; for SAM2, this includes the decoder's high-resolution projection layers. All methods use AdamW with learning rate 10^{-4}, weight decay 0.01, $\beta = (0.9, 0.999)$, and $\epsilon = 10^{-8}$ unless noted otherwise.

The eight targets are *aeroplane*, *monitor*, *bicycle*, *bottle*, *cat*, *sheep*, *bird*, and *chair*. Retained examples are drawn from a separate candidate pool after filtering images with target-concept exposure, and VOC void pixels (label 255) are excluded from formal IoU computation.

State-Bank training, checkpoint selection, and cohort construction use only the designated training and development data. The 893-pair held-out audit set is used for the post-qualification analyses. The access and model-evolution analyses use the same audit queries.

A.2 Unlearning Methods and Standardized Controls

The State Bank contains SCU (Du et al., 2026), the direct prior method for selective concept unlearning in segmentation foundation models, and six standardized forgetting/retention controls implemented in the same segmentation backend.

SCU. SCU combines negative distillation, target-region attention suppression, target-output suppression, and non-target preservation (Du et al., 2026). We use a local implementation following this training logic.

Standardized controls. Gradient Ascent maximizes BCE–Dice loss on the forget examples. Random Label replaces each forget target with an area-preserving random mask while retained examples use the true mask. Boundary Shrink and Boundary Expand use deterministic eroded and dilated forget masks, respectively. L1-Sparse combines retained-data fine-tuning with a linearly decayed ℓ_1 penalty on decoder parameters with base coefficient 5×10^{-4}. FT performs retain-only BCE–Dice fine-tuning without an explicit forgetting objective.

A.3 Qualification, Checkpoint Selection, and Sensitivity

A candidate model–target state qualifies under the canonical route when

$$T_{\text{seen}} \leq 0.50, \qquad T_{\text{held}} \leq 0.60, \qquad U_{\text{ret}} \geq 0.75. \tag{12}$$

The continuous scores are retained for analysis; the binary criterion is used for cohort routing.

For each of the 336 model–method–concept–seed runs, we select the earliest checkpoint satisfying the full criterion. If none exists, the earliest Access-Fail checkpoint is retained for the separate SAM2 access-diagnostic cohort; otherwise, step 50 is used for failure classification. Under these thresholds, 28 runs form the primary qualified cohort (27 SCU and one Random Label), 10 form the access-diagnostic cohort, and the remaining 298 comprise 161 forgetting failures, 54 utility failures, and 83 mixed failures.

Threshold sensitivity. Table 4 recomputes routing from the stored development trajectories without retraining, varying one threshold at a time while holding the other two fixed. The cohort is unchanged for seen-target thresholds from 0.45 to 0.60 and is more sensitive to the held-out-target and retained-utility criteria.

Joint strictness and checkpoint persistence. For joint settings $(0.40, 0.50, 0.85)$, $(0.45, 0.55, 0.80)$, $(0.50, 0.60, 0.75)$, $(0.55, 0.65, 0.70)$, and $(0.60, 0.70, 0.65)$, the canonical rule selects 16, 21, 28, 36, and 59 states, respectively; the corresponding fixed-S50 counts are 9, 12, 15, 21, and 39. The gap between canonical and fixed-S50 counts shows that qualification at

Table 3: Frozen State-Bank configuration. The held-out audit set is not used for checkpoint selection or cohort construction.

Component	Configuration
Backbones	SAM1 ViT-L; original SAM2 Hiera-L
Trainable parameters	Full mask decoder; image/prompt encoders frozen
Optimizer	AdamW, lr 10^{-4}, wd 0.01
Seeds	3 per model–method–concept configuration
Target concepts	8
Forget train	120 pairs (8×15)
Qualification dev	320 pairs (8×40)
Post-qualification audit	893 pairs / 773 images
Formal overlap metric	IoU with VOC void pixels ignored

Table 4: Qualification-threshold sensitivity from the stored State-Bank development trajectories. "Canonical" uses the checkpoint-selection rule; "S50" evaluates the fixed step-50 checkpoint.

