PROMPT VARIATION ALIGNMENT FOR ROBUST SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Promptable segmentation models such as the Segment Anything Model (SAM) family provide convenient interfaces for interactive mask generation through clicks, bounding boxes, text prompts, or their combinations. However, SAM-based segmentation models are often sensitive to the user prompt: clicks placed on different parts of an object, or different text prompts for the same object, can substantially degrade the predicted mask — making the model unreliable in practice, especially in critical applications such as medical imaging. To address this, we propose Best-Teaches-Rest (BTR), a post-training objective that enforces the alignment of predictions across prompts that refer to the same target object. BTR groups prompts that refer to the same object, identifies the highest-IoU prediction as the teacher, and trains the remaining predictions to match it — with only the teacher receiving ground-truth supervision. Unlike standard supervised fine-tuning where prompt invariance can only emerge implicitly across independent prompt–mask pairs, BTR enforces it directly through the loss. BTR loss can be applied uniformly to click, box, text, and joint-modality prompt variation. We post-train SAM 3 with BTR on EntitySeg and observe consistent gains in both mask quality and prompt consistency over pretrained SAM 3. These gains transfer zero-shot to six object-centric instance segmentation datasets and to the out-of-distribution RAOS benchmark (Luo et al., 2024) for medical image segmentation.

1 INTRODUCTION

Promptable segmentation models in the Segment Anything family (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) return a mask for the target indicated by a user-supplied click, bounding box, text prompt, or by any combination of the three. The implicit assumption is that prompts referring to the same target under a fixed segmentation criteria (e.g. instance vs. semantic) yield the same mask – a property that, when violated, directly undermines the model's reliability in deployment.

In practice, however, SAM-based models are fragile to prompt variation even when the target and the segmentation criteria remain fixed: slight perturbations in click locations, bounding box sizes, or noun phrases can substantially change the predicted mask – with IoU against the ground-truth mask shifting by tens of points. Standard evaluations of segmentation models often report IoU under canonical prompts such as region-centered clicks, obscuring how performance varies across prompt variations. The IoU a downstream user experiences is closer to the mean across the user prompts, and this unreliability is particularly consequential in safety-critical applications such as medical imaging.

One canonical approach is to fine-tune on more diverse prompts: for each target object, sample several click locations or noun phrases and add the resulting prompt–mask pairs to the training set. However, this augmented supervised fine-tuning does not impose prompt invariance as a constraint on the model: each prompt–mask pair contributes an independent gradient toward the ground-truth mask, and the loss can decrease without strong per-prompt agreement of the masks. The constraint that the model's prediction does not depend on the supplied valid prompt is a structural property of the model, not of the data distribution.

To address this, we propose Best-Teaches-Rest (BTR), a post-training objective that enforces alignment of predictions across prompts that refer to the same target. BTR groups such prompts and uses the highest-IoU prediction in each group to supervise the remaining predictions. The prompt variations are constructed to mirror realistic interaction patterns: clicks sampled from across the

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.

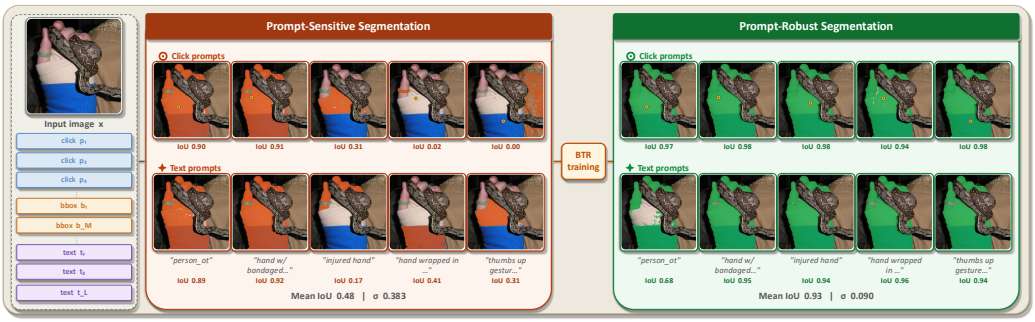


Figure 1: **Prompt variation alignment.** Multiple clicks or text prompts can refer to the same object, but promptable segmentation models may return different masks for each prompt. Best-Teaches-Rest post-trains the model so predictions from valid prompt variations agree on the target.

object's extent rather than only its center, and alternative noun phrases a user might plausibly use to refer to the same object. The same loss applies seamlessly to click, box, and text variation, as well as their joint click+text and click+box+text variation; only the composition of the group changes.

A naive single-stage application of BTR can align predictions toward an unreliable teacher early in training, producing a wrong-and-consistent failure mode. Therefore, we follow a training recipe with two components: (i) a one-epoch warmup stage in which only the ground-truth anchor is active, so the consistency and reference-KL terms turn on only after the best mask in each group is reliable enough to teach the rest; and (ii) selective adaptation, freezing the vision, language, and geometry prompt encoders and updating the transformer encoder-decoder, segmentation head, and scoring head. We ablate these components in Section 5.3.

We post-train SAM 3 with BTR on EntitySeg (Qi et al., 2023) — a dataset of high-quality dense mask annotations spanning diverse image domains — and evaluate against the frozen pretrained SAM 3. BTR substantially improves mean, worst-prompt, and stability of IoU across click prompts, and yields modest gains across noun-phrase prompts. These click-variation gains transfer zero-shot to LVIS (Gupta et al., 2019) for object-centric instance segmentation and to the out-of-distribution medical-imaging benchmark RAOS (Luo et al., 2024).

Our contributions are threefold:

1. We propose Best-Teaches-Rest (BTR), a post-training objective that enforces prompt invariance as a structural property of the model by coupling predictions across prompts through the loss, and applies unchanged to click, box, text, and joint prompt variation (Section 4).
2. We identify a wrong-and-consistent failure mode of naive BTR and introduce a two-component recipe – warmup and selective adaptation – that prevents it (Section 4.4).
3. We post-train SAM 3 with BTR on EntitySeg and show large gains over the frozen pretrained segmentation models in mean, maximum, and minimum IoU and in IoU stability across click prompts with zero-shot transfer to LVIS and the out-of-distribution medical benchmark RAOS (Section 5).

2 Related Work

SAM variants and promptable segmentation. The Segment Anything Model (SAM) family (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026) introduced general-purpose promptable segmentation models; subsequent work adapts SAM via varied backbone architectures (HQ-SAM (Ke et al., 2023)), parameter-efficient adapters (Chen et al., 2023; Wu et al., 2025; Chen et al., 2025), and medical-imaging specialization (Ma et al., 2024; Cheng et al., 2023; Wang et al., 2024; Li et al., 2025). In our work, we keep the architecture fixed and propose a post-training objective; the same loss can be applied to any of these regimes.

For text-conditioned segmentation (Hu et al., 2016; Mao et al., 2016; Liu et al., 2023a; Lai et al., 2024), robustness to paraphrased variants has emerged as a challenge: R^2-Bench (Li et al., 2024b) and Ref-L4 (Chen et al., 2024) provide paraphrase-augmented benchmarks while SPARTA (Zinkovich

et al., 2026) documents adversarial-paraphrase vulnerability and SelfEQ (He et al., 2024) enforces paraphrase consistency on grounding attention. Our BTR objective applies to text robustness as well as clicks, which we demonstrate with SAM 3.

Click robustness and disambiguation in interactive segmentation. Iterative click-based segmentation (Maninis et al., 2018; Jang & Kim, 2019; Sofiiuk et al., 2022; Chen et al., 2022; Liu et al., 2023b) reports IoU as a function of click count, treating it as a budget. A separate line of work studies click *position*: TETRIS (Moskalenko et al., 2024) formalizes click-position robustness via best- vs. worst-case click trajectories, and RClicks (Antonov et al., 2024) shows interactive segmentation models are not robust under realistic click distributions. Recent SAM-side methods address this with a deformable-attention plugin (Fan et al., 2025), a variance-insensitive refinement module (Fang et al., 2024), perturbed-prompt fine-tuning (Rahman et al., 2024), a noise-tolerant click formulation (Zhang et al., 2025), and a probabilistic prompt encoder (Wang et al., 2025). In contrast, BTR is a post-training objective that improves robustness across different prompt types and generalizes zero-shot to unseen benchmarks.

