TINY CLICKS TRIGGER LARGE CONSEQUENCES: EXPLORING SENSITIVITY TO SPATIAL-PROMPT NOVELTY IN SAM-BASED PROMPTABLE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Click-based point prompts provide a simple and intuitive interface for promptable segmentation, where additional correct clicks are generally expected to refine the prediction. However, the number of clicks does not necessarily reflect how much new spatial evidence they provide. We term this property Spatial-Prompt Novelty and ask a fundamental question: how sensitive are SAM-based segmentation models to different levels of spatial evidence novelty in additional clicks? We investigate this question across three progressively relaxed regimes: (1) At the Zero-Novelty limit, exact repetition adds no new spatial location yet can reduce mean IoU by up to 22.1 points. (2) Under Controlled Low-Novelty, the failure persists even when all added clicks are distinct but closely spaced, while performance consistently recovers as their spatial dispersion increases. (3) Finally, under Naturalistic Low-Novelty, 89,924 valid human interaction events show that nearby click additions occur in recorded interaction and are associated with smaller immediate gains. Further experiments reveal that seemingly minor differences in the spatial novelty of correct clicks can trigger disproportionately large changes in model behavior, exposing spatial-prompt novelty as an important but overlooked factor in SAM-based promptable segmentation.

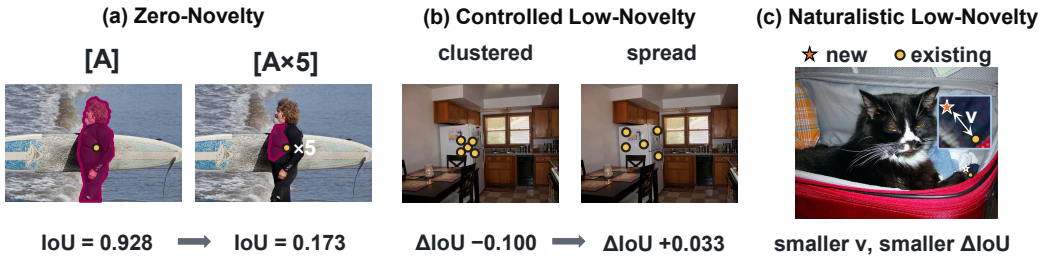


Figure 1: Spatial-Prompt Novelty in three settings. (a) *Zero-Novelty*: exact repetition of one correct click changes SAM3 IoU from 0.928 to 0.173 without adding a new spatial location. (b) *Controlled Low-Novelty*: five distinct correct clicks are compared under clustered and spread configurations; for the illustrated target, ΔIoU changes from -0.100 to $+0.033$. (c) *Naturalistic Low-Novelty*: in a recorded PointPrompt interaction, the new correct positive click (star) lies near existing positive evidence (yellow circle). Here $\nu = \nu_{\text{target}}$ denotes normalized distance to the nearest existing positive point; smaller ν corresponds to lower spatial novelty. This panel is illustrative; Section 5 reports the association between smaller distance and smaller immediate ΔIoU across the corpus.

1 INTRODUCTION

Click-based point prompts provide a simple and intuitive interface for promptable segmentation, where additional correct clicks are generally expected to refine an imperfect prediction (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026; Sofiiuk et al., 2022; Liu et al., 2023; Wang et al., 2025). Prior work has shown that segmentation quality depends on point placement, spatial

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.

distribution, informativeness, and prompt selection or refinement (Quesada et al., 2024b; Liu et al., 2025; 2026; Chen et al., 2025b). These studies change which prompt locations are supplied to the model by adding, moving, or selecting points. However, more correct clicks do not necessarily introduce more new spatial evidence. If added clicks repeat existing locations or cluster around them, prompt count can increase while little new spatial information is added.

We define *Spatial-Prompt Novelty* as the spatial evidence contributed by an additional prompt beyond what is already present. A correct click has low spatial novelty if it repeats an existing point or falls close to existing positive evidence. This leads to our central question: *how sensitive are SAM-based segmentation models to the spatial-prompt novelty of additional correct clicks?*

In this paper, we study Spatial-Prompt Novelty at three levels. *Zero-Novelty* uses exact repetition, where an additional click introduces no new spatial location. *Controlled Low-Novelty* relaxes this limit with distinct but closely spaced correct clicks in controlled spatial configurations. *Naturalistic Low-Novelty* examines recorded human interaction, where a new correct click may fall near positive evidence already present. By *sensitivity*, we mean how strongly the prompt representation, predicted mask, or task utility changes as additional correct clicks provide different amounts of new spatial evidence. Figure 1 illustrates these three cases.

Across COCO and VOC2012, fivefold exact repetition reduces mean IoU by up to 22.1 points for SAM2.1 and SAM3. A comparison of prompt interfaces shows that identical repeated evidence can enter the model differently. The effect also persists when all five correct clicks are distinct but closely spaced, while utility improves as their spatial spread increases. Nearby additions likewise arise as interaction state evolves in a stateful sequential study. Among 89,924 valid PointPrompt additions, nearby human clicks are associated with smaller immediate gains. These results show that the effect of additional correct clicks depends not only on how many are added but also on how much new spatial evidence they provide.

Our contributions are threefold:

- We define *Spatial-Prompt Novelty* to distinguish prompt count from the new spatial evidence added by correct clicks. Exact repetition provides the zero-novelty limit, where *Duplicate Consistency* separates responses at the representation, output, and utility levels.
- We establish controlled sensitivity in the zero-novelty and controlled low-novelty settings. The effect persists after duplicate coordinates are removed: five distinct correct clicks can still underperform a single click when closely spaced, while utility improves as their spatial spread increases.
- We connect the controlled findings to interaction settings through stateful sequential correction and human PointPrompt trajectories, where nearby additions are associated with smaller immediate gains.

2 Background and Related Work

2.1 Promptable and Interactive Segmentation

Promptable segmentation includes the SAM family and broader interfaces for points, boxes, masks, and semantic prompts (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026; Ke et al., 2023; Zou et al., 2023; Li et al., 2024; Chen et al., 2025a; Liu et al., 2024). Interactive methods use successive clicks to select and refine a target (Xu et al., 2016; Maninis et al., 2018; Jang & Kim, 2019; Sofiiuk et al., 2020; Lin et al., 2020; Chen et al., 2021; Sofiiuk et al., 2022; Chen et al., 2022; Liu et al., 2023). Recent systems also carry masks, focus regions, queries, or other state across rounds (Wei et al., 2023; Du et al., 2023; Rana et al., 2023; Huang et al., 2023; Lee et al., 2024; Wang et al., 2025; Zhang et al., 2025). These systems motivate evaluating not only final accuracy but also how predictions respond to successive prompts.

2.2 Prompt Selection, Diversity, and Information

Prior work establishes that point placement, spatial coverage, diversity, redundancy, and prompt quality matter. PointPrompt records human sequential prompting trajectories and corresponding

SAM outputs (Quesada et al., 2024a). Benchmarking Human and Automated Prompting compares human prompting with several automated spatial sampling strategies (Quesada et al., 2024b), and Stable-SAM evaluates multiple randomly selected positive points (Fan et al., 2025). Methods such as PPO, Attack for Defense, and AoP-SAM move, select, suppress, or filter point prompts (Liu et al., 2025; 2026; Chen et al., 2025b). Our study separates click count from added spatial evidence. Exact repetition provides a fixed-state counterfactual with zero new spatial support; later experiments relax this condition to distinct but closely spaced configurations.

2.3 Consistency and Robustness of Prompt Interfaces

SAM robustness studies examine prompt perturbations, degraded inputs, and inference choices (Wang et al., 2024; Chen et al., 2024). Behavioral testing uses controlled invariance checks to expose failures hidden by aggregate metrics (Ribeiro et al., 2020). This motivates our zero-novelty test under a fixed state, where the image, target, label, interaction step, and state remain fixed while prompt multiplicity changes.

3 Spatial-Prompt Novelty: Definition and Experiments

3.1 Defining Spatial-Prompt Novelty

We now formalize *Spatial-Prompt Novelty*. Prompt count alone does not determine novelty: adding a click may introduce a new location, repeat an existing one, or add a nearby location that contributes little new spatial evidence.

Let a point prompt be $p = (x, y, \ell)$, with image coordinate (x, y) and label ℓ. We describe a prompt configuration as a multiset

$$\mathcal{P} = \{\{p_1, \ldots, p_m\}\}, \tag{1}$$

while allowing the model to encode the supplied prompts internally as an ordered tensor. Its unique spatial support is

$$\mathrm{supp}_{xy}(\mathcal{P}) = \{(x, y) : (x, y, \ell) \in \mathcal{P}\}. \tag{2}$$

For an exact duplicate configuration $\mathcal{P}^{(m)} = \{\{p, \ldots, p\}\}$, we have $|\mathcal{P}^{(m)}| = m$ but $|\,\mathrm{supp}_{xy}(\mathcal{P}^{(m)})| = 1$. Exact repetition is therefore the strict zero-novelty limit. Distinct coordinates increase spatial support, but closely spaced coordinates can still provide limited new evidence. Beyond exact repetition, novelty is described by controlled spatial spread or normalized distance, depending on the setting; we do not impose a single scalar score across all three levels.

3.2 Sensitivity and Consistency

Sensitivity describes how model behavior changes as the novelty of additional correct clicks changes. Let I be an image, S an interaction state, $E_\theta(\mathcal{P})$ the model-facing prompt representation, and

$$\hat{Y} = F_\theta(I, \mathcal{P}; S) \tag{3}$$

the predicted mask. We consider three response levels: representation, output, and utility.

The zero-novelty limit gives the strictest form of this analysis because the image, coordinate, label, interaction step, and state can all be held fixed while only prompt multiplicity changes. We call stability under this intervention *Duplicate Consistency*:

$$\text{Representation Consistency: } E_\theta(\mathcal{P}^{(m)}) = E_\theta(\mathcal{P}^{(1)}), \tag{4}$$

$$\text{Output Consistency: } F_\theta(I, \mathcal{P}^{(m)}; S) = F_\theta(I, \mathcal{P}^{(1)}; S), \tag{5}$$

$$\text{Utility Consistency: } U(F_\theta(I, \mathcal{P}^{(m)}; S), Y) = U(F_\theta(I, \mathcal{P}^{(1)}; S), Y), \tag{6}$$

where U is a deterministic task utility such as target-mask IoU. These levels need not coincide: a changed prompt representation can still produce the same mask, and different masks can have similar IoU. They also organize the analysis as novelty is relaxed. Exact equality is meaningful at Zero-Novelty; the controlled and naturalistic settings instead examine how output and utility change as additional clicks contribute more or less new spatial evidence. Representation analysis

is used separately to explain how prompt interfaces handle exact repetition. Duplicate Consistency is distinct from permutation invariance: an order-invariant system can still react to duplicated set elements or cardinality (Zaheer et al., 2017; Lee et al., 2019).

3.3 Three Levels of Spatial-Prompt Novelty

The experiments move from the strict zero-novelty limit to controlled relaxations and then to recorded interaction.

Zero-Novelty. Exact repetition compares one correct click with multiple copies of the same coordinate and label while the image and prompt state are held fixed. Because no new spatial location is introduced, this setting directly tests Duplicate Consistency at the representation, output, and utility levels. We use each SAM generation's released inference path and analyze how repeated evidence is represented and propagated to the reported mask.