Threshold varied	Value	Canonical qualified	S50 qualified
Seen target	0.40	27	15
	0.45	28	15
	0.50	**28**	**15**
	0.55	28	15
	0.60	28	15
Held-out target	0.50	21	10
	0.55	24	12
	0.60	**28**	**15**
	0.65	32	21
	0.70	38	27
Retained utility	0.65	33	17
	0.70	30	15
	0.75	**28**	**15**
	0.80	26	15
	0.85	21	13

the selected checkpoint need not persist to step 50. The main analysis uses the default setting $(0.50, 0.60, 0.75)$.

B Additional Access and Model-Evolution Results

B.1 Additional Task-Native Access Results

For the access analysis, the qualified model state and target audit queries are fixed while the model-supported route varies. **Box** uses the per-instance box prompt with the deployed output-resolution rule. **Point** uses a deterministic interior positive point and returns a single mask. **MM-top** uses the same point with multimask output and returns the candidate with the highest model-predicted mask-quality score; it does not use ground-truth overlap. These routes follow the promptable and multimask interfaces of SAM-family models (Kirillov et al., 2023; Ravi et al., 2025).

Table 5 also includes MM-oracle, which selects the best mask from the same multimask candidate family using ground-truth overlap, and Click10, a deterministic point-only condition in which each corrective click is placed at the center of the largest false-negative or false-positive error region. No mask-input feedback is used. Click10 uses the ground-truth error map to determine the next click. Among instances for which the original model reaches the corresponding criterion within ten clicks, the interactive elicitation rates are 86.9% at IoU 0.85 and 75.6% at IoU 0.90.

The access effect is present in both backbones (Table 6).

For SAM2, the image, box prompt, target query, and model weights are fixed while the output-resolution rule varies. The controlled route returns the token-0 single-mask output, whereas the

Table 5: Task-native access results for the 28 qualified model-target states. Residual capability is normalized by the original model under the same route. MM-oracle uses ground-truth overlap for candidate selection; Click10 uses ground-truth-guided corrective clicks.

Access condition	Mean residual	Difference from Box
Box	0.548	—
Point	0.846	+0.298
MM-top	0.944	+0.396
MM-oracle	0.950	+0.402
Click10	0.940	+0.392

Table 6: Backbone-stratified access results for the qualified cohort.

Backbone	States	Box	Point	MM-top
SAM1 ViT-L	18	0.544	0.807	0.953
SAM2 Hiera-L	10	0.556	0.915	0.928

deployed route follows the released inference policy used in our evaluation. The mean deployed-minus-controlled IoU gap is 0.0013 in the original model, 0.203 across the 10 qualified SAM2 states, and 0.349 in the 10-state access-diagnostic cohort (Table 7).

B.2 Additional Target-Clean Model-Evolution Results

Each of the 28 qualified states is evaluated under three non-overlapping five-example target-clean support episodes, producing 84 independently restarted trajectories evaluated at $t \in \{0, 1, 5, 10, 20, 50\}$. The three episodes are repeated conditions of the same model-target state, so summaries first average within state and then aggregate across states. Target-clean support contains no direct target-positive segmentation supervision.

The state-first mean GRR at steps $0/1/5/10/20/50$ is $0.000/0.341/0.772/0.851/0.888/0.896$, with a state-level 95% confidence interval of $[0.859, 0.926]$ at step 50. Mean target IoU rises from 0.500 at the post-unlearning starting state to 0.871 at step 50, compared with 0.912 for the original-model reference. Overall, 82 of 84 trajectories reach $\mathrm{GRR} \geq 0.5$, including 76 within five updates; 59 of 84 reach $\mathrm{GRR} \geq 0.9$. At step 50, 83 of 84 trajectories satisfy the joint utility guard on retained and support-task performance.

Return magnitude and support sensitivity vary across concepts (Table 8). No qualified *bird* state is available in this cohort.