A complementary thread treats one-click-many-masks as a feature rather than a robustness liability: SAM's three-mask head ranks alternative interpretations of an ambiguous click by predicted IoU, while PiClick (Yan et al., 2024), Semantic-SAM (Li et al., 2024a), GraCo (Zhao et al., 2024), and CPOT (Liu et al., 2024) similarly rank or condition on candidate masks; our objective is the converse – all valid prompts for a given object should yield consistent masks.

Auxiliary regularizers: cross-view consistency and KL-anchored fine-tuning. BTR is structurally a consistency objective, but differs from cross-view consistency in semi-supervised segmentation (Chen et al., 2021; Ouali et al., 2020; Liu et al., 2022; Wang et al., 2022; Yang et al., 2023; French et al., 2020; Yang et al., 2022; Zhao et al., 2023) in two respects: (i) those methods enforce symmetric agreement across views, whereas BTR ranks predictions by quality and supervises asymmetrically; (ii) their views are stochastic perturbations during training, whereas BTR's prompt groups reflect genuine variation in user-supplied inputs.

The forward-KL term serves as a standard anti-forgetting anchor (Li & Hoiem, 2016), motivated by evidence that fine-tuning can degrade out-of-distribution performance (Kumar et al., 2022; Wortsman et al., 2022), and is used in BTR to limit deviation from the pretrained reference.

3 Problem Formulation

3.1 Promptable Segmentation

A promptable segmentation model with parameters θ takes an image $x \in \mathbb{R}^{3\times H\times W}$ and a prompt q that refers to a target object, and predicts a binary mask $\hat{m} = f_\theta(x, q) \in [0,1]^{H\times W}$. Prompts can be visual (e.g. clicks or bounding boxes), textual (e.g. nouns or referring expressions), or a combination.

We consider single-interaction promptable segmentation, where the user supplies one prompt per object – distinct from iterative interactive segmentation (Sofiiuk et al., 2022; Chen et al., 2022; Liu et al., 2023b), where a mask is refined through repeated input. SAM models are trained to support multiple valid segmentation criteria for the same prompt – e.g., part, object, or surrounding region – which is a useful capability but a confounding source of variation when measuring prompt invariance. We therefore fix the segmentation criterion to instance-level masks and consider any variation across valid prompts undesired. Under this scope, the choice of prompt is incidental to the user's goal: a click on the chest of a dog, a click on its leg, the phrase "the dog," and the phrase "the brown dog on the right" all refer to the same target. The output should depend on (x, m^*), where m^* denotes the ground-truth mask of the target, not on the specific element of the equivalence class of valid prompts.

3.2 Prompt Variation Alignment

We denote $\mathcal{G} = \{q_1, \ldots, q_{|\mathcal{G}|}\}$ as the *prompt group*: a set of prompts that all refer to the same target object. *Prompt variation alignment* is the property that the predictions $\{\hat{m}_k = f_\theta(x, q_k)\}_{k=1}^{|\mathcal{G}|}$ agree pixelwise. We consider three prompt modalities and their compositions. In *click variation*, the group consists of K clicks $\{p_1, \ldots, p_K\}$ sampled inside m^*. In *box variation*, the group

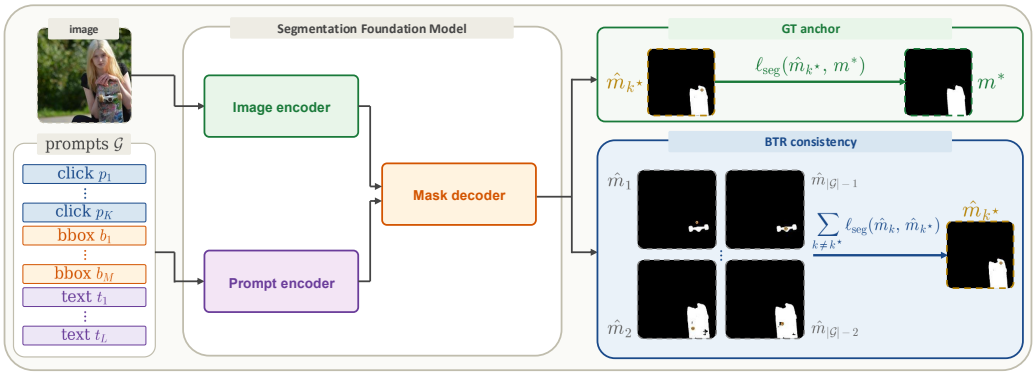


Figure 2: **Best-Teaches-Rest (BTR) overview.** For each target object, we form a prompt group $\mathcal{G}$ of clicks $\{p_1, \ldots, p_K\}$, bounding boxes $\{b_1, \ldots, b_M\}$, and noun phrases $\{t_1, \ldots, t_L\}$ that all refer to the same target. The segmentation foundation model produces a predicted mask $\hat{m}_k$ for each prompt. The *GT anchor* term supervises only the highest-IoU prediction $\hat{m}_{k^\star}$ against the ground-truth mask m^*, while the *BTR consistency* term aligns the remaining predictions to $\hat{m}_{k^\star}$ under stop-gradient. The model architecture is shown schematically; the specific modules frozen and updated for each model configuration are described in Section 4.4 and Appendix A.

consists of M bounding boxes $\{b_1, \ldots, b_M\}$ referring to the same target. In *text variation*, the group consists of L alternative noun phrases $\{t_1, \ldots, t_L\}$ for the same object. In *joint variation*, the group is the union of prompts from multiple modalities. For click+text variation, the group is $\{p_1, \ldots, p_K\} \cup \{t_1, \ldots, t_L\}$ with $|\mathcal{G}| = K + L$. For click+text+box variation, the group is $\{p_1, \ldots, p_K\} \cup \{b_1, \ldots, b_M\} \cup \{t_1, \ldots, t_L\}$ with $|\mathcal{G}| = K + M + L$. The procedures we use to construct these groups are described in Section 4.2.

3.3 Evaluation

Reporting IoU under a single canonical prompt — e.g., a region-centered click or a tight ground-truth box — could overstate deployment performance, since users may supply a range of prompts in deployment. We instead construct prompt group $\mathcal{G}$ with prompts that reflect realistic interaction patterns (Section 4.2), and report the *mean*, *maximum*, *minimum*, and *standard deviation* of $\text{IoU}(\hat{m}_k, m^*)$ across $k \in \mathcal{G}$. The mean reflects the IoU a user can expect on average; the minimum captures the worst observed behavior among the evaluated prompt variants; the maximum indicates whether the model can produce a correct mask for at least one prompt in the group; and the standard deviation measures consistency across the group.

4 Method

4.1 Setup

For each input image x and target object with ground-truth mask m^*, we construct a prompt group $\mathcal{G} = \{q_1, \ldots, q_{|\mathcal{G}|}\}$ in which every prompt refers to the same target. Let $\hat{m}_k = f_\theta(x, q_k)$ denote the model's predicted mask under prompt q_k, and $l_k^{(i,j)}$ the corresponding pre-sigmoid logit at pixel (i, j), so $\hat{m}_k^{(i,j)} = \sigma(l_k^{(i,j)})$. We write $\pi_{\theta,k}$ and $\pi_{\text{ref},k}$ for the per-pixel Bernoulli predictions of the current model and the frozen pretrained reference under prompt q_k.

4.2 Prompt Group Construction

We construct prompt variants from each ground-truth mask, choosing variants that mirror realistic interaction patterns. Click variants are sampled across the mask's interior, spanning the spectrum from the most-centered point to points near the boundary. For box variation, bounding-box variants are obtained by perturbing the tight ground-truth box through translations and changes in its extent and scale. For text variation, we use five noun phrases per object: the original category name and

four variants generated by Qwen3-VL-8B (Qwen Team, 2025). To ground the VLM on the target, all non-target pixels are replaced by a checkerboard pattern.