Controlled Low-Novelty. We relax exact repetition in two steps. First, one duplicate in $[A, A, A, A, A]$ is replaced by a distinct correct point B whose distance from A increases across predefined bands. Second, all duplicate coordinates are removed: five correct points $[A, B, C, D, E]$ remain distinct while their spatial spread is varied from clustered to dispersed configurations. Both studies use paired targets, a common single-mask contract, fresh state, and the same frozen coordinates across models. These configurations vary several geometric properties together and are treated as controlled spatial comparisons. The two steps test whether the response persists as exact duplication is weakened and then removed.

Naturalistic Low-Novelty. Recorded human prompting provides the naturalistic setting. Before that analysis, a stateful sequential study serves as a bridge from the controlled setting by testing whether nearby additions also arise when interaction state evolves. The sequential study uses prediction-normalized distance ν_{pred}, whereas PointPrompt uses target-normalized distance ν_{target} to the nearest prior positive point. The two normalizations answer related but different questions and are analyzed separately. The human analysis relates normalized distance to the immediate change in IoU.

Full inference, eligibility, construction, and statistical details are provided in the appendix.

4 Controlled Sensitivity to Spatial-Prompt Novelty

4.1 Zero-Novelty: Exact Repetition

Across the full populations, fivefold repetition reduces mean IoU for SAM2.1 and SAM3 on both datasets, with the largest drop reaching 22.13 points for SAM3 on VOC2012 (Table 1). All four image-cluster 95% confidence intervals lie below zero. SAM1 is substantially more stable: its COCO mean changes slightly upward, while the VOC2012 interval includes zero. Figure 2 summarizes these endpoint effects across datasets and inference contracts. Repeated prompts therefore change model behavior even though the direction and magnitude of the IoU effect differ across SAM generations. A consistency violation can occur without a utility loss.

Output changes are much more widespread than utility changes. In the predefined A_0 validation subset, at least 99.7% of five-copy selected masks differ pixelwise from their one-point counterparts, while many changed masks remain within $|\Delta\mathrm{IoU}| \leq 0.01$ (Figure 3). Output Consistency and Utility Consistency measure different types of change. Utility alone would therefore miss many output-level violations: repeated prompts can alter the predicted mask even when target IoU changes only slightly.

4.2 Where Does the Change Enter the Prediction Pipeline?

The official selected mask reflects both candidate generation and predicted-quality selection. Exact repetition changes more than the selector: the oracle-best returned candidate, chosen retrospectively by ground-truth IoU, degrades in all six model–dataset settings. SAM1's relatively stable official

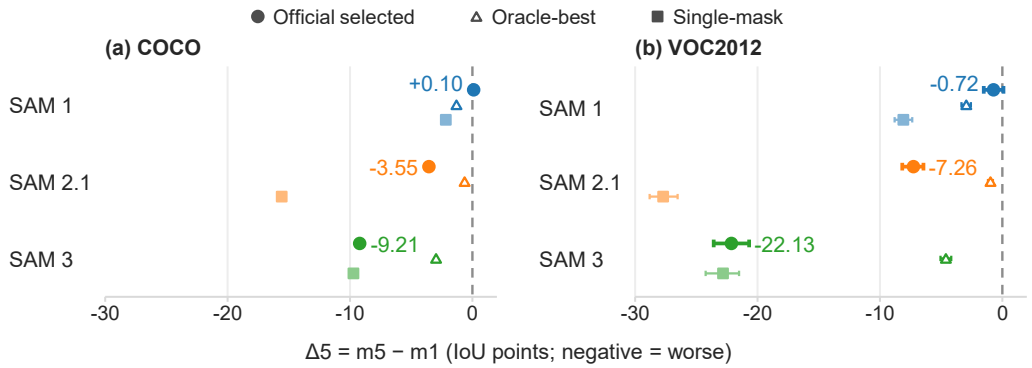


Figure 2: Population-scale utility under exact repetition on COCO and VOC2012. Points show mean $\Delta_5 = m_5 - m_1$ with image-cluster 95% CIs for the official selected output, oracle-best candidate, and forced single-mask path. Negative values indicate worse IoU after fivefold repetition; oracle-best degradation shows that the effect is not solely due to candidate selection.

output occurs despite worse available candidates. For SAM2.1 and SAM3, selection behavior further exposes the degradation. Appendix Table 8 reports the corresponding selection-regret accounting.

The measured magnitude also depends on the inference contract. Forcing single-mask output makes repetition more harmful in all six paired comparisons, reaching -15.58 and -27.71 IoU points for SAM2.1 on COCO and VOC2012. Restoring the released SAM2.1 postprocessing narrows the gap without removing it (Appendix Table 12). The exact intervention stays fixed, but the observed effect changes with the way candidates are produced and converted into the reported mask.

4.3 How Prompt Interfaces Handle Exact Repetition

The same visible click need not produce the same computation. Under the main SAM setting, repeated coordinates remain distinct sparse point tokens. Fixed-state SimpleClick (Liu et al., 2023) instead forms same-sign click maps by minimum distance, so repeating the exact coordinate leaves the model input map unchanged. Across 300 targets, all 1,200 comparisons for $m = 2, 3, 4, 5$ produce identical evidence maps, logits, probabilities, and masks. Sequential SimpleClick processes the same visible point with updated interaction state, and 299/300 masks change. Figure 4 summarizes how the interfaces differ.

Table 1: Exact-repetition results on COCO and VOC2012. Values are IoU-point changes from $m = 1$ to $m = 5$; brackets denote image-cluster 95% CIs. Harmed is $P(\Delta_5 < 0)$. Oracle-best reports the best returned candidate by ground-truth IoU; Single reports the forced single-mask path.

Model	Official		Harmed (%)	Oracle-best		Single Δ_5
	Δ_5	[95% CI]		Δ_5	[95% CI]	
COCO						
SAM1	$+0.10$	$[+0.03, +0.17]$	49.5	-1.30	$[-1.33, -1.27]$	-2.17
SAM2.1	-3.55	$[-3.65, -3.47]$	58.7	-0.64	$[-0.66, -0.61]$	-15.58
SAM3	-9.21	$[-9.34, -9.09]$	65.4	-2.97	$[-3.01, -2.93]$	-9.72
VOC2012						
SAM1	-0.72	$[-1.48, +0.00]$	47.1	-2.95	$[-3.34, -2.57]$	-8.09
SAM2.1	-7.26	$[-8.09, -6.46]$	52.4	-0.97	$[-1.17, -0.77]$	-27.71
SAM3	-22.13	$[-23.44, -20.73]$	67.1	-4.61	$[-5.05, -4.17]$	-22.82

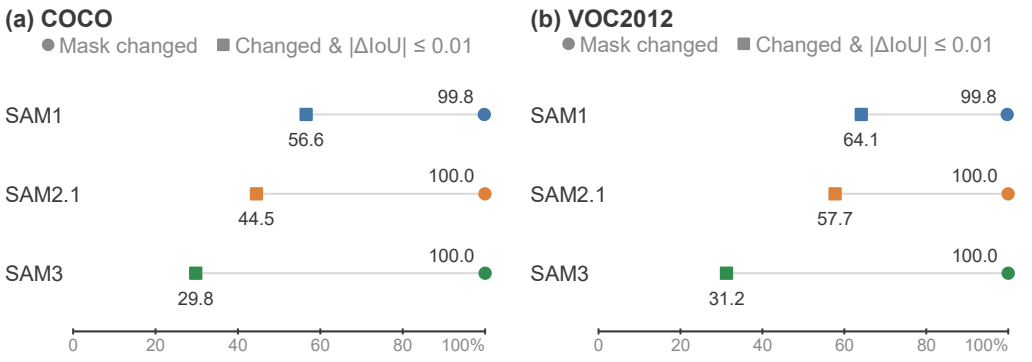


Figure 3: Output consistency under exact repetition on the predefined A_0 validation subset of COCO and VOC2012. (a) COCO; (b) VOC2012. Circles show the fraction of selected masks that change after fivefold repetition, while squares show the joint fraction that both change and satisfy $|\Delta\text{IoU}| \leq 0.01$. Output changes are nearly universal, although many remain small in utility.

For fixed image I and state S, write $F_\theta(I, \mathcal{P}; S) = G_\theta(I, E_\theta(\mathcal{P}); S)$ with deterministic G_θ. A duplicate-insensitive representation is then sufficient for output consistency under exact repetition:

$$E_\theta(\mathcal{P}^{(m)}) = E_\theta(\mathcal{P}^{(1)}) \implies F_\theta(I, \mathcal{P}^{(m)}; S) = F_\theta(I, \mathcal{P}^{(1)}; S). \tag{7}$$

The converse need not hold. As a constructive check, canonicalizing $[A \times m]$ to $[A]$ before prompt encoding restores SAM's $m = 1$ behavior for all 300 tested targets. This explains why exact duplicates can behave differently across interfaces. Both controls show that duplicate handling at the prompt-representation stage matters: collapsing duplicate multiplicity restores the one-click output under a fixed state. The following experiments move from a two-point duplicate boundary case to progressively relaxed and finally all-distinct configurations.

4.4 Relaxing the Zero-Novelty Limit

A two-point boundary test asks whether the effect appears after adding only one duplicate. On the predefined 300-target COCO single-mask cohort, the $[A] \to [A, A]$ comparison changes 97.3–100% of masks, but only SAM2.1 has a 95% confidence interval below zero; the SAM1 and SAM3 intervals include zero (Appendix Table 5). The controlled relaxation below therefore uses five prompts.

Utility recovers as the distinct point B moves away from A. On the common 195-target cohort, the D6-minus-D0 recovery exceeds 20 IoU points for all three models and reaches 29.23 points for SAM2.1. Four exact copies of A remain in every condition, so this relaxation still contains duplicate

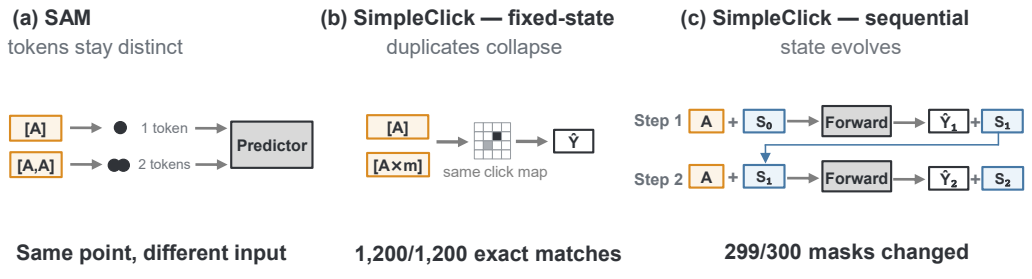


Figure 4: Different prompt interfaces can process the same repeated evidence differently. (a) SAM preserves repeated coordinates as distinct sparse tokens. (b) Fixed-state SimpleClick collapses exact same-sign duplicates and yields 1,200/1,200 exact matches. (c) Sequential SimpleClick advances interaction state and changes 299/300 masks.

coordinates. Substantial recovery already occurs while four copies of A remain, so exact duplication alone cannot explain the change.