The model-evolution effect is present in both backbones: mean GRR@50 is 0.912 for the 18 qualified SAM1 states and 0.868 for the 10 qualified SAM2 states. The utility guard is satisfied by 53 of 54 SAM1 trajectories and all 30 SAM2 trajectories.

C Fine-Grained Extension and Matched Diagnostic

C.1 Fine-Grained Object-Part Setting and Results

We evaluate SAM1 ViT-L on five PASCAL-Part object-part capabilities derived from VOC2010 images (Chen et al., 2014): *horse:leg*, *dog:leg*, *person:arm*, *bicycle:handlebar*, and *aeroplane:body*.

All evaluations use a single positive point placed at the Euclidean-distance-transform center of the target-part mask with single-mask output. The image and prompt encoders are frozen and the full mask decoder is trainable. Target-clean adaptation contains no direct target-positive supervision and additionally excludes the target parent category; for example, *horse:leg* is adapted without direct *horse:leg* or *horse* supervision.

Table 9 reports the complete step-50 results summarized in Figure 5a. Behavior is target dependent: *horse:leg*, *person:arm*, and *dog:leg* show positive return, *bicycle:handlebar* increases with poor util-

Table 7: SAM2 fixed-input output-resolution comparison. The reported quantity is the mean IoU difference between the deployed and controlled routes; image, box prompt, target query, and model weights are fixed within each comparison.

Evaluation group	Units	Mean deployed–controlled gap
Original SAM2	8 concepts	0.0013
Qualified SAM2 states	10 states	0.203
Access-diagnostic SAM2 states	10 states	0.349

Table 8: Concept-level model-evolution results within the primary qualified cohort. "Support spread" summarizes variation across repeated support conditions for states associated with each concept.

Concept	Qualified states	Mean GRR@50	Mean support spread
aeroplane	7	0.936	0.067
bicycle	4	0.732	0.269
bottle	2	0.935	0.114
cat	6	0.932	0.069
chair	3	0.907	0.096
monitor	4	0.880	0.073
sheep	2	0.954	0.030

ity preservation, and *aeroplane:body* shows no sustained positive mean change under either support family.

C.2 Matched Horse Diagnostic

For *horse:leg*, we compare the post-unlearning state U with a matched control C. Both lineages start from the same pretrained SAM1 base B. One lineage undergoes explicit target-positive acquisition to obtain A and then canonical SCU to obtain U; state C is trained from B with matched target-clean control data:

$$B \to A \to U, \qquad B \to C. \tag{13}$$

State C serves as the control for this lineage.

The two states begin at nearly matched target capability and receive the same 20 target-clean adaptation episodes, with identical support, optimizer, update budget, and target evaluation queries within each pair. The post-unlearning lineage has greater trajectory-integrated target gain in all 20 episodes; the corresponding starting values and step-50 gains are reported in Table 10.

Table 9: Fine-grained target behavior at step 50. F1 and F2 denote the two target-clean support families used in the development extension. The utility-pass column reports episodes satisfying the pre-specified utility criterion. *Horse:leg* uses a separate 20-episode confirmatory protocol.

Target	**Step-50 target change**	**Episodes**	**Utility pass**
horse:leg	$+21.1$	20	14/20
person:arm	$+13.7$ / $+16.8$	6 / 6	6/6 / 4/6
dog:leg	$+6.3$ / $+6.5$	6 / 6	2/6 / 1/6
bicycle:handlebar	$+18.4$ / $+9.9$	6 / 6	0/6 / 1/6
aeroplane:body	-1.5 / -6.4	6 / 6	4/6 / 0/6

Table 10: Matched *horse:leg* diagnostic. U and C begin at nearly identical target capability and undergo the same target-clean adaptation.

	Post-unlearning U	**Matched control** C
Initial target IoU	0.2581	0.2592
Adaptation episodes	20	20
Mean step-50 target gain	+0.211	-0.009
Greater integrated gain, paired episodes	20/20 for U over C	