For single-modality training, we use $K = 10$ clicks, $M = 10$ boxes, or $L = 5$ text prompts per object. For joint training, click+text uses $K = 10$ clicks and $L = 5$ text prompts, while click+text+box uses five prompts from each modality, yielding $|\mathcal{G}| = 15$ in both joint settings.

We manually inspect 400 phrases and trigger a verification stage with Qwen3-VL-32B-AWQ to filter and regenerate if the error rate exceeds 3%. Detailed construction procedures for box and text variants are provided in Appendix F.

4.3 Best-Teaches-Rest Objective

Given a prompt group $\mathcal{G}$, we identify the highest-IoU prediction

$$k^\star = \arg \max_{k \in \{1,\dots,|\mathcal{G}|\}} \mathrm{IoU}(\hat{m}_k, m^*), \tag{1}$$

and define the Best-Teaches-Rest loss consisting of three terms:

$$\mathcal{L}_{\mathrm{BTR}} = \lambda_1 \underbrace{\ell_{\mathrm{seg}}(\hat{m}_{k^\star}, m^*)}_{\text{GT anchor}} + \lambda_2 \underbrace{\frac{1}{|\mathcal{G}|-1} \sum_{k \neq k^\star} \ell_{\mathrm{seg}}\big(\hat{m}_k, \mathrm{sg}(\hat{m}_{k^\star})\big)}_{\text{BTR consistency}} + \lambda_{D_{\mathrm{KL}}} \underbrace{\frac{1}{|\mathcal{G}|} \sum_{k=1}^{|\mathcal{G}|} D^{\mathrm{px}}_{D_{\mathrm{KL}}}\big(\pi_{\mathrm{ref},k} \| \pi_{\theta,k}\big)}_{\text{reference KL}}. \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient, $\ell_{\mathrm{seg}} = \ell_{\mathrm{focal}} + \ell_{\mathrm{Dice}}$ combines per-pixel sigmoid focal loss (Lin et al., 2017) with DICE loss (Milletari et al., 2016) to balance pixel-level supervision with region-level overlap under class imbalance, and $D^{\mathrm{px}}_{D_{\mathrm{KL}}}$ is the per-pixel Bernoulli KL between the reference and current model (full form in Appendix B). We analyze the gradient contributions of the three terms in Appendix C. We use $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_{D_{\mathrm{KL}}} = 0.2$ (Appendix A). For the $|\mathcal{G}| = 1$ ablation, we define the BTR consistency term as zero.

The three terms have distinct roles. The *GT anchor* provides the only direct ground-truth signal and applies to a single member of the group: at each training step, we compute $k^\star$ as the current highest-IoU prediction and supervise only that mask against m^*, keeping the learning signal focused on the most reliable teacher in the group. The *BTR consistency* term pulls the remaining predictions toward the current best mask, directly penalizing disagreement among prompt-conditioned predictions. The stop-gradient on the best mask blocks the direct consistency gradient through the teacher target branch; we analyze this gradient effect in Appendix C.2. The *reference KL* prevents the post-trained model from drifting too far from the reference, which is intended to help retain zero-shot transfer.

4.4 Training Recipe

A naive single-stage minimization of equation 2 can align predictions toward an unreliable teacher early in training, producing wrong-and-consistent failures. We address this with a two-component recipe.

1. **Warmup.** The first epoch is standard SFT (i.e. $\lambda_1 = 1$, $\lambda_2 = \lambda_{D_{\mathrm{KL}}} = 0$); BTR consistency and reference-KL activate from epoch two.
2. **BTR training.** All three loss terms are active. For the main SAM 3 experiments, we freeze the vision, language, and geometry prompt encoders and update only the transformer encoder–decoder, segmentation head, and scoring head. For SAM, we freeze the image encoder and update the prompt encoder and mask decoder. Architecture-specific settings for the SAM 3 interactive pathway used in the RClicks evaluation are provided in Appendix A.

5 Experiments

5.1 Setup

Baselines and variants. We apply BTR on top of two base models, SAM (Kirillov et al., 2023) and SAM 3 (Carion et al., 2026). On SAM, we post-train with click-variation prompt groups only. On

Table 1: Robustness evaluation under instance-level (click, box) and semantic (noun phrase) perturbations. For each instance, we evaluate T perturbations (e.g., $T = 15$ clicks), and compute per-instance statistics: the mean IoU over perturbations, the best (maximum) IoU, the worst (minimum) IoU, and the standard deviation across perturbations. We then report the dataset-level average of these per-instance metrics. Specifically, $\mathrm{IoU}_{\mathrm{avg}}$ denotes the mean of per-instance average IoUs, $\mathrm{IoU}_{\mathrm{max}}$ and $\mathrm{IoU}_{\mathrm{min}}$ denote the mean of per-instance best and worst IoUs, respectively, and σ_{IoU} denotes the mean of per-instance standard deviations. RAOS and LVIS are zero-shot evaluation sets.

Dataset	Method	15 Clicks				11 Boxes				5 Noun Phrases			
		$\mathrm{IoU}_{\mathrm{avg}}\uparrow$	$\mathrm{IoU}_{\mathrm{max}}\uparrow$	$\mathrm{IoU}_{\mathrm{min}}\uparrow$	$\sigma_{\mathrm{IoU}}\downarrow$	$\mathrm{IoU}_{\mathrm{avg}}\uparrow$	$\mathrm{IoU}_{\mathrm{max}}\uparrow$	$\mathrm{IoU}_{\mathrm{min}}\uparrow$	$\sigma_{\mathrm{IoU}}\downarrow$	$\mathrm{IoU}_{\mathrm{avg}}\uparrow$	$\mathrm{IoU}_{\mathrm{max}}\uparrow$	$\mathrm{IoU}_{\mathrm{min}}\uparrow$	$\sigma_{\mathrm{IoU}}\downarrow$
EntitySeg	SAM 3 (Frozen)	47.84	55.85	37.66	6.78	45.80	75.89	15.93	22.41	47.18	65.73	20.72	19.06
	SAM 3 + BTR (click)	81.97	84.93	76.04	2.79	77.32	**87.36**	51.22	11.85	50.99	69.02	24.16	19.05
	SAM 3 + BTR (text)	44.89	50.05	39.29	3.95	52.28	72.44	25.40	17.76	**52.60**	67.66	**30.54**	**15.80**
	SAM 3 + BTR (box)	74.63	83.21	50.49	11.20	**83.32**	87.16	**73.86**	**4.52**	45.60	65.67	19.30	19.60
	SAM 3 + BTR (click+text)	**82.25**	**85.06**	**77.08**	**2.50**	77.58	87.34	50.54	12.12	52.07	68.46	28.75	16.83
	SAM 3 + BTR (click+text+box)	81.13	83.94	75.23	2.88	82.86	87.22	71.98	5.23	51.01	**69.34**	25.49	18.55
RAOS	SAM 3 (Frozen)	29.19	37.47	20.90	5.44	52.70	70.10	32.27	12.44	-	-	-	-
	SAM 3 + BTR (click)	**52.38**	57.84	**43.87**	4.43	61.49	72.47	46.47	8.01	-	-	-	-
	SAM 3 + BTR (text)	38.52	47.95	27.05	7.36	57.40	70.60	39.64	9.79	-	-	-	-
	SAM 3 + BTR (box)	50.29	**61.77**	34.81	9.26	**64.57**	72.08	**53.77**	**5.67**	-	-	-	-
	SAM 3 + BTR (click+text)	51.83	57.16	43.40	**4.40**	60.42	72.08	43.43	8.89	-	-	-	-
	SAM 3 + BTR (click+text+box)	48.15	54.35	39.91	5.01	64.40	**72.87**	52.11	6.50	-	-	-	-
LVIS	SAM 3 (Frozen)	58.52	66.13	48.76	6.41	-	-	-	-	**57.98**	75.47	**26.31**	**20.58**
	SAM 3 + BTR (click+text)	**73.37**	**77.58**	**67.44**	**3.26**	-	-	-	-	55.83	**75.50**	23.96	21.55

SAM 3, we post-train under the five prompt-group compositions defined in Section 3.2: click, box, text, joint click+text and joint click+box+text. For each (base model, composition) pair we compare BTR against the frozen pretrained model evaluated under the same prompt protocol. Training follows the recipe of Section 4.4; more details such as hyperparameters and hardware requirements are in Appendix A.