4.5 All-Distinct Closely Spaced Prompts

The same sensitivity remains after duplicate coordinates are removed. On the paired 194-target cohort, all three models are evaluated from the clustered C1 condition to the more dispersed C6 condition; the construction is detailed in Appendix B.1.

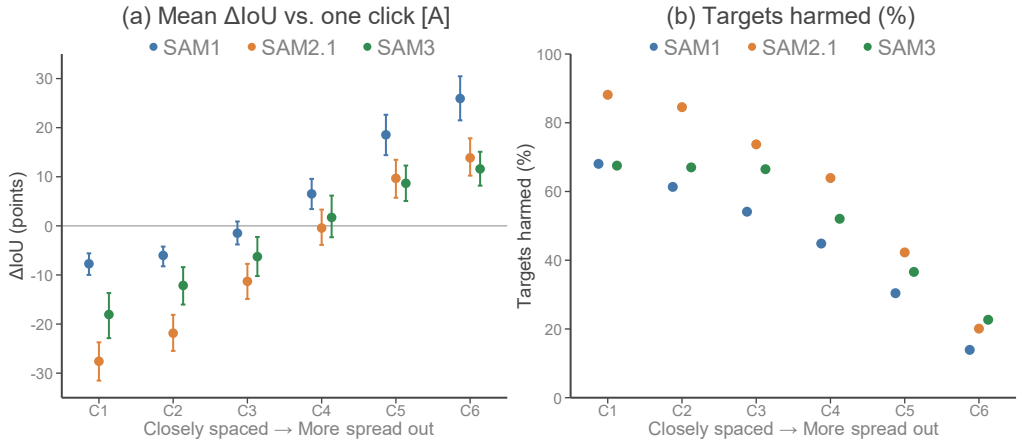


Figure 5: All-distinct recovery under the controlled single-mask contract. Each condition contains five correct, mutually distinct positive points on the same 194 targets that satisfy the geometry constraints. (a) Mean ΔIoU relative to $[A]$ with 95% CIs; (b) fraction of harmed targets. From C1 to C6, utility rises and harm falls across all three SAM generations.

Figure 5 shows the full recovery trajectory. Mean ΔIoU increases monotonically from C1 to C6 for all three models: the mean change is negative for every model at C1 and positive for every model at C6. For SAM2.1, for example, mean ΔIoU changes from -27.58 points at C1 to $+13.86$ points at C6. Over the same sequence, the harmed fraction falls from 67.5–88.1% to 13.9–22.7%. Because all five coordinates are distinct, this trend cannot be attributed to exact duplicate multiplicity. Appendix Table 4 reports the exact C1–C6 values.

5 Spatial-Prompt Novelty in Interaction

The controlled studies above use fresh prompt states. We now turn to stateful and recorded interaction.

5.1 Stateful Bridge: Controlled Sequential Interaction

On VOC2012, a GT-oracle policy runs for up to 20 rounds on 300 targets per model, following standard click-based correction protocols (Sofiiuk et al., 2022; Liu et al., 2023; Chen et al., 2022; Wang et al., 2025). Using the prediction-normalized distance ν_{pred}, introduced in Section 3 and defined in Appendix A.3, the threshold $\nu_{\text{pred}} < 0.08$ was selected on a separate DEV set and fixed before this analysis. Before first reaching IoU 0.90, nearby events account for 30.3–35.6% of eligible positive events across the three SAM generations. Because the simulated clicker does not revisit previously clicked pixels, these are distinct nearby additions rather than exact duplicates. In the adjusted comparison within this protocol, these events are associated with lower marginal ΔIoU and a higher probability of a harmful update (Figure 6; Appendix Tables 13 and 14). Appendix C reports a sensitivity analysis using continuous distance. The reported rates apply to this controlled policy, not to human clicking.

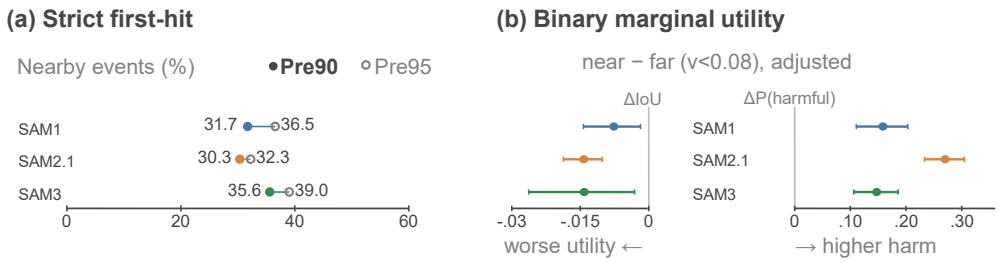


Figure 6: Nearby prompting under controlled VOC2012 sequential interaction. Here $\nu = \nu_{\text{pred}}$ is the normalized distance to the nearest prior positive point. (a) Occurrence before first reaching IoU 0.90/0.95. (b) Adjusted differences between smaller and larger normalized distance in marginal ΔIoU and the probability of a harmful update.

5.2 Naturalistic Low-Novelty: Human Interaction

PointPrompt provides recorded trajectories for the naturalistic setting (Quesada et al., 2024a). Eligible events are adjacent timestep pairs that add exactly one correct positive point, preserve all previous positives, leave the negative set unchanged, and place the new point inside the target. This yields 89,924 valid events. Using the target-normalized distance ν_{target}, introduced in Section 3 and defined in Appendix D.1, 21.14% of these events satisfy $\nu_{\text{target}} \leq 0.04$, showing that nearby additions occur beyond exact duplicates. Exact repeats account for only 3.06% of valid events, so most events within this low-distance range are distinct nearby additions rather than repeated coordinates. Appendix D reports the full distance distribution.

Immediate utility. The primary comparison defines Near as $0 < \nu_{\text{target}} \leq 0.02$ and Reference as $0.04 < \nu_{\text{target}} \leq 0.16$. After stratifying by dataset, annotator, interaction round, number of previous positives, and current IoU bin, a trajectory-cluster bootstrap estimates that the immediate IoU gain for Near additions is 0.97 points lower (95% CI: 0.70–1.25 points lower). Appendix D reports the adjusted sample details and regression check. This association is observational, not causal.

6 Discussion and Limitations

6.1 Implications

Evaluations based on click count should also report how added prompts change the spatial evidence supplied to the model. Exact repetition shows why this matters: the count increases while spatial support does not. Evaluations should also distinguish consistency tests under a fixed state from stateful interactive refinement.

The results for exact repetition also expose an interface design choice. SAM preserves repeated coordinates as separate sparse tokens, while SimpleClick under a fixed state maps exact duplicates with the same sign to one evidence representation. In our controlled test, canonicalizing exact SAM duplicates to one copy restores the behavior with one point on all tested targets. This addresses the zero-novelty case, but not the closely spaced all-distinct result. An interface can detect exact duplicates and collapse them before the model processes them separately when duplicate count has no intended meaning.

Duplicate handling alone is not enough. Performance can still drop when every click is unique, and nearby additions also arise in the sequential and PointPrompt analyses.

6.2 Scope and Limitations

Exact repetition is a controlled limiting case rather than a claim about typical user behavior. Its value is that spatial support, interaction step, and state remain fixed while multiplicity alone changes.

The all-distinct result is a paired comparison within the controlled single-mask setting, not a population estimate over all object scales or inference procedures. Its 194 targets satisfy geometry constraints that favor larger objects, and the constructed C1–C6 sequence changes several geometric properties together. We interpret it as a comparison of spatial configurations, not as the causal effect of distance alone.

The PointPrompt analysis is observational. Available metadata do not fully specify the model configuration behind every released score, so the result is limited to an association between nearby additions and immediate utility.

7 Conclusion

Spatial-Prompt Novelty separates how many correct clicks are provided from how much new spatial evidence they add. At the zero-novelty limit, exact repetition can substantially change SAM predictions and reduce utility. In the controlled all-distinct study, the effect persists with distinct but closely spaced clicks and weakens as their spatial spread increases. Sequential and human analyses further show that nearby additions also arise during interaction, with nearby human clicks associated with smaller immediate gains. Prompt count alone therefore does not specify the spatial evidence available to a promptable segmentation model.

AI USE STATEMENT

Generative AI tools were used to assist with code development, literature organization, feedback on experiments, and manuscript drafting and editing. All AI-assisted material included in the final work was reviewed and verified by the authors against the underlying experiments, code, results, and cited sources. The authors take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The main paper studies zero-novelty and controlled low-novelty settings, followed by a stateful sequential bridge and the naturalistic PointPrompt analysis. Duplicate Consistency defines the fixed-state test at the zero-novelty limit. The appendix documents datasets, eligibility rules, checkpoints, inference settings, resampling, controlled constructions, state handling, and the statistical details of the sequential and PointPrompt analyses.

REFERENCES

Anton Antonov, Andrey Moskalenko, Denis Shepelev, Alexander Krapukhin, Konstantin Soshin, Anton Konushin, and Vlad Shakhuro. RClicks: Realistic click simulation for benchmarking interactive segmentation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 127673–127710. Curran Associates, Inc., 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Pengfei Chen, Lingxi Xie, Xinyue Huo, Xuehui Yu, Xiaopeng Zhang, Yingfei Sun, Zhenjun Han, and Qi Tian. SAM-CP: Marrying SAM with composable prompts for versatile segmentation. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/34c34987232565ca726b99a9304bb144-Abstract-Conference.html.

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 (CVPR)*, pp. 4081–4091, 2024.

Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7345–7354, 2021.

Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1300–1309, 2022.

Yi Chen, Muyoung Son, Chuanbo Hua, and Joo-Young Kim. AoP-SAM: Automation of prompts for efficient segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2284–2292, 2025b. doi: 10.1609/aaai.v39i2.32228. URL https://ojs.aaai.org/index.php/AAAI/article/view/32228.

Fei Du, Jianlong Yuan, Zhibin Wang, and Fan Wang. Efficient mask correction for click-based interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22773–22782, 2023. doi: 10.1109/CVPR52729.2023.02181.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/78288ef33b18a351c3cd679dc9a15c8d-Abstract-Conference.html.

You Huang, Hao Yang, Ke Sun, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, and Rongrong Ji. InterFormer: Real-time interactive image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22301–22311, 2023.

Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5306, 2019.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Chaewon Lee, Seon-Ho Lee, and Chang-Su Kim. MFP: Making full use of probability maps for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4051–4059, 2024.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, 2024. URL https://eccv.ecva.net/virtual/2024/poster/2265.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/90800f46f84381b7891e1378ee850013-Abstract-Conference.html.

Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13339–13348, 2020.

Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. SimpleClick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22290–22300, 2023. doi: 10.1109/ICCV51070.2023.02037.

Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3773–3782, 2024. doi: 10.1109/CVPR52733.2024.00362.

Xueyu Liu, Rui Wang, Yexin Lai, Guangze Shi, Feixue Shao, Fang Hao, Jianan Zhang, Jia Shen, Yongfei Wu, and Wen Zheng. Plug-and-play PPO: An adaptive point prompt optimizer making SAM greater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4332–4342, 2025. doi: 10.1109/CVPR52734.2025.00409.

Xueyu Liu, Xiaoyi Zhang, Meilin Liu, Guangze Shi, Jia Shen, Yujie Wang, Cai Zhao, Ziyuan He, Yongfei Wu, Mingqiang Wei, and Yongle Chen. 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 (CVPR)*, pp. 6591–6600, 2026.

Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 616–625, 2018.

Stephen Price, Elke A. Rundensteiner, and Danielle L. Cote. Mastering SAM prompts: A large-scale empirical study in segmentation refinement for scientific imaging. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=cWcTQMpqv6.

Jorge Quesada, Mohammad Alotaibi, Mohit Prabhushankar, and Ghassan Alregib. PointPrompt: A multi-modal prompting dataset for segment anything model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1604–1610, 2024a. doi: 10.1109/CVPRW63382.2024.00167.

Jorge Quesada, Zoe Fowler, Mohammad Alotaibi, Mohit Prabhushankar, and Ghassan AlRegib. Benchmarking human and automated prompting in the segment anything model. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1625–1634, 2024b. doi: 10.1109/BigData62323.2024.10825731. URL https://arxiv.org/abs/2410.22048.

Amit Kumar Rana, Sabarinath Mahadevan, Alexander Hermans, and Bastian Leibe. DynaMITe: Dynamic query bootstrapping for multi-object interactive segmentation transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1043–1052, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45c1f6a8cbf2da59ebf2c802b4f742cd-Abstract-Conference.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. F-BRS: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8623–8632, 2020.

Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 3141–3145. IEEE, 2022. doi: 10.1109/ICIP46576.2022.9897365.

Bin Wang, Anwesa Choudhuri, Meng Zheng, Zhongpai Gao, Benjamin Planche, Andong Deng, Qin Liu, Terrence Chen, Ulas Bagci, and Ziyan Wu. Order-aware interactive segmentation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9d7ec703f135a77b3be95e0f2548afb-Abstract-Conference.html.

Yuqing Wang, Yun Zhao, and Linda Petzold. An empirical study on the robustness of the Segment Anything Model (SAM). *Pattern Recognition*, 155:110685, 2024. doi: 10.1016/j.patcog.2024.110685.

Qiaoqiao Wei, Hui Zhang, and Jun-Hai Yong. Focused and collaborative feedback integration for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18643–18652, 2023.

Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S. Huang. Deep interactive object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 373–381, 2016.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Chenyi Zhang, Ting Liu, Xiaochao Qu, Luoqi Liu, Yao Zhao, and Yunchao Wei. NTClick: Achieving precise interactive segmentation with noise-tolerant clicks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8921–8930, 2025.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Experimental Setup and Reproducibility

A.1 Models, checkpoints, and inference procedures

We evaluate fixed checkpoints of SAM1, SAM2.1, and base SAM3. SAM1 uses ViT-H with `sam_vit_h_4b8939.pth` (SHA256 `a7bf3b02f3ebf1267aba913ff637d9a2d5c33d3173bb679e46d9f338c26f262e`); the vendored `segment_anything` source used by this project has no git metadata or recorded package version, so we do not assign one. We instead identify that source snapshot by a SHA256 digest over the sorted Python source files (`a98e89ebf7474e953c7f208a383bb3f3e315803a3618b2eae115394626806d3a`); the directly relevant `predictor.py` and `modeling/prompt_encoder.py` hashes are `2f3f34e5437e79d2424d796e1b06838b466b5e17af2b0139023202b7ad24ca65` and `bd8f2f5903f0647f55c21c608bcbb334eec2901ebe76aa41061b1c513df4a06f`, respectively. SAM2.1 uses `sam2.1_hiera_large.pt` (SHA256 `2647878d5dfa5098f2f8649825738a9345572bae2d4350a2468587ece47dd318`) with `sam2.1_hiera_l.yaml` from revision `2b90b9f5ceec907a1c18123530e92e794ad901a4`. SAM3 uses the base `sam3.pt` checkpoint from the official repository (`https://github.com/facebookresearch/sam3`) with SHA256 `9999e2341ceef5e136daa386eecb55cb414446a00ac2b55eb2dfd2f7c3cf8c9e` and revision `8f0b7f4d4e7eda2ed606ebde6702c93359ad01da`; it is not the SAM3.1 multiplex checkpoint.

The main SAM setting evaluates exact repetition through the official native-multimask path with fresh-state positive-point prompting. Each multiplicity condition is evaluated independently and inherits no prior prompt set, mask input, logits, memory, or iterative state. Image-feature caches may be reused. Each backend returns its native candidate set and selects the output with the highest predicted quality score. Across all 1,657,014 recorded calls on this path, covering three models, both datasets, and $m \in \{1, 5\}$, every call returned exactly three candidates. On a fixed cohort of 3,000 model–target pairs, this explicit path and the official default call produced identical scores and selected masks for both $m = 1$ and $m = 5$ (6,000 comparisons, zero mismatches). The binary output is the selected backend mask with threshold > 0 when a non-boolean tensor is returned. Utility is target-mask IoU.

The two-point, D0–D6, and C1–C6 studies use the same frozen model backends under a separate controlled single-mask contract. Every condition is an independent call with fresh state, `multimask_output=False`, exactly one returned mask, no candidate ranking, and no mask input, logits, memory, or carried interaction state. SAM2.1 uses `apply_postprocessing=False`. Within each study, SAM1, SAM2.1, and SAM3 receive the same target set and frozen prompt coordinates. A dedicated SAM2.1 sensitivity analysis restores `apply_postprocessing=True`, including the released dynamic-multimask stability fallback. An independent logit check confirmed that thresholding captured logits at zero reproduces the corresponding binary masks.

A.2 DATASETS, POPULATIONS, AND STATISTICAL UNITS

COCO uses `val2014` with `instances_val2014.json`; VOC2012 uses the official segmentation `val` list with instance masks from `SegmentationObject` and classes from `SegmentationClass`. For each eligible target instance, the primary anchor A is the pixel maximizing the Euclidean distance transform inside the ground-truth mask; ties use the first row-major maximum. The construction is deterministic and uses one anchor per target.

For COCO, exact-valid targets exclude crowd annotations, empty/invalid segmentations, non-positive boxes or areas, decode failures, masks with fewer than 64 foreground pixels, and invalid deepest-interior anchors. For VOC2012, instance IDs must be valid, masks must contain at least 256 pixels, the majority semantic class must be one of the 20 VOC classes, and the anchor depth must be at least 1.5 pixels. The resulting exact-valid populations contain 272,977 COCO targets from 40,120 images and 3,192 VOC2012 targets from 1,442 images. Table 2 reports pipeline-first attrition.

Table 2: Exact-valid population attrition using pipeline-first attribution: a candidate failing multiple rules is charged only to its first failing stage. Zero-trigger branches are retained to reflect the implemented eligibility logic.

Dataset	Pipeline stage / rule	Before	Excluded	Remaining
COCO	Raw universe: `instances_val2014.json` annotations	291,875	–	291,875
COCO	`iscrowd==1`	291,875	3,460	288,415
COCO	Empty segmentation	288,415	0	288,415
COCO	Non-positive bbox width/height	288,415	1	288,414
COCO	`area<=0`	288,414	0	288,414
COCO	Mask decode failure	288,414	0	288,414
COCO	Foreground pixels < 64	288,414	15,437	272,977
COCO	Invalid deepest-interior anchor / $d \leq 0$	272,977	0	**272,977**
VOC2012	Raw universe: `val.txt` × unique `SegmentationObject` values	6,319	–	6,319
VOC2012	Non-instance values (background/void)	6,319	2,892	3,427
VOC2012	Foreground pixels < 256	3,427	230	3,197
VOC2012	Majority class outside the 20 VOC classes	3,197	0	3,197
VOC2012	Invalid anchor / depth < 1.5 px	3,197	5	**3,192**

The primary analysis uses the official native-multimask path and evaluates $m \in \{1, 5\}$ on the complete valid populations. A complementary validation analysis uses a deterministic hash sample of 2,000 COCO targets / 1,916 images and all 3,192 VOC2012 targets / 1,442 images. It evaluates $m \in \{1, 2, 5\}$ at the deepest-interior anchor A_0 and a predefined alternative anchor A_1. The intersections valid for A_1 contain 1,961 COCO and 3,187 VOC2012 targets. The complete-population sensitivity analysis under the controlled single-mask setting evaluates $m \in \{1, 2, 3, 5\}$. Targets are equally weighted and pairing within each target is preserved. The simulated interaction analysis is limited to VOC2012 and follows Appendix C.

Canonical image-cluster bootstrap. For the complete exact-repetition, validation, paired-contract, and fixed-S_1 comparisons, let d_i be the paired target-level difference, $g(i)$ its image, n_g the number of eligible targets in image g, and $S_g = \sum_{i:g(i)=g} d_i$. Each replicate samples the L unique images with replacement and computes the target-weighted ratio

$$\widehat{\Delta}^{(b)} = \frac{\sum_{g \in \mathcal{G}^{(b)}} S_g}{\max\left(\sum_{g \in \mathcal{G}^{(b)}} n_g, 1\right)}. \tag{8}$$

Drawing an image retains all of its eligible targets. The replicate estimand remains the equally target-weighted mean, not an image-equal mean. Reported percentile 95% intervals use 1,000 draws from one `default_rng(20260908)` stream consumed in the row order of the corresponding result table.

A.3 STATE HANDLING AND SIMULATED INTERACTION

Fixed-state SimpleClick comparisons hold the interaction state constant while varying duplicate evidence; sequential comparisons advance the interaction and carry state. The state-conditioned

SAM comparison in Appendix F freezes the same carried state across independent second-stage forwards.

The simulated-interaction analysis uses 300 VOC2012 targets per model under the tested GT-oracle clicker. Starting from an empty prediction, each round places a click in the largest current error region; false-negative regions yield positive clicks and false-positive regions negative clicks. The budget is 20 rounds, with early stopping only when prediction and ground truth are exactly equal. For positive events with prior positive evidence, define $\nu_{\text{pred}} = d_{\min}/s_{\text{pred}}$, where $d_{\min}$ is the minimum Euclidean distance to a prior positive click and s_{pred} is the pre-click predicted-mask bounding-box diagonal, falling back to the image diagonal when the predicted scale is absent or below one pixel. The binary threshold $\nu_{\text{pred}} < 0.08$ was selected on a separate 32-target DEV set by balanced accuracy and frozen before the sequential analysis reported here. Pre90/Pre95 count events up to and including the first post-click attainment of IoU 0.90/0.95. Bootstrap seed is 20260909. Replay, hash-alignment, postprocessing, and implementation checks are summarized in Appendix H; no human-frequency inference is made from these trajectories.

A.4 Reproducibility and reporting conventions

Mask round-trip, metric reconstruction, runner consistency, equivalence with the official path, eligibility replay, and SimpleClick replay were verified before the reported analyses. Appendix H gives the corresponding implementation details.

The analyses of exact repetition, all-distinct prompts, state-conditioned repetition, simulated interaction, and human events use distinct target universes and are reported separately. Fixed-state results are not pooled with results from sequential interaction. Unless stated otherwise, 95% intervals are descriptive intervals for each comparison; no family-wise multiplicity correction is applied. In appendix tables, "null" means that the reported 95% interval includes zero, not that equivalence has been established.