Datasets. We post-train on EntitySeg (Qi et al., 2023) and evaluate on three datasets: the EntitySeg held-out split (in-distribution), and zero-shot transfer to RAOS (Luo et al., 2024) and LVIS (Gupta et al., 2019). RAOS comprises 413 CT volumes covering 17–19 abdominal organs per patient, with challenging cases such as partial or total organ excision and cross-modality scans, and tests transfer to a far-OOD medical domain unseen during post-training. LVIS is a large-vocabulary object-centric segmentation benchmark whose long-tail noun-phrase distribution tests text-prompt robustness in the natural-image regime. Across all datasets, we filter out instances whose mask area is either too small or big: small masks make click-variant generation unreliable, while very large masks dominate the image and degenerate the prompt-conditioning signal (see Appendix A.3).

Evaluation protocols. We evaluate SAM 3 under click, box and noun-phrase perturbations to measure robustness within each prompt modality as well as transfer across modalities. SAM uses click and bounding-box perturbations only, as it has no text encoder. Each SAM 3 instance receives $T_{\text{click}} = 15$ clicks and $T_{\text{text}} = 5$ noun-phrase paraphrases (Section 4.2), plus $T_{\text{box}} = 11$ bounding boxes obtained by shifting and scaling the ground-truth box (see Appendix F.4); SAM uses the same click and box protocols. For each (instance, modality, perturbation) we compute IoU against the ground-truth mask, aggregate to per-instance statistics (mean, best, worst, standard deviation), and report dataset-level averages. On RAOS we omit noun-phrase perturbations because SAM 3's text encoder has limited medical-terminology coverage; on LVIS, compute limits restrict evaluation to frozen SAM 3 and the BTR (click+text) variant under click and noun-phrase perturbations only.

We additionally evaluate on RClicks (Antonov et al., 2024), which measures click-position robustness across five interactive-segmentation benchmarks (GrabCut, Berkeley, DAVIS, COCO-MVal, TETRIS) using crowdsourced real-user clicks. We report User@1, the per-user mean and standard deviation of IoU, in the main comparison. For the RClicks comparison, we additionally apply BTR to SAM 3's native interactive pathway; its optimization configuration is described in Appendix A.1.

5.2 Results

On EntitySeg in-distribution (Table 1), BTR (click+text) lifts click $\mathrm{IoU}_{\mathrm{avg}}$ from 47.84 to 82.25, while BTR (box) raises box $\mathrm{IoU}_{\mathrm{avg}}$ from 45.80 to 83.32. The text-only variant performs best on noun-phrase prompts ($52.60/15.80$ for $\mathrm{IoU}_{\mathrm{avg}}/\sigma_{\mathrm{IoU}}$). Spatial robustness learned from click and box

Table 2: One-click User@1 IoU (mean±std) under the RClicks protocol. Top-3 baselines are reproduced from Antonov et al. (2024) for reference; all baselines are listed in Appendix F.6. SAM 3-interactive + BTR adds the grouped-prompt objective to the native SAM 3 interactive pathway, with no change to architecture or inference.

Method	GrabCut	Berkeley	DAVIS	COCO-MVal	TETRIS	Avg.
AdaptClickLin et al. (2024)	89.26±4.47	83.45±5.56	74.92±7.09	77.17±6.49	82.16±6.15	81.39±5.95
SimpleClickLiu et al. (2023b)	88.51±4.48	83.66±4.88	74.35±6.71	80.34±6.16	83.47±6.98	82.07±5.84
CFR-ICLSun et al. (2024)	88.95±4.72	85.05±5.88	75.48±8.23	**81.18±6.55**	85.92±7.02	83.32±6.48
SAM (ViT-H)	75.19±16.43	74.13±17.30	54.88±16.88	67.26±17.00	63.88±20.39	67.07±17.60
SAM (ViT-H) + BTR	83.43±15.63	80.19±16.48	69.37±15.87	71.40±18.13	78.12±20.14	76.50±17.25
SAM 3-interactive	74.94±16.26	80.45±12.26	70.37±12.98	72.87±15.26	47.89±19.17	69.30±15.19
SAM 3-interactive + BTR	**89.28±3.38**	**86.59±3.32**	**76.68±7.41**	79.42±6.72	**88.54±5.99**	**84.10±5.36**

Table 3: Robustness evaluation on SAM (ViT-H) under click and box perturbations on EntitySeg (in-distribution) and RAOS (zero-shot), following the same protocol as Table 1. We compare the frozen pretrained SAM against SAM post-trained with BTR on click-variation prompt groups.

Dataset	Method	15 Clicks				11 Boxes			
		$\mathrm{IoU}_{\mathrm{avg}}\uparrow$	$\mathrm{IoU}_{\mathrm{max}}\uparrow$	$\mathrm{IoU}_{\mathrm{min}}\uparrow$	$\sigma_{\mathrm{IoU}}\downarrow$	$\mathrm{IoU}_{\mathrm{avg}}\uparrow$	$\mathrm{IoU}_{\mathrm{max}}\uparrow$	$\mathrm{IoU}_{\mathrm{min}}\uparrow$	$\sigma_{\mathrm{IoU}}\downarrow$
EntitySeg	SAM (Frozen)	55.93	70.04	32.30	**12.94**	75.40	82.70	58.07	8.30
	SAM + BTR (click)	**68.12**	**82.74**	**35.28**	15.97	**78.04**	**84.09**	**63.46**	**6.74**
RAOS	SAM (Frozen)	**27.67**	32.10	**21.83**	**3.33**	**61.28**	**68.87**	**52.04**	**5.26**
	SAM + BTR (click)	15.87	**33.67**	4.50	11.08	56.46	65.07	46.23	5.70

variation transfers bidirectionally between the two prompt types, whereas transfer involving text is asymmetric.

For zero-shot transfer to RAOS, BTR (click) gains $+23.19$ on click and $+8.79$ on box $\mathrm{IoU}_{\mathrm{avg}}$. Box-only BTR similarly improves both box ($52.70 \rightarrow 64.57$) and click ($29.19 \rightarrow 50.29$) robustness, showing that the bidirectional spatial transfer also persists out of domain. On LVIS, click $\mathrm{IoU}_{\mathrm{avg}}$ rises $58.52 \rightarrow 73.37$ and σ_{IoU} more than halves ($6.41 \rightarrow 3.26$), whereas noun-phrase performance trails the frozen baseline by $\sim$2 points.

On RClicks (Table 2), applying BTR to SAM 3's native interactive pathway raises the five-dataset average User@1 IoU from 69.30 to 84.10, while reducing its standard deviation from 15.19 to 5.36. The resulting model is competitive with established interactive-segmentation methods across the five benchmarks, providing benchmark-level context for BTR under realistic user clicks.

To check that BTR is not specific to SAM 3, we apply BTR (click) on SAM (ViT-H) and evaluate on EntitySeg (Table 3) and on RClicks (Table 2). On EntitySeg, click and box $\mathrm{IoU}_{\mathrm{avg}}$ both improve ($55.93 \rightarrow 68.12$ and $75.40 \rightarrow 78.04$). On RClicks, SAM + BTR improves User@1 mean IoU on all five benchmarks, mirroring the SAM 3 pattern at lower absolute scores. On RAOS, however, SAM + BTR underperforms the frozen baseline (click $\mathrm{IoU}_{\mathrm{avg}}$ $27.67 \rightarrow 15.87$), suggesting that BTR is less reliable when the base model has limited coverage of the far-OOD target domain.

Qualitative results. Figures 3, 6, 7, and 8 show qualitative comparisons between the frozen SAM 3 baseline and SAM 3 + BTR under click, noun-phrase, and bounding-box perturbations, respectively. Frozen SAM 3 often produces masks that drift or fragment as the prompt varies — click position shifts the target boundary, paraphrased noun phrases yield semantically inconsistent regions, and box perturbations change the mask scale. SAM 3 + BTR produces substantially more stable predictions across all three modalities.

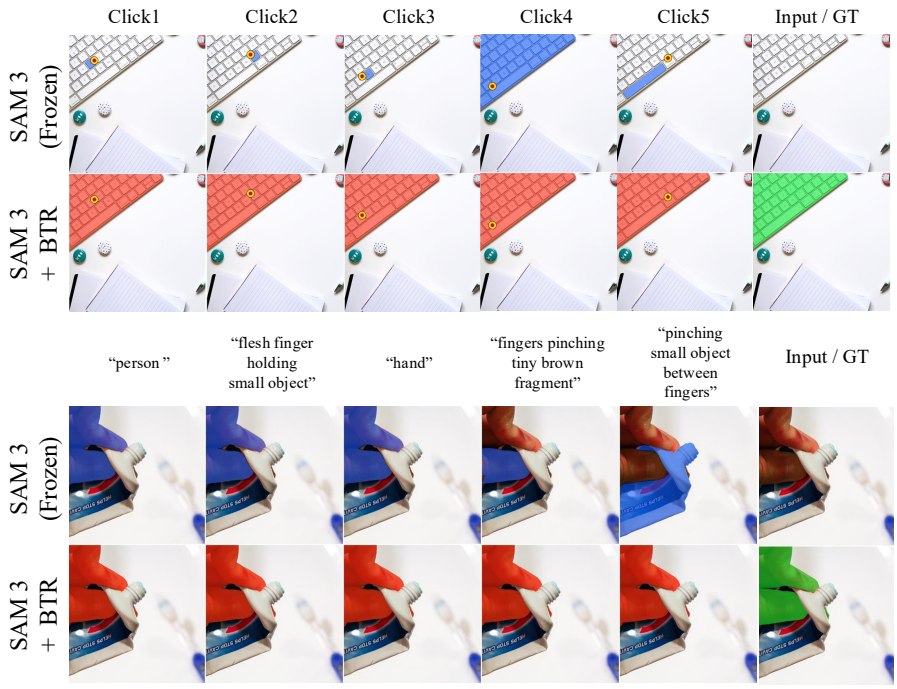


Figure 3: Qualitative comparison under subset of click and noun phrase variations. SAM 3 (frozen) exhibits inconsistent predictions across interactions, while SAM 3 + BTR produces stable and consistent masks. The green mask denotes the ground truth.

5.3 Ablations

Matched augmented-SFT. We compare BTR with augmented-SFT baselines matched in training data, prompt groups, optimization steps, model configuration, and prompt–mask exposure. BTR improves $\mathrm{IoU}_{\mathrm{avg}}$ by 1.86–2.03 points over the two matched click baselines while reducing prompt-dependent variation (Table 4(a)). For text variation, BTR similarly improves $\mathrm{IoU}_{\mathrm{avg}}$ from 50.09 to 52.60, raises the worst-prompt IoU from 26.13 to 30.54, and reduces σ_{IoU} from 17.76 to 15.80, while the best-prompt IoU remains comparable (68.03 vs. 67.66). The exact augmented-SFT objectives and additional matched click/text comparisons are provided in Appendix A.2.

Prompt-group size. Varying $K \in \{1, 3, 5, 10\}$ shows that most of the gain is obtained by $K = 3$, with only marginal improvements from larger groups (Table 4(b)).

Training Recipe. We ablate the warmup stage and trainable-module configuration on EntitySeg under click-prompt variation (Table 4(c)). Removing warmup reduces $\mathrm{IoU}_{\mathrm{avg}}$ from 81.97 to 81.32. Freezing only the backbone lowers it to 80.35, while additionally freezing the transformer encoder further reduces $\mathrm{IoU}_{\mathrm{avg}}$ to 79.10 and increases σ_{IoU} from 2.79 to 5.22. These results support both warmup and adaptation of the transformer encoder in the full BTR recipe.

6 Discussion and Limitations

Noun-phrase prompts are ambiguous at the instance level. In an image containing two monitors, the phrase "monitor" may validly refer to either instance or to their union. We address this in two ways: text-only post-training is evaluated at the semantic-mask level (which absorbs per-instance ambiguity), and the joint click+text variant uses the click signal to disambiguate the target. The LVIS noun-phrase regression – BTR (click+text) trailing the frozen baseline by ~ 2 points (Table 1) – may reflect specialization to the noun-phrase distribution of EntitySeg during post-training.

Cross-modality transfer is asymmetric. Click-variation post-training transfers to box prompts: BTR (click) raises EntitySeg box $\mathrm{IoU}_{\mathrm{avg}}$ from 45.80 to 77.32 even though boxes are not in the

Table 4: Ablations on EntitySeg. All results use the click-prompt evaluation protocol unless otherwise specified.

Ablation	Prompt	Method / Setting	$\mathrm{IoU}_{\mathrm{avg}} \uparrow$	$\mathrm{IoU}_{\mathrm{max}} \uparrow$	$\mathrm{IoU}_{\mathrm{min}} \uparrow$	$\sigma_{\mathrm{IoU}} \downarrow$
(a) Matched Aug-SFT	Click	Aug-SFT	79.94	83.83	72.56	3.68
	Click	Aug-SFT-Equal-λ	80.11	83.62	73.12	3.46
	Click	BTR ($K = 10$)	**81.97**	**84.93**	**76.04**	**2.79**
	Text	Aug-SFT	50.09	**68.03**	26.13	17.76
	Text	BTR	**52.60**	67.66	**30.54**	**15.80**
(b) Prompt-group size	Click	BTR ($K = 1$)	80.04	83.98	72.15	3.96
	Click	BTR ($K = 3$)	81.74	84.75	75.58	2.92
	Click	BTR ($K = 5$)	81.90	84.93	75.83	2.87
	Click	BTR ($K = 10$)	**81.97**	**84.93**	**76.04**	**2.79**
(c) Training recipe and modules	Click	Full recipe	**81.97**	**84.93**	**76.04**	**2.79**
	Click	w/o warmup	81.32	84.48	75.00	3.11
	Click	Freeze backbone only	80.35	84.06	73.25	3.56
	Click	Freeze backbone + transformer encoder	79.10	84.01	68.20	5.22

training prompt group (Table 1). Text-variation post-training, by contrast, lowers click accuracy ($47.84 \rightarrow 44.89$). Click and box prompts both condition on a localized image region, so invariance to one tends to imply invariance to the other. Text prompts condition on a category label whose pixel support is not spatially localized; forcing image-conditioned predictions to align with a noisier text-conditioned mask drags them away from the target. The joint click+text variant recovers click and box performance while improving text invariance.

Consistency does not guarantee correctness. BTR substantially reduces unstable and inaccurate predictions during training, but a non-negligible stable-but-wrong population remains. From the end of warmup to the final checkpoint, the fraction of stable-correct instances increases from 60.12% to 68.93%, while 14.36% of instances remain stable but wrong at the final checkpoint. This shows that improved cross-prompt consistency does not by itself guarantee segmentation correctness. A detailed state-transition analysis and representative severe failures are provided in Appendix G.1.

BTR's gains depend on base-model coverage of the target domain. On RAOS, SAM 3 + BTR raises click $\mathrm{IoU}_{\mathrm{avg}}$ from 29.19 to 52.38 (Table 1), but SAM + BTR drops from 27.67 to 15.87 (Table 3). The reference-KL anchor of Section 4.3 preserves whatever pretrained coverage exists; it does not create coverage from scratch. SAM 3 was pretrained on a broader and more diverse image distribution than SAM (ViT-H), which appears to retain enough structure for the medical-CT domain to benefit from BTR. BTR is therefore most reliable when the base model already has at least weak prior coverage of the deployment domain.