B Additional Controlled Results

B.1 All-distinct construction and C1–C6 values

Construction and cohort. The Controlled Low-Novelty all-distinct study reuses the 195-target D0–D6 cohort and its deepest-interior ground-truth anchor A. Let s_{gt} be the ground-truth bounding-box diagonal and $\rho(P) = \|A - P\|_2/s_{\text{gt}}$. C1–C6 use the predefined bands $(0, 0.01]$, $(0.01, 0.02]$, $(0.02, 0.04]$, $(0.04, 0.08]$, $(0.08, 0.16]$, and $(0.16, 0.32]$. Within each band, we select four ground-truth foreground pixels that are distinct from A and one another. Each lies at least $\max(2\text{ px}, 0.01 s_{\text{gt}})$ from the ground-truth boundary. The first added point is closest to the band centre in ρ. Subsequent points greedily maximize their minimum Euclidean distance to the selected set, with deterministic tie breaking. Coordinates are frozen before model inference and shared across SAM1, SAM2.1, and SAM3. Requiring all six configurations to be feasible retains 194 targets and favors larger objects. Every condition is evaluated independently with fresh state and `multimask_output=False`; SAM2.1 uses `apply_postprocessing=False`. The observation unit is the target, and 95% intervals use target bootstrap with seed 20260908.

Table 3: Geometry of the all-distinct constructed configurations. The bands use $\rho = \|A - P\|_2/s_{\text{gt}}$; mean pairwise distance is the recorded pixel-space average on the common 194-target cohort.

Condition	ρ band	Mean pairwise distance (px)
C1	$(0, 0.01]$	3.58
C2	$(0.01, 0.02]$	8.09
C3	$(0.02, 0.04]$	16.51
C4	$(0.04, 0.08]$	32.93
C5	$(0.08, 0.16]$	63.17
C6	$(0.16, 0.32]$	106.04

Table 4: All-distinct five-point results on the common cohort of 194 targets that satisfy the geometry constraints. ΔIoU is relative to the single-point $[A]$ baseline in IoU points; brackets are 95% CIs. Harm is $P(\Delta < 0)$.

Model	Metric	C1	C2	C3	C4	C5	C6
SAM1	ΔIoU [95% CI]	-7.72 $[-9.99, -5.60]$	-6.02 $[-8.24, -4.22]$	-1.49 $[-3.79, +0.89]$	$+6.52$ $[+3.42, +9.57]$	$+18.55$ $[+14.40, +22.63]$	$+25.94$ $[+21.49, +30.46]$
	Harm (%)	68.04	61.34	54.12	44.85	30.41	13.92
SAM2.1	ΔIoU [95% CI]	-27.58 $[-31.52, -23.72]$	-21.86 $[-25.47, -18.12]$	-11.30 $[-14.89, -7.74]$	-0.44 $[-3.88, +3.31]$	$+9.68$ $[+5.72, +13.46]$	$+13.86$ $[+10.23, +17.83]$
	Harm (%)	88.14	84.54	73.71	63.92	42.27	20.10
SAM3	ΔIoU [95% CI]	-18.06 $[-22.87, -13.68]$	-12.14 $[-16.03, -8.41]$	-6.27 $[-10.22, -2.26]$	$+1.72$ $[-2.32, +6.16]$	$+8.67$ $[+5.06, +12.28]$	$+11.60$ $[+8.19, +15.09]$
	Harm (%)	67.53	67.01	66.49	52.06	36.60	22.68

B.2 Two-point boundary case

The two-point boundary test compares $[A]$ with $[A, A]$ on the predefined 300-target COCO cohort under the controlled single-mask setting. The observation unit is the target, and the reported 95% intervals use target bootstrap with seed 20260908. Table 5 shows that output changes are nearly ubiquitous, while the average utility loss is resolved only for SAM2.1. This experiment uses a different cohort and inference contract from the 2,000-target official-default A_0 validation in Table 9; the two analyses are complementary, not identical replications.

Table 5: Two-point exact-duplicate boundary case on the 300-target COCO single-mask discovery cohort. ΔIoU is $[A, A] - [A]$ in IoU points. Mask change is the fraction whose binary output changes pixelwise.

Model	Mean ΔIoU [95% CI]	$P(\Delta < 0)$	Mask change
SAM1	-0.18 $[-1.07, +0.57]$	41.7%	97.3%
SAM2.1	-1.85 $[-2.90, -0.90]$	60.3%	100.0%
SAM3	-1.29 $[-2.73, +0.08]$	50.0%	99.7%

B.3 Gradual relaxation from exact repetition: D0–D6

Protocol. The D0–D6 study reuses the frozen 300-target COCO cohort and its deepest-interior ground-truth anchor A. Let s_{gt} be the ground-truth bounding-box diagonal and $\rho(B) = \|A - B\|_2 / s_{\mathrm{gt}}$. D1–D6 use the bands $(0, 0.01]$, $(0.01, 0.02]$, $(0.02, 0.04]$, $(0.04, 0.08]$, $(0.08, 0.16]$, and $(0.16, 0.32]$. In each band, B is a ground-truth foreground pixel distinct from A and at least $\max(2\text{ px}, 0.01 s_{\mathrm{gt}})$ from the ground-truth boundary. Among eligible pixels, we select the one whose ρ is closest to the band centre, with deterministic tie breaking. Requiring all six bands to be feasible retains 195 targets. D0 is $[A, A, A, A, A]$, and D1–D6 are $[A, A, A, A, B]$. All conditions use the same frozen coordinates across models and the same controlled single-mask contract with fresh state as the C1–C6 study. Every condition is compared with the single-point $[A]$ baseline, and 95% intervals use target bootstrap with seed 20260908. Four exact copies of A remain throughout, so the experiment moves one prompt gradually rather than creating an all-distinct configuration.

Table 6: Gradual relaxation from fivefold exact repetition on the common 195-target cohort. Values are mean ΔIoU relative to the single-point $[A]$ baseline in IoU points with 95% CIs. D0 is the exact-repetition endpoint; D1–D6 progressively displace the fifth correct point while retaining four copies of A.

Model	D0	D1	D2	D3	D4	D5	D6
SAM1	-8.34 $[-10.83, -5.99]$	-8.32 $[-10.78, -6.28]$	-7.78 $[-10.15, -5.77]$	-7.08 $[-9.39, -5.02]$	-5.71 $[-8.36, -3.35]$	$+0.80$ $[-1.98, +3.63]$	$+11.81$ $[+8.32, +15.24]$
SAM2.1	-29.86 $[-34.02, -25.54]$	-29.50 $[-33.41, -25.49]$	-28.43 $[-32.46, -24.32]$	-25.29 $[-29.05, -21.56]$	-20.37 $[-23.96, -17.21]$	-11.12 $[-14.35, -7.60]$	-0.63 $[-3.30, +2.08]$
SAM3	-19.94 $[-24.44, -15.68]$	-19.36 $[-24.11, -14.61]$	-17.78 $[-22.23, -13.42]$	-15.31 $[-19.60, -11.44]$	-11.63 $[-15.90, -7.52]$	-4.06 $[-8.30, +0.14]$	$+2.92$ $[-0.28, +6.26]$

The paired D6-minus-D0 recovery is $+20.15$ $[+16.59, +23.75]$, $+29.23$ $[+25.42, +33.13]$, and $+22.87$ $[+18.24, +27.17]$ IoU points for SAM1, SAM2.1, and SAM3, respectively. The mono-

tonic mean recovery across D0–D6 motivates the stronger all-distinct test above, where duplicate coordinates are removed entirely.

B.4 Complete-population exact repetition

Table 7 reports the full-population endpoint on the official native-multimask path with predicted-quality selection. Values are IoU points; intervals use the image-cluster bootstrap in Appendix A.

Table 7: Official-path exact repetition over the complete exact-valid populations. COCO has 272,977 exact-valid targets / 40,120 images; VOC2012 has 3,192 / 1,442. Oracle-best reports the highest ground-truth IoU among returned candidates; Official reports the model-selected output. Deltas are computed from unrounded values.

Model	Dataset	selected $m1 \to m5$	selected Δ_5 [95% CI]	oracle-best Δ_5 [95% CI]
SAM1	COCO	$60.49 \to 60.60$	$+0.10$ $[+0.03, +0.17]$	-1.30 $[-1.33, -1.27]$
SAM1	VOC2012	$62.49 \to 61.76$	-0.72 $[-1.48, +0.00]$	-2.95 $[-3.34, -2.57]$
SAM2.1	COCO	$67.31 \to 63.75$	-3.55 $[-3.65, -3.47]$	-0.64 $[-0.66, -0.61]$
SAM2.1	VOC2012	$71.79 \to 64.52$	-7.26 $[-8.09, -6.46]$	-0.97 $[-1.17, -0.77]$
SAM3	COCO	$68.26 \to 59.05$	-9.21 $[-9.34, -9.09]$	-2.97 $[-3.01, -2.93]$
SAM3	VOC2012	$73.85 \to 51.71$	-22.13 $[-23.44, -20.73]$	-4.61 $[-5.05, -4.17]$

The oracle-best candidate change is negative in all six model–dataset settings. The target-paired selection-regret accounting used in Section 4.2 is reported below.

Table 8: Target-paired comparisons on the existing full-population raw outputs. The single-minus-official contrast is $\Delta_{\text{single}} - \Delta_{\text{official}}$; negative values mean the forced single-mask path is more harmful. With O_m denoting oracle-best IoU and V_m official-selected IoU, $\Delta R = (O_5 - O_1) - (V_5 - V_1)$ is the selection-regret change; its per-target accounting identity has zero residual in all six cells. Values are IoU points with image-cluster 95% CIs.

Model	Dataset	$\Delta_{\text{single}} - \Delta_{\text{official}}$ [95% CI]	ΔR [95% CI]
SAM1	COCO	-2.28 $[-2.36, -2.20]$	-1.40 $[-1.48, -1.34]$
SAM1	VOC2012	-7.36 $[-8.44, -6.37]$	-2.22 $[-3.10, -1.45]$
SAM2.1	COCO	-12.03 $[-12.14, -11.92]$	$+2.92$ $[+2.83, +3.00]$
SAM2.1	VOC2012	-20.45 $[-21.76, -19.19]$	$+6.30$ $[+5.51, +7.20]$
SAM3	COCO	-0.51 $[-0.56, -0.46]$	$+6.24$ $[+6.12, +6.36]$
SAM3	VOC2012	-0.69 $[-1.03, -0.36]$	$+17.52$ $[+16.16, +18.89]$

B.5 Anchor and low-multiplicity validation

The complementary validation grid tests whether the endpoint replicates at a second pre-defined anchor and whether a single added duplicate already changes utility. Table 9 reports the paired mean changes; Table 10 separates pixelwise output change from utility change on the same grid.

B.6 Alternative inference contracts

The forced single-mask path uses the same targets and point prompts but returns one mask without candidate ranking. Table 11 reports the complete-population contrasts. For SAM2.1, Table 12 restores the released postprocessing setting while retaining forced single-mask output.

Table 9: Additional anchor and multiplicity validation under official default inference. A_0 rows use the deterministic 2,000-target COCO / 3,192-target VOC pool; A_1 rows use the pre-defined A_1-valid intersection (1,961 / 3,187 targets). Values are mean IoU-point changes with image-cluster 95% CIs.