Scaling prompt-group size. Each training step requires $|\mathcal{G}|$ prompt-conditioned forward passes per instance; the image features are shared, so overhead scales linearly but moderately with prompt-group size. In our largest configurations, both joint settings use 15 prompts per instance.

Object categories. BTR yields gains on structurally challenging categories such as wires, ropes, cables, and branches. However, they fall below the overall average, suggesting that a degree of shape-dependent sensitivity remains. A per-category breakdown is provided in Appendix G.2.

7 Conclusion

We address *prompt variation alignment*: predictions for prompts that refer to the same target object should agree, regardless of the user prompt. We propose Best-Teaches-Rest, a post-training objective that imposes this constraint by letting the highest-IoU prediction in a prompt group supervise the rest. With a training recipe including a warmup stage, post-training SAM 3 with BTR on EntitySeg substantially improves spatial-prompt robustness across click and box variations, while also providing in-domain gains under noun-phrase variation. The spatial robustness gains transfer zero-shot to LVIS and the out-of-distribution RAOS benchmark, whereas cross-dataset transfer for text prompts remains more limited.

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 (NeurIPS)*, 2024. URL https://arxiv.org/abs/2410.11722.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, 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 (ICLR)*, 2026.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2406.16866.

Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. SAM fails to segment anything? – SAM-Adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2023. URL https://arxiv.org/abs/2304.09148.

Tianrun Chen, Runlong Cao, Xinda Yu, Lanyun Zhu, Chaotao Ding, Deyi Ji, Cheng Chen, Qi Zhu, Chunyan Xu, Papa Mao, and Ying Zang. SAM3-Adapter: Efficient adaptation of segment anything 3 for camouflage object segmentation, shadow detection, and medical image segmentation. *arXiv preprint arXiv:2511.19425*, 2025. URL https://arxiv.org/abs/2511.19425.

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.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. URL https://arxiv.org/abs/2106.01226.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. SAM-Med2D. *arXiv preprint arXiv:2308.16184*, 2023. URL https://arxiv.org/abs/2308.16184.

Qi Fan, Xin Tao, Lei Ke, Mingqiao Ye, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, Yu-Wing Tai, and Chi-Keung Tang. Stable segment anything model. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2311.15776.

Chaowei Fang, Ziyin Zhou, Junye Chen, Hanjing Su, Qingyao Wu, and Guanbin Li. Variance-insensitive and target-preserving mask refinement for interactive image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2312.14387.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *British Machine Vision Conference (BMVC)*, 2020. URL https://arxiv.org/abs/1906.01916.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5356–5364, 2019.

Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C. Berg, and Vicente Ordonez. Improved visual grounding through self-consistent explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2312.04554.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision (ECCV)*, volume 9905, pp. 108–124, 2016.

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 (NeurIPS)*, 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.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2202.10054.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Chengyin Li, Prashant Khanduri, Yao Qiang, Rafi Ibn Sultan, Indrin Chetty, and Dongxiao Zhu. AutoProSAM: Automated prompting SAM for 3D multi-organ segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. URL https://arxiv.org/abs/2308.14936.

Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-SAM: Segment and recognize anything at any granularity. In *European Conference on Computer Vision (ECCV)*, 2024a. URL https://arxiv.org/abs/2307.04767.

Xiang Li, Kai Qiu, Jinglu Wang, Xiaohao Xu, Rita Singh, Kashu Yamazaki, Hao Chen, Xiaonan Huang, and Bhiksha Raj. R^2-Bench: Benchmarking the robustness of referring perception models under perturbations. In *European Conference on Computer Vision (ECCV)*, 2024b. URL https://arxiv.org/abs/2403.04924.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision (ECCV)*, 2016. URL https://arxiv.org/abs/1606.09282.

Jiacheng Lin, Jiajun Chen, Kailun Yang, Alina Roitberg, Siyu Li, Zhiyong Li, and Shutao Li. Adaptiveclick: Click-aware transformer with adaptive focal loss for interactive image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5759–5773, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, 2023a.

Jie Liu, Haochen Wang, Wenzhe Yin, Jan-Jakob Sonke, and Efstratios Gavves. Click prompt learning with optimal transport for interactive segmentation. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/04816.pdf.

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, 2023b.

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2111.12903.

Xiangde Luo, Zihan Li, Shaoting Zhang, Wenjun Liao, and Guotai Wang. Rethinking abdominal organ segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15009, pp. 531–541, 2024.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. doi: 10.1038/s41467-024-44824-z. URL https://arxiv.org/abs/2304.12306.

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2016.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 416–423, 2001.

Kevin McGuinness and Noel E. O'Connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

F. Milletari, N. Navab, and S.A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in Hilbert space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Andrey Moskalenko, Vlad Shakhuro, Anna Vorontsova, Anton Konushin, Anton Antonov, Alexander Krapukhin, Denis Shepelev, and Konstantin Soshin. TETRIS: Towards exploring the robustness of interactive segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2402.06132.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://arxiv.org/abs/2003.09005.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4047–4056, 2023.

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Md Mostafijur Rahman, Mustafa Munir, Debesh Jha, Ulas Bagci, and Radu Marculescu. PP-SAM: Perturbed prompts for robust adaptation of segment anything model for polyp segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. URL https://arxiv.org/abs/2405.16740.

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 (ICLR)*, 2025.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut": Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH*, volume 23, pp. 309–314, 2004.

Konstantin Sofiiuk, Ilia A. Petrov, and Anton Konushin. f-BRS: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Defines the COCO-MVal evaluation subset.

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

Shoukun Sun, Min Xian, Fei Xu, Luca Capriotti, and Tiankai Yao. Cfr-icl: Cascade-forward refinement with iterative click loss for interactive image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5017–5024, 2024.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. In *European Conference on Computer Vision Workshops (BIC)*, 2024. URL https://arxiv.org/abs/2310.15161.

Xiaoqi Wang, Clint Sebastian, Wenbin He, and Liu Ren. ProSAM: Enhancing the robustness of SAM-based visual reference segmentation with probabilistic prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. URL https://arxiv.org/abs/2506.21835.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fan, Yuxi Wang, Chang Xu, Xinyi Le, and Wei Zhang. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2203.03884.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2109.01903.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical Image Analysis*, 2025. URL https://arxiv.org/abs/2304.12620.

Cilin Yan, Haochen Wang, Jie Liu, Xiaolong Jiang, Yao Hu, Xu Tang, Guoliang Kang, and Efstratios Gavves. PiClick: Picking the desired mask from multiple candidates in click-based interactive segmentation. *Neurocomputing*, 2024. URL https://arxiv.org/abs/2304.11609.

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://arxiv.org/abs/2106.05095.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2208.09910.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_NTClick_Achieving_Precise_Interactive_Segmentation_With_Noise-tolerant_Clicks_CVPR_2025_paper.html.

Yian Zhao, Kehan Li, Zesen Cheng, Pengchong Qiao, Xiawu Zheng, Rongrong Ji, Chang Liu, Li Yuan, and Jie Chen. GraCo: Granularity-controllable interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2405.00587. Highlight.

Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2211.11335.

Viktoriia Zinkovich, Anton Antonov, Andrei Spiridonov, Denis Shepelev, Andrey Moskalenko, and Daria Pugacheva. SPARTA: Evaluating reasoning segmentation robustness through black-box adversarial paraphrasing in text autoencoder latent space. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026. URL https://arxiv.org/abs/2510.24446.

A Experiment Settings

A.1 Training Configurations and Hyperparameters

SAM 3-BTR. All SAM 3-BTR variants are initialized from the same pretrained SAM 3 checkpoint and post-trained on the filtered-and-verified EntitySeg training split described below. We freeze the vision, language, and geometry encoders and update the remaining trainable modules, including the transformer encoder/decoder, segmentation head, and scoring head. Training uses image batch size 1, image resolution 1008, bfloat16 mixed precision, AdamW, weight decay 0.1, gradient clipping at norm 0.1, and an inverse-square-root learning-rate schedule with 20 warmup and cooldown steps. Within each image batch, the model processes all retained target instances in the image, and each instance is evaluated under the perturbations in its prompt group. The active trainable parameter groups use learning rate 8×10^{-5}.

We use loss weights 1.0, 0.5, and 0.2 for the GT-anchor, BTR consistency, and reference-KL terms, respectively. The consistency and reference-KL terms are enabled after the first epoch, so epoch 0 serves as the GT-anchor warmup. Each run is trained on a single NVIDIA A100 40GB or RTX Pro 6000 Blackwell 96GB GPU. For the largest prompt group considered, containing up to 15 prompts per target instance, training took 52 hours and used 46GB peak VRAM.

SAM 3-interactive-BTR. For the RClicks comparison, we apply BTR to SAM 3's native interactive pathway. We update the interactive prompt encoder and mask decoder while leaving the underlying architecture and inference procedure unchanged.

SAM-BTR Because SAM does not take text prompts, we apply SAM-BTR to SAM using click prompts only. We start from SAM ViT-H, freeze the image encoder, and train the prompt encoder and mask decoder with $K = 10$ random positive clicks per instance. The training recipe follows SAM 3-BTR: bfloat16 mixed precision, AdamW, learning rate 8×10^{-5}, weight decay 0.1, gradient

Table 5: Matched augmented-SFT comparisons on EntitySeg under click and text prompt variation. Aug-SFT uses the same prompt groups and training configuration as the corresponding BTR variant.

Prompt	Method	$\mathrm{IoU}_{\mathrm{avg}} \uparrow$	$\mathrm{IoU}_{\max} \uparrow$	$\mathrm{IoU}_{\min} \uparrow$	$\sigma_{\mathrm{IoU}} \downarrow$
Click	SAM 3 (Frozen)	47.84	55.85	37.66	6.78
	SAM 3 + Aug-SFT	79.94	83.83	72.56	3.68
	SAM 3 + BTR	**81.97**	**84.93**	**76.04**	**2.79**
Text	SAM 3 (Frozen)	47.18	65.73	20.72	19.06
	SAM 3 + Aug-SFT	50.09	**68.03**	26.13	17.76
	SAM 3 + BTR	**52.60**	67.66	**30.54**	**15.80**

clipping at norm 0.1, and an inverse-square-root schedule with 20 warmup and cooldown steps. The loss weights are again 1.0, 0.5, and 0.2 for the GT-anchor, BTR consistency, and reference-KL terms, with the latter two enabled after epoch 0. The SAM instantiation uses resolution 1024.

A.2 Matched Augmented-SFT Baselines

To isolate the gains of BTR from simply training with augmented prompts, we compare BTR with augmented-SFT baselines matched in training data, prompt groups, optimization steps, model configuration, and prompt–mask exposure.

Unlike BTR, augmented SFT directly supervises every prompt-conditioned prediction with the ground-truth mask. For a controlled comparison, we retain the same best-prompt selection $k^\star$ and loss weights as BTR, but replace the BTR teacher target for the remaining predictions with the ground truth:

$$\mathcal{L}_{\text{Aug-SFT}} = \lambda_1 \ell_{\text{seg}}\left(\hat{m}_{k^\star}, m^\star\right) + \lambda_2 \frac{1}{|\mathcal{G}| - 1} \sum_{k \neq k^\star} \ell_{\text{seg}}\left(\hat{m}_k, m^\star\right) + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}. \tag{3}$$

To remove the asymmetric weighting induced by $k^\star$, we additionally consider an equal-weight variant,

$$\mathcal{L}_{\text{Aug-SFT-Equal-}\lambda} = \lambda_1 \frac{1}{|\mathcal{G}|} \sum_{k \in \mathcal{G}} \ell_{\text{seg}}\left(\hat{m}_k, m^\star\right) + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}, \tag{4}$$

which assigns the same segmentation-loss weight to all predictions in the prompt group.

Table 5 reports matched comparisons for click and text prompt variation. Under matched prompt augmentation, BTR improves $\mathrm{IoU}_{\mathrm{avg}}$ over Aug-SFT by 2.03 points for clicks and 2.51 points for text, while also reducing prompt-dependent variation in both settings.

A.3 Datasets

We list more details about the datasets we use to evaluate the zero-shot transfer of BTR post-training.

EntitySeg filtering and text prompts. We construct the EntitySeg training and validation splits to contain instances for which both click perturbations and noun-phrase perturbations are meaningful. Starting from EntitySeg annotations with valid instance masks, we first retain masks whose normalized area lies in $[0.005, 0.8]$, removing very small objects that make click sampling unreliable and very large masks that are less informative for prompt robustness. For each remaining mask, we generate four noun-phrase variants with Qwen3-VL-8B using dual-image conditioning, and pair them with the original EntitySeg category name to form a five-prompt semantic perturbation set.

We then run the pretrained SAM 3 model separately with each of the five text prompts and compute the standard deviation of the resulting IoUs against the ground-truth mask. We retain instances whose text-prompt IoU standard deviation under the pretrained SAM 3 model exceeds 0.1, filtering out cases where all text prompts already produce nearly identical predictions. This filtering pipeline

yields $20{,}430$ training instances and $5{,}858$ validation instances, from the original $76{,}969$ training and $11{,}899$ validation annotations, respectively.

Table 6: Evaluation benchmarks used to evaluate the zero-shot transfer of BTR post-training. The first 6 benchmarks consist of object-centric instance segmentation annotations, while RAOS (Luo et al., 2024) consists of medical images with organ annotations which is completely out-of-domain.

Dataset	**Images**	**Instances**
GrabCut (Rother et al., 2004)	50	50
Berkeley (Martin et al., 2001; McGuinness & O'Connor, 2010)	96	100
DAVIS (Perazzi et al., 2016)	345	345
COCO-MVal (Lin et al., 2014; Sofiiuk et al., 2020)	800	800
TETRIS (Moskalenko et al., 2024)	2,000	2,531
LVIS (Gupta et al., 2019)	164k	$\sim$2M
RAOS (Luo et al., 2024)	413	$\sim$8k

GrabCut (Rother et al., 2004) contains 50 images each with a single foreground object and was created specifically for interactive segmentation. Berkeley (Martin et al., 2001) is a boundary detection dataset (BSDS300) adapted for per-instance interactive segmentation evaluation (McGuinness & O'Connor, 2010). DAVIS (Perazzi et al., 2016) provides 345 frames sampled from video sequences originally collected for video object segmentation. COCO-MVal (Lin et al., 2014; Sofiiuk et al., 2020) is a subset of COCO val2017 split containing 10 instances from each of the 80 object categories. TETRIS (Moskalenko et al., 2024) is a high-resolution benchmark (12 MP median) with 2,000 manually labeled images spanning people, transportation, wildlife, architecture, and other common categories. LVIS (Gupta et al., 2019) covers 1000+ entry-level categories with a long-tail distribution and is used for training in combination with COCO. For evaluation, we use the official validation split for LVIS. RAOS (Luo et al., 2024) comprises 413 CT scans ($\sim$80k slices) with 17–19 abdominal organs annotated per patient, including clinically challenging excision and partial-excision cases. All benchmarks are instance/object-centric, and we filter out masks that are less than 0.3% or larger than 80% of the original image resolution for cleaner experimental signals.