Model	Dataset	Anchor	Contrast	Mean Δ [95% CI]
SAM1	COCO	A_0	$m2 - m1$	$+0.32\,[-0.14, +0.81]$
SAM1	COCO	A_0	$m5 - m1$	$+0.44\,[-0.27, +1.20]$
SAM1	COCO	A_1	$m2 - m1$	$+0.39\,[-0.08, +0.88]$
SAM1	COCO	A_1	$m5 - m1$	$+0.31\,[-0.43, +1.11]$
SAM1	VOC2012	A_0	$m2 - m1$	$-0.11\,[-0.52, +0.25]$
SAM1	VOC2012	A_0	$m5 - m1$	$-0.72\,[-1.48, +0.00]$
SAM1	VOC2012	A_1	$m2 - m1$	$+0.02\,[-0.45, +0.48]$
SAM1	VOC2012	A_1	$m5 - m1$	$-0.73\,[-1.53, +0.09]$
SAM2.1	COCO	A_0	$m2 - m1$	$+0.04\,[-0.41, +0.49]$
SAM2.1	COCO	A_0	$m5 - m1$	$-3.23\,[-3.98, -2.42]$
SAM2.1	COCO	A_1	$m2 - m1$	$-0.26\,[-0.74, +0.22]$
SAM2.1	COCO	A_1	$m5 - m1$	$-6.96\,[-7.94, -6.02]$
SAM2.1	VOC2012	A_0	$m2 - m1$	$-0.61\,[-0.97, -0.25]$
SAM2.1	VOC2012	A_0	$m5 - m1$	$-7.26\,[-8.09, -6.46]$
SAM2.1	VOC2012	A_1	$m2 - m1$	$-0.59\,[-1.00, -0.18]$
SAM2.1	VOC2012	A_1	$m5 - m1$	$-9.69\,[-10.57, -8.73]$
SAM3	COCO	A_0	$m2 - m1$	$-1.56\,[-2.24, -0.94]$
SAM3	COCO	A_0	$m5 - m1$	$-8.22\,[-9.35, -7.14]$
SAM3	COCO	A_1	$m2 - m1$	$-2.48\,[-3.19, -1.84]$
SAM3	COCO	A_1	$m5 - m1$	$-12.16\,[-13.25, -10.97]$
SAM3	VOC2012	A_0	$m2 - m1$	$-3.86\,[-4.53, -3.26]$
SAM3	VOC2012	A_0	$m5 - m1$	$-22.13\,[-23.44, -20.73]$
SAM3	VOC2012	A_1	$m2 - m1$	$-4.48\,[-5.19, -3.81]$
SAM3	VOC2012	A_1	$m5 - m1$	$-22.37\,[-23.83, -21.10]$

Table 10: Same-run selected-output equality and selection results for the additional validation grid. Mask equality is direct pixelwise `np.array_equal` against $m = 1$; hashes are persistence checks only. Changed∧small denotes changed masks with $|\Delta\text{IoU}| \leq 0.01$. Candidate-index switch is reported separately because index identity is not treated as mask-semantic identity.

Model	Dataset	Anchor	Contrast	Mask equal	Changed∧harmful	Changed∧small	Index switch
SAM1	COCO	A_0	$m2 - m1$	3.0%	42.0%	77.3%	12.3%
SAM1	COCO	A_0	$m5 - m1$	0.2%	49.6%	56.6%	36.5%
SAM1	COCO	A_1	$m2 - m1$	3.5%	42.4%	77.7%	12.3%
SAM1	COCO	A_1	$m5 - m1$	0.3%	48.1%	58.0%	36.3%
SAM1	VOC2012	A_0	$m2 - m1$	1.6%	40.6%	87.5%	12.4%
SAM1	VOC2012	A_0	$m5 - m1$	0.3%	47.1%	64.1%	43.5%
SAM1	VOC2012	A_1	$m2 - m1$	1.6%	40.5%	87.0%	11.6%
SAM1	VOC2012	A_1	$m5 - m1$	0.3%	45.7%	64.2%	43.5%
SAM2.1	COCO	A_0	$m2 - m1$	1.5%	50.8%	73.4%	11.3%
SAM2.1	COCO	A_0	$m5 - m1$	0.1%	56.6%	44.5%	33.2%
SAM2.1	COCO	A_1	$m2 - m1$	1.2%	52.9%	74.5%	11.0%
SAM2.1	COCO	A_1	$m5 - m1$	0.1%	60.7%	41.8%	38.1%
SAM2.1	VOC2012	A_0	$m2 - m1$	0.4%	43.2%	84.2%	9.9%
SAM2.1	VOC2012	A_0	$m5 - m1$	0.0%	52.4%	57.7%	30.0%
SAM2.1	VOC2012	A_1	$m2 - m1$	0.4%	44.4%	83.7%	10.7%
SAM2.1	VOC2012	A_1	$m5 - m1$	0.0%	55.1%	53.5%	35.7%
SAM3	COCO	A_0	$m2 - m1$	0.2%	56.1%	60.0%	22.1%
SAM3	COCO	A_0	$m5 - m1$	0.0%	64.3%	29.8%	54.2%
SAM3	COCO	A_1	$m2 - m1$	0.3%	55.5%	58.7%	23.9%
SAM3	COCO	A_1	$m5 - m1$	0.0%	70.6%	25.0%	56.6%
SAM3	VOC2012	A_0	$m2 - m1$	0.1%	46.2%	69.6%	23.7%
SAM3	VOC2012	A_0	$m5 - m1$	0.0%	67.1%	31.2%	57.0%
SAM3	VOC2012	A_1	$m2 - m1$	0.2%	48.3%	69.6%	23.2%
SAM3	VOC2012	A_1	$m5 - m1$	0.0%	69.9%	31.3%	55.8%

Table 11: Complete-population contrasts under the controlled single-mask setting. These values show sensitivity to the inference path and are not selected-output results from the official native-multimask path.

Model	Dataset	$m2 - m1$ [95% CI]	$m5 - m2$ [95% CI]	$m5 - m1$ [95% CI]
SAM1	COCO	$+0.59\,[+0.56, +0.61]$	$-2.76\,[-2.81, -2.71]$	$-2.17\,[-2.23, -2.12]$
SAM1	VOC2012	$+0.04\,[-0.17, +0.23]$	$-8.12\,[-8.81, -7.47]$	$-8.09\,[-8.80, -7.36]$
SAM2.1	COCO	$-2.61\,[-2.67, -2.56]$	$-12.97\,[-13.08, -12.85]$	$-15.58\,[-15.72, -15.44]$
SAM2.1	VOC2012	$-4.06\,[-4.46, -3.65]$	$-23.66\,[-24.76, -22.56]$	$-27.71\,[-28.81, -26.53]$
SAM3	COCO	$-1.82\,[-1.88, -1.76]$	$-7.90\,[-8.01, -7.80]$	$-9.72\,[-9.85, -9.60]$
SAM3	VOC2012	$-4.08\,[-4.81, -3.52]$	$-18.74\,[-20.05, -17.49]$	$-22.82\,[-24.24, -21.51]$

Table 12: SAM2.1 forced-single-mask comparison with released postprocessing restored. The single-mask path uses `multimask_output=False` and `apply_postprocessing=True`; the official path uses native multimask prediction, predicted-quality selection, and the same released postprocessing. Values are complete-population $m5-m1$ mean IoU-point changes with 1,000-draw image-cluster bootstrap 95% CIs (seed 20260908). The final column is the per-target paired contrast, not a subtraction of independent intervals; negative values mean the single-mask+postprocessing duplicate effect is more harmful than the official path.

Dataset	Targets	Official Δ_5 [95% CI]	Single+postproc Δ_5 [95% CI]	Paired contrast [95% CI]
COCO	272,977	-3.553 $[-3.648, -3.467]$	-9.115 $[-9.231, -9.008]$	-5.562 $[-5.637, -5.492]$
VOC2012	3,192	-7.264 $[-8.095, -6.464]$	-13.544 $[-14.562, -12.559]$	-6.280 $[-6.959, -5.607]$

C Nearby Prompts in Simulated Interaction

This appendix supports the stateful bridge in Section 5.1. All results use VOC2012 and the simulated GT-oracle interaction protocol, so the reported rates apply to this controlled policy rather than human clicking.

C.1 Strict first-hit Pre90/Pre95 occurrence

Events are nested within model–target sessions. The analysis bootstraps complete target sessions, retaining all events within a sampled session; image-cluster sensitivity is reported separately. Each model uses 300 targets from 300 images, giving 900 sessions overall. Events are not treated as independent bootstrap samples. Across 18,000 recorded rounds, the exact same-coordinate, same-label duplicate count is zero because the clicker excludes already-clicked pixels; this protocol therefore evaluates nearby additions without exact repeated coordinates.

The strict definition uses the first click index whose post-click prediction reaches the threshold; eligible events include clicks up to and including that first hit. This differs from merely selecting events whose pre-click IoU is below threshold because trajectories can be non-monotonic. Table 13 reports the first-hit occurrence summary.

Table 13: Strict first-hit nearby-event occurrence on VOC2012 under the simulated GT-oracle protocol. All-target prevalence uses the same target pool.

Model	Pre90 eligible	Pre90 nearby	Pre90 rate	Pre90 prevalence	Pre95 rate	Pre95 prevalence
SAM1	1163	369	31.7%	32.0%	36.5%	54.7%
SAM2.1	1358	412	30.3%	34.0%	32.3%	63.3%
SAM3	941	335	35.6%	30.3%	39.0%	57.3%

The Pre95 eligible/nearby counts are 1877/686, 2548/822, and 1794/700 for SAM1/SAM2.1/SAM3. In rounds 2–5 alone, nearby-event rates are 13.6%, 9.2%, and 10.1%, confirming that the pattern is not confined to the final interaction rounds.

Overall occurrence and late-stage amplification. Across all eligible positive events, nearby-event rates are 43.5%, 33.1%, and 42.4%, with all-target prevalence 85.6%, 98.0%, and 90.0%. These supplemental overall rates differ from rates measured before a quality threshold is first reached. Of all nearby events, approximately 72.3%, 69.4%, and 73.3% occur after first-hit IoU90 (approximately 69–73% across models). Late-stage refinement therefore raises the overall event prevalence. The main text reports the stricter first-hit Pre90/Pre95 summaries.

C.2 Marginal utility of nearby prompts

Binary controlled comparison. Across all eligible positive events, the binary comparison uses target fixed effects and predefined controls for click index, pre-click IoU, pre-click FN ratio, pre-click predicted-mask area ratio, and prior negative-click count. Figure 6 summarizes the contrast between $\nu_{\text{pred}} < 0.08$ and $\nu_{\text{pred}} \geq 0.08$, and Table 14 reports the numerical results. The threshold comes from the separate DEV-set selection described in Appendix A; it was not selected from these

sequential outcomes. A complementary continuous analysis tests the same normalized distance without thresholding.

Table 14: VOC2012-only binary near-versus-far normalized-distance association under the simulated GT-oracle protocol. These are within-protocol adjusted associations.

Model	adjusted ΔIoU near−far [95% CI]	harmful-probability near−far [95% CI]
SAM1	-0.007828 [-0.014332,-0.001838]	+0.156962 [+0.110627,+0.202866]
SAM2.1	-0.014379 [-0.018676,-0.010176]	+0.268815 [+0.233151,+0.304475]
SAM3	-0.014342 [-0.026313,-0.003131]	+0.145808 [+0.106082,+0.185917]

Continuous regressions and sensitivity. In the complementary continuous analysis, the $\nu_{\mathrm{pred},t} \rightarrow \Delta$IoU coefficient has the same positive direction in all three models but 95% interval support in only SAM2.1 and SAM3; SAM1 is null. For harmful probability, the continuous slope is negative for SAM1 and SAM2.1 and approximately zero for SAM3, with interval support only for SAM1. An image-diagonal normalization sensitivity analysis does not reverse the main directional association.

Interaction burden. The burden analysis uses early nearby-event exposure from rounds 2–5 and measures only later clicks-to-threshold/failure outcomes. NoC@90 directions are mixed across models and no model has a confidence interval excluding zero; NoC@95 point directions are positive in all three, but intervals again include zero. No stable global interaction-burden effect was established.

Relation to human data. Available human-click resources record different forms of interaction. RClicks (Antonov et al., 2024) contains real first-round clicks and later corrections collected against model-generated error masks. It does not contain live trajectories from the same user under the tested SAM1/SAM2.1/SAM3 pipelines. PointPrompt (Quesada et al., 2024a) records live prompt progression and corresponding SAM masks under a different annotation interface and dataset collection. Neither resource provides a directly comparable estimate of human-session frequency for our VOC2012 stateful contract. These simulated occurrence rates are specific to the tested protocol.

First-hit definition. Overall event frequencies mix early correction with refinement after a target has crossed a common quality threshold. The strict first-hit construction asks whether nearby additions appear before the trajectory first reaches that threshold. The late-stage analysis shows why overall percentages would overstate prevalence before the threshold.

D Human Interaction Analysis

This appendix provides the event construction and statistical details for the Naturalistic Low-Novelty analysis in Section 5.2. The analysis is observational and uses the released PointPrompt scores; it does not validate all three SAM generations.

D.1 Data source and event construction

The audit uses the PointPrompt data available through the `gOLIVES/SAM_PointPrompt_Dataset` mirror. The audited mirror contains 17 subsets / 6,400 rows, whereas the accompanying official repository lists 16 datasets / 6,000 rows. The additional 400 mirror rows are all from the `Pizza` subset and contain no `st*_points` or `st*_scores` fields, so they contribute zero events to the analysis. We therefore do not claim bitwise identity between the mirror and the original archive.

Events are constructed within recorded interaction rounds and never across round boundaries. We retain pairs of adjacent timesteps only when exactly one positive point is added, all previous positive points are retained, the negative set is unchanged, the added point lies inside the target, and both timesteps contain valid scores. Target membership is decoded from the official foreground label values before this eligibility test. The corrected filter produces 89,924 valid events from 16,397 human trajectories, with one positive point added in each event. The released score is treated as

the mask IoU at each timestep. The available metadata do not fully verify the exact model and configuration that generated every score, so this analysis provides observational evidence rather than validation across model generations.

For each added positive point, let

$$\nu_{\text{target}} = \frac{d_{\min}}{s_{\text{target}}},$$

where $d_{\min}$ is its Euclidean distance to the nearest previous positive point and s_{target} is the target bounding-box diagonal. This normalization is distinct from ν_{pred} in the simulated sequential analysis, so their numerical thresholds are not directly comparable. Exact repeats have $\nu_{\text{target}} = 0$.

D.2 Distance distribution and immediate utility

Exact repeats account for 3.06% of the corrected valid events. Table 15 reports the seven nonzero-distance bins. Including exact repeats, the corrected cumulative prevalences are 5.11% for $\nu_{\text{target}} \leq 0.01$, 9.38% for $\nu_{\text{target}} \leq 0.02$, 21.14% for $\nu_{\text{target}} \leq 0.04$, 45.23% for $\nu_{\text{target}} \leq 0.08$, and 74.88% for $\nu_{\text{target}} \leq 0.16$. These estimates describe the audited PointPrompt corpus under the stated event filter; they are not population estimates of human clicking behavior.

Table 15: PointPrompt events with one added positive click, grouped by normalized distance to the nearest prior positive. Shares are computed over all 89,924 corrected valid events; exact repeats ($\nu_{\text{target}} = 0$) account for 3.06% and are omitted from the rows below. Mean ΔIoU is the immediate score change after the added click, in raw IoU units.

Distance bin	Share (%)	Mean ΔIoU
$0 < \nu_{\text{target}} \leq 0.01$	2.06	-0.002750
$0.01 < \nu_{\text{target}} \leq 0.02$	4.27	-0.003910
$0.02 < \nu_{\text{target}} \leq 0.04$	11.76	-0.002670
$0.04 < \nu_{\text{target}} \leq 0.08$	24.09	$+0.001640$
$0.08 < \nu_{\text{target}} \leq 0.16$	29.65	$+0.018370$
$0.16 < \nu_{\text{target}} \leq 0.32$	19.17	$+0.056050$
$\nu_{\text{target}} > 0.32$	5.95	$+0.128770$

The three nearest nonzero-distance bins have small negative mean changes; the mean becomes positive for $\nu_{\text{target}} > 0.04$ and increases across the remaining bins.

D.3 Adjusted Association

The primary adjusted comparison defines Near as $0 < \nu \leq 0.02$ and Reference as $0.04 < \nu \leq 0.16$. Events are stratified by dataset, annotator, interaction round, previous-positive count, and a predefined current-IoU bin; only strata containing both groups are retained. The corrected adjusted sample contains 4,540 Near and 20,973 Reference events, yielding 25,513 events across 2,183 usable strata from 6,204 trajectories.

The event-weighted adjusted Near–Reference gap is -0.00978 raw IoU. A trajectory-cluster bootstrap gives a mean gap of -0.00967 with 95% CI $[-0.01248, -0.00699]$. A clustered regression controlling for the same measured factors gives a Near coefficient of -0.010997 with cluster-robust SE 0.001294, $t = -8.50$, and 95% CI $[-0.01353, -0.00846]$. The raw means are -0.00433 for Near and $+0.00643$ for Reference, a raw gap of -0.01076. These results support an observational association between smaller normalized distance and smaller immediate gain; they do not establish causality.

E Formal Properties of Duplicate Consistency

E.1 Definitions

For fixed image I and state S, let $E(\mathcal{P})$ be the model-facing prompt representation, $F(I, \mathcal{P}; S)$ the prediction, and $U(F, Y)$ a deterministic task utility for fixed ground truth Y; we suppress θ for

readability. Exact-duplicate consistency can be evaluated separately at representation, output, and utility layers. Closely spaced distinct prompts are not exact duplicates and therefore fall outside the exact consistency condition.

Sufficient implication. Assume $F(I, \mathcal{P}; S) = G(I, E(\mathcal{P}); S)$ and deterministic G for fixed I, S. If $E(\mathcal{P}^{(m)}) = E(\mathcal{P}^{(1)})$, both evaluations pass the same I, S, and evidence representation to the same deterministic map G. Their outputs are therefore identical. Applying the same deterministic utility $U(\cdot, Y)$ to identical outputs also gives identical utility. Under these conditions, Representation Consistency $\Rightarrow$ Output Consistency $\Rightarrow$ Utility Consistency. The condition is sufficient, not necessary.

Counterexamples to reverse implications. Output Consistency does not imply Representation Consistency: two distinct evidence representations may be absorbed or ignored by downstream computation and map to the same output. Utility Consistency does not imply Output Consistency: two distinct masks may achieve the same scalar IoU or other deterministic task utility. These counterexamples clarify the logical relationship among the three consistency levels.

E.2 Consequences and examples

Interface cases. We separate three cases: (i) representations that collapse duplicates; (ii) representations that preserve multiplicity; and (iii) stateful iterative contracts. A representation that preserves multiplicity exposes duplicate count to later computation and lacks the sufficient guarantee above. Later computation may still ignore the count, cancel its effect, improve the output, or make it worse. Keeping duplicates separate does not by itself predict the direction of the effect.

Consequences. The structural predictions are:

1. **Duplicate collapse.** If the effective representation is unchanged by an exact duplicate and downstream computation is deterministic under fixed state, fixed-state output should be exactly invariant.
2. **Multiplicity preservation.** If duplicates remain distinct at the representation layer, exact output consistency is not guaranteed; output may change, but neither harm direction nor harm rate is predicted.
3. **Stateful interaction.** A sequential repeated click that changes state/step is not a pure multiplicity intervention; the valid duplicate-consistency counterfactual must hold state fixed.
4. **Canonicalization.** Canonicalizing exact duplicates before multiplicity-dependent computation should restore exact fixed-state output consistency in a deterministic system.

The fixed-state and sequential SimpleClick comparison supports the predictions for duplicate collapse and stateful interaction. The SAM source supports multiplicity preservation, and exact canonicalization supports the canonicalization prediction.

E.3 Scope of the sufficient condition

The sufficient condition guarantees equality only when exact duplicates collapse to the same model-facing evidence under fixed state. When duplicates remain distinct, downstream computation may ignore their multiplicity, change the prediction without changing utility, improve utility, or reduce it. Sequential interactions that change state and closely spaced distinct prompts therefore fall outside this condition. Empirically, the direction varies with checkpoint and inference procedure: the effect on the official native-multimask path is negative for SAM2.1 and SAM3, while SAM1 shows a small positive COCO change and a VOC2012 interval that includes zero. The controlled single-mask analysis is negative in all six settings.

F Interface and State Controls

F.1 SimpleClick fixed-state equality

The tested COCO setting uses `PlainVitModel`, ViT-B at 448×448, and the SBD-trained checkpoint `sbd_vit_base.pth` (SHA256 `ebc4b7631b01853d70dffd04528a4d8c93b40bdbe29f79d6ca0c222a528ef824`) on the predefined 300-target/300-image COCO manifest. Preprocessing uses `ZoomIn(target=(448,448), skip_clicks=-1)`, horizontal-flip augmentation, sigmoid conversion, positional-embedding interpolation to 448, and binary threshold 0.49. The previous sigmoid prediction enters as a fourth input channel.

Duplicates collapse in the minimum-distance click-map encoding before network input. With identical initial state, equality is checked at zero numerical tolerance for the effective evidence map, carried state, network input, logits, probabilities, and binary masks. Across 300 targets and $m = 2, 3, 4, 5$, all 1,200 comparisons with the one-click reference are exactly equal at every checked layer.

F.2 Sequential repetition and SAM prompt semantics

Under sequential SimpleClick, 299/300 masks change after the repeated visible coordinate; the detailed 20-target trace records state changes in 20/20 cases and ZoomIn-state changes in 14/20. Because interaction step and state change together, this is a different intervention from fixed-state repetition.

For SAM1/SAM2.1/SAM3, repeated identical positive points remain as separate sparse prompt tokens in the tested fresh-state path. Exact input canonicalization maps $[A{\times}m]$ to $[A]$ before the prompt encoder and restores the $m = 1$ behavior on all 300 tested targets for each checkpoint. These checks establish multiplicity preservation for SAM and duplicate collapse for fixed-state SimpleClick without assigning a universal harm direction to either representation.

F.3 State-conditioned fixed-S_1 comparison

We do not use a mask-only reference with no new evidence as the state-conditioned counterfactual because it drops the previous sparse-point evidence. Instead, a normal first step with $[A]$ produces the carried state S_1. We freeze this same S_1 across independent second-stage forwards for $[A] + S_1$, $[A, A] + S_1$, $[A, A, A] + S_1$, and $[A \times 5] + S_1$. The primary contrast is $\Delta_{\text{state},2} = \text{IoU}([A, A] + S_1) - \text{IoU}([A] + S_1)$.

The protocol covers 2,000 COCO and 3,192 VOC2012 targets per model. Table 16 reports the primary two-copy contrast; four of six intervals lie below zero and two include zero.

Table 16: State-conditioned fixed-S_1 duplicate counterfactual under the forced single-mask contract. Values are $\Delta_{\text{state},2}$ in IoU points ($100\times$ raw IoU difference) with canonical target-weighted image-cluster bootstrap 95% CIs (Equation 8). Four of six intervals lie below zero, two include zero, and none lies entirely above zero.

Model	Dataset	$\Delta_{\text{state},2}$ [95% CI]
SAM1	COCO	$+0.017$ $[-0.302, +0.331]$
SAM1	VOC2012	-0.751 $[-1.047, -0.463]$
SAM2.1	COCO	-3.893 $[-4.396, -3.419]$
SAM2.1	VOC2012	-6.648 $[-7.187, -6.160]$
SAM3	COCO	-0.351 $[-0.801, +0.085]$
SAM3	VOC2012	-0.971 $[-1.411, -0.538]$

G Comparison with Prior Work for the Zero-Novelty Test

Prior work has studied prompt count, over-prompting, placement, composition, redundancy, selection, and refinement. Table 17 compares the studies most directly related to our zero-novelty

exact-repetition test. Existing conditions with more or optimized prompts generally introduce new point locations, change the selected prompt set, or advance interaction state. Our exact-repetition setting holds the coordinate, label, step, and state fixed while varying only multiplicity.

Table 17: Comparison with closest prior work. "New spatial evidence" asks whether the higher-count or optimized condition changes/introduces point locations. "Exact duplicate" requires repetition of the same coordinate and label as the manipulated variable, not merely nearby or redundant prompts.

Work	Main prompt manipulation	New spatial evidence	Fixed step/state	Exact coordinate+label duplicate	Relation to exact repetition
PointPrompt (Quesada et al., 2024a)	Human sequential point-prompting trajectories	Yes	No	Not reported	An evolving interaction trajectory; exact fixed-state multiplicity is not isolated.
Benchmarking Human/Automated Prompting (Quesada et al., 2024b)	Six sampling strategies and searched inclusion/exclusion counts	Yes	No isolated multiplicity counterfactual	Not reported	Valid policy/count benchmark; count is coupled to sampled coordinates.
Stable-SAM (Fan et al., 2025)	1/3/5/10 positive points randomly selected from GT mask	Yes	No isolated multiplicity counterfactual	Not reported	Studies count robustness with newly sampled points.
Mastering SAM Prompts (Price et al., 2025)	Broad point/box/mask prompt configurations	Yes / composition changes	No isolated multiplicity counterfactual	Not reported	Large-scale prompt-configuration study, not exact repetition with zero new spatial support.
PPO (Liu et al., 2025)	Optimizes point-prompt distribution in feature/physical spaces	Yes / moved or selected points	No	Not reported	Prompt optimization changes the selected or moved evidence.
Attack for Defense (Liu et al., 2026)	Selects/suppresses disruptive points from coarse prompt sets	Yes / subset changes	No	Not reported	Prompt optimization/defense changes the selected evidence.
AoP-SAM (Chen et al., 2025b)	Generates/filters prompt candidates to reduce redundant refinements	Yes / candidate selection	No	Not reported	Redundancy filtering/efficiency, not exact duplicate consistency.
SAMRefiner (Lin et al., 2025)	Multi-prompt excavation from a coarse mask (points, elastic boxes, Gaussian-style masks)	Yes / modality and location change	No isolated multiplicity counterfactual	Not reported	Multi-prompt refinement, not fixed-evidence repetition.
Ours	Repeat identical positive point m times	**No**	**Yes**	**Yes**	Isolates multiplicity with zero new spatial support at fixed step and state.

H Additional Examples and Implementation Details

H.1 Qualitative interface examples

Figure A1 complements the SAM example in Figure 1 with three SimpleClick targets selected in fixed dataset order, independent of visual outcome. The top block holds the image, evidence, interaction step, and state fixed. The $[A]$ and $[A \times 5]$ conditions produce identical masks, and the black panels are exact zero-difference maps. The bottom block shows official sequential interaction with repeated A: the interaction step and carried state advance, and the output changes. The examples show the distinction between fixed-state and sequential interfaces; all quantitative claims come from the experiments above.

H.2 Implementation and reproducibility details

The recorded execution environment provides four NVIDIA A30 GPUs (24 GB each; driver 550.163.01), Python 3.12.14, PyTorch 2.6.0+cu124, CUDA 12.4, and cuDNN 90100. SAM1 inference uses fp32 without autocast; SAM2.1 and SAM3 use CUDA bfloat16 autocast for image setup and prediction, while all three stored checkpoints use fp32 weights. For the 2,000-target COCO scope pool, eligible target UIDs were sorted by `SHA256('R2_PHASE0_20260911' | uid)` and the first 2,000 were selected. An eligibility replay on CPU reused the original manifest builder, constants, and filter order and exactly reconstructed both final UID sets; anchor coordinates and

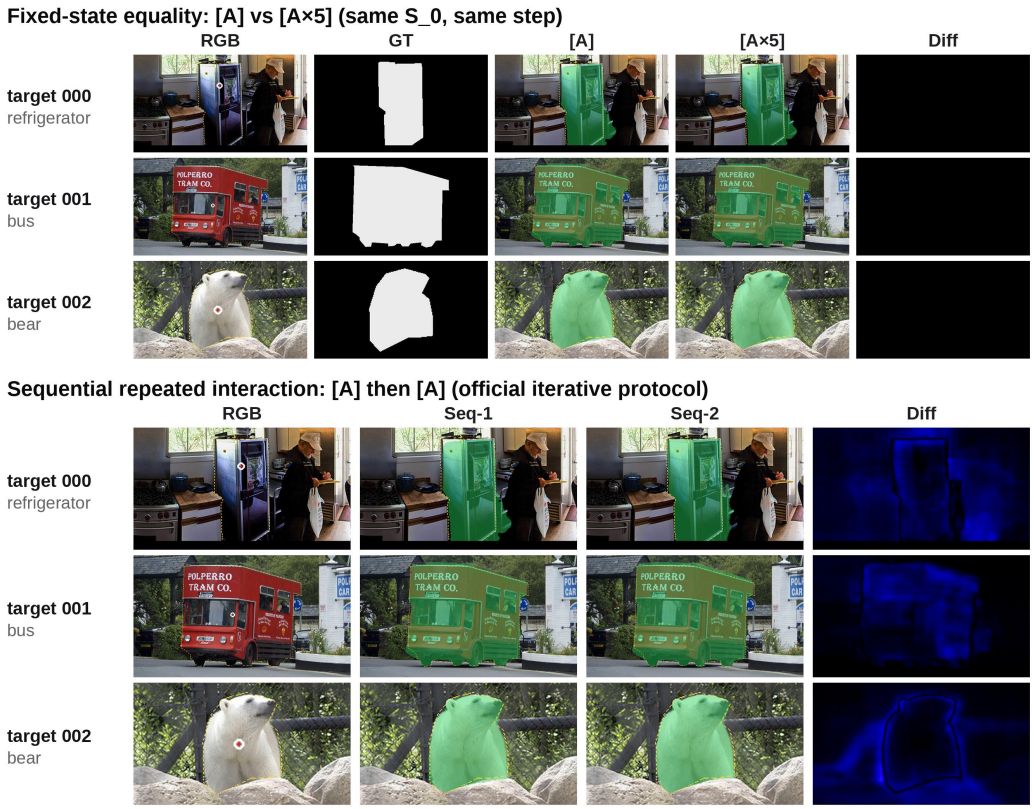


Figure A1: Qualitative SimpleClick fixed-state and sequential controls on the first three targets of the predefined 20-target pool. Top: with the same image, step, and state, $[A]$ and $[A \times 5]$ yield identical outputs; black difference panels denote exactly zero pixelwise difference. Bottom: under the official sequential interaction, repeating the visible point A advances the interaction step and changes carried state, and the output evolves. The comparison separates fixed-state repeated evidence from a sequential interaction in which step and state evolve together.

foreground-pixel counts also matched with zero discrepancies. The implementation checks include consistency under exact repetition, reconstruction from official logits to masks, deterministic raw reforwarding, inspection of the SAM and SimpleClick prompt interfaces, SimpleClick replay, clicker snapshots, target/ground-truth hash alignment, and exact fixed-state traces. Table 18 summarizes the checks most directly tied to the reported outputs.

Validation detected an early grouped-runner issue that mixed outputs across models. All affected intermediate outputs were excluded, and every analysis reported in this paper uses outputs from the corrected pipeline.

H.3 Postprocessing Details

The main SAM multiplicity results use the official native-multimask path with predicted-quality selection, as specified in Appendix A. The dynamic-multimask stability fallback is a single-mask decoder branch and is not reached on this path. For SAM2.1, the controlled single-mask procedure sets `apply_postprocessing=False`. The additional single-mask comparison restores the released setting and reduces the loss without removing it (Table 12). In the independent logit check, thresholding captured logits at zero reproduces the corresponding binary masks. SimpleClick's official iterative evaluation includes its released postprocessing and refinement stack; the fixed-state comparison holds state fixed. Cross-system comparisons are limited to computational variables that are aligned.

Table 18: Reproducibility checks. The official default equivalence row documents the cohort construction, equality criteria, and mismatch rates for the API default path used in the main text.

Check	Cohort / procedure	Equality criterion	Result
Official-default equivalence	SHA256-ordered fixed cohort: 500 COCO + 500 VOC targets per model; three models; $m \in \{1, 5\}$ (6,000 condition comparisons)	Candidate scores: `np.allclose` with `atol=0, rtol=0`; selected index exact; selected mask `np.array_equal`	Scores $6000/6000$ equal; selected-index mismatches 0; masks $6000/6000$ bitwise equal. Logits/probabilities not recorded
Eligibility replay	Original manifest builder, original COCO/VOC inputs, and original filter order; pipeline attrition accounting	Final target UID sets exact; replayed `a_x/a_y/gt_mask_px` exact on the intersection	COCO $272{,}977/272{,}977$ and VOC $3{,}192/3{,}192$ UID intersection; missing 0, extra 0, field mismatches 0
Mask round-trip	Pack and unpack stored masks	Bitwise mask equality	$1200/1200$ identical, 0 mismatches
Metric reconstruction	Recompute metrics from unpacked masks	Exact stored metric reconstruction	$1200/1200$ reproduced, 0 mismatches
Runner consistency	Grouped runner vs. per-target runner	Mask/metric exact; max absolute logit difference	1820 comparisons/model, 0 mask mismatch, 0 metric mismatch, max logit difference 0.0
SimpleClick replay	Re-run fixed-state/sequential outputs	Stored summary / trace replay	Stored summary and trace outputs reproduced