B Per-Pixel Bernoulli KL

The reference-KL term in equation 2 uses the per-pixel Bernoulli KL between the frozen pretrained reference and the current model. Let $\pi_{\theta,k}, \pi_{\text{ref},k} \in [0,1]^{H\times W}$ denote the per-pixel sigmoid outputs of the current model and the frozen reference under prompt q_k, with $\pi_{\theta,k}^{(i,j)} = \sigma(l_k^{(i,j)})$. Then

$$D_{D_{\mathrm{KL}}}^{\mathrm{px}}\left(\pi_{\text{ref},k} \,\|\, \pi_{\theta,k}\right) = \frac{1}{HW}\sum_{i,j}\left[\pi_{\text{ref},k}^{(i,j)} \log \frac{\pi_{\text{ref},k}^{(i,j)}}{\pi_{\theta,k}^{(i,j)}} + (1-\pi_{\text{ref},k}^{(i,j)}) \log \frac{1-\pi_{\text{ref},k}^{(i,j)}}{1-\pi_{\theta,k}^{(i,j)}}\right]. \tag{5}$$

The forward direction $D_{D_{\mathrm{KL}}}(\pi_{\text{ref}} \,\|\, \pi_\theta)$ is mode-covering: it penalizes the post-trained model for assigning low confidence where the pretrained reference was confident.

C Gradient Analysis

C.1 Decomposition

The gradient of $\mathcal{L}_{\text{BTR}}$ with respect to θ decomposes as

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{BTR}}}{\partial \theta} = \lambda_1 \underbrace{\frac{\partial \ell_{\text{seg}}(\hat{m}_{k^\star}, m^*)}{\partial \hat{m}_{k^\star}} \frac{\partial f_\theta(x, q_{k^\star})}{\partial \theta}}_{G_1\text{: GT anchor}} \\ + \lambda_2 \underbrace{\frac{1}{|\mathcal{G}|-1} \sum_{k \neq k^\star} \frac{\partial \ell_{\text{seg}}(\hat{m}_k, \text{sg}(\hat{m}_{k^\star}))}{\partial \hat{m}_k} \frac{\partial f_\theta(x, q_k)}{\partial \theta}}_{G_2\text{: BTR consistency}} \\ + \lambda_{D_{\text{KL}}} \underbrace{\frac{\partial}{\partial \theta} \overline{D}^{\text{px}}_{D_{\text{KL}}}}_{G_3\text{: reference KL}},\end{aligned} \tag{6}$$

where $\overline{D}^{\text{px}}_{D_{\text{KL}}}$ denotes the per-pixel KL averaged over the group.

G_1 **(GT signal).** A standard segmentation gradient applied at one prompt per object per training step. It improves the highest-IoU mask toward m^*.

G_2 **(consistency signal).** For a non-best prediction, let

$$p_j = \sigma(l_{k,j}), \qquad t_j = \text{sg}(\sigma(l_{k^\star,j}))\,,$$

where j indexes pixels. The implemented segmentation loss is

$$\ell_{\text{seg}} = \ell_{\text{focal}} + \ell_{\text{Dice}}.$$

Hence,

$$\frac{\partial \ell_{\text{seg}}(\hat{m}_k, \text{sg}(\hat{m}_{k^\star}))}{\partial l_{k,j}} = \frac{\partial \ell_{\text{focal}}}{\partial l_{k,j}} + \frac{\partial \ell_{\text{Dice}}}{\partial l_{k,j}}, \qquad k \neq k^\star. \tag{7}$$

For sigmoid focal loss, define

$$\begin{aligned} c_j &= -t_j \log p_j - (1-t_j)\log(1-p_j), \\ p_{t,j} &= p_j t_j + (1-p_j)(1-t_j), \end{aligned}$$

and

$$\alpha_{t,j} = \alpha t_j + (1-\alpha)(1-t_j).$$

Then

$$\frac{\partial \ell_{\text{focal}}}{\partial l_{k,j}} = \frac{\alpha_{t,j}}{HW}\left[(p_j - t_j)(1-p_{t,j})^\gamma - \gamma c_j (1-p_{t,j})^{\gamma-1}(2t_j - 1)p_j(1-p_j)\right]. \tag{8}$$

For the soft Dice loss, let

$$N = 2\sum_r p_r t_r + 1, \qquad D = \sum_r p_r + \sum_r t_r + 1.$$

Then

$$\frac{\partial \ell_{\text{Dice}}}{\partial l_{k,j}} = p_j(1-p_j)\frac{N - 2t_j D}{D^2}. \tag{9}$$

Three properties are important.

1. *Teacher-directed signal.* Both components use the stop-gradient best prediction as their target, so the non-best prediction is optimized relative to the selected teacher.
2. *No direct GT signal.* The consistency term contains no direct dependence on m^*; ground-truth supervision enters only through G_1 and the selection of $k^\star$.
3. *Coupling through shared weights.* Because θ is shared across all $f_\theta(x, q_k)$ for $q_k \in \mathcal{G}$, the update from G_2 adjusts one shared function so that predictions from different prompt conditions become more aligned.

G_3 **(regularization).** The forward-KL gradient pulls π_θ toward π_{ref}, preserving the pretrained model's coverage on inputs not seen during post-training.

C.2 Role of Stop-Gradient

The stop-gradient operator $\mathrm{sg}(\cdot)$ in the BTR consistency term is essential. Without it, the loss $\ell_{\text{seg}}(\hat{m}_k, \hat{m}_{k^\star})$ would propagate gradient through both arguments, yielding an additional gradient through the teacher branch that updates $\hat{m}_{k^\star}$ according to the non-best predictions. The unstopped consistency objective can then conflict with the GT anchor G_1: G_1 directly optimizes $\hat{m}_{k^\star}$ toward m^*, while the consistency term would simultaneously update the teacher according to student-dependent consistency gradients, which contain no direct ground-truth information. Stop-gradient resolves this conflict by blocking the direct consistency gradient through the teacher branch: under G_2, gradients are propagated only through the $|\mathcal{G}| - 1$ non-best prediction branches, toward a teacher that is itself anchored to m^* by G_1. This follows the design pattern of self-distillation methods such as SimSiam (Chen & He, 2021), which similarly block gradients through the target branch to avoid degenerate solutions.

C.3 Training Trajectory

The interaction between G_1 and G_2 produces a three-phase trajectory.

1. *Early.* The pretrained model produces reasonable but inconsistent masks. G_1 improves the best mask toward m^*. G_2 produces large gradients on inconsistent prompts.
2. *Mid.* The best mask approaches m^*. Non-best masks are pulled toward it via G_2. The consistency gap narrows as the stop-gradient target improves.
3. *Late.* Predictions in the group become increasingly aligned, reducing the discrepancy-driven consistency signal.

This trajectory is the formal motivation for the warmup phase of Section 4.4: G_2 is most useful once G_1 has improved the best mask enough to be a reliable teacher.

C.4 Comparison With Augmented SFT

In augmented SFT on K prompt–mask pairs, each prediction is supervised independently against the same ground-truth mask:

$$\frac{\partial \ell_{\text{seg}}(\hat{m}_k, m^*)}{\partial l_{k,j}} = \frac{\partial \ell_{\text{focal}}(\hat{m}_k, m^*)}{\partial l_{k,j}} + \frac{\partial \ell_{\text{Dice}}(\hat{m}_k, m^*)}{\partial l_{k,j}}, \tag{10}$$

with no loss term that directly couples predictions from different prompts. Thus, although all prompts share the same target mask, cross-prompt agreement is encouraged only indirectly through shared parameters and common supervision.

In BTR, the $|\mathcal{G}| - 1$ non-best predictions receive no direct GT supervision: their direct alignment signal is G_2, whose target $\mathrm{sg}(\hat{m}_{k^\star})$ changes as G_1 improves the selected teacher. BTR therefore explicitly couples predictions within each prompt group, encouraging non-best predictions to align with the current best prediction rather than optimizing each prompt–mask pair independently.

D Proof of Loss-Level Optimality

Proposition 1 (Loss-level optimality of prompt invariance)**.** *Let $\mathcal{L}^{(0)}_{BTR}$ denote the BTR loss in equation 2 with $\lambda_{D_{\mathrm{KL}}} = 0$. Assume realizability: there exists θ_0 such that $f_{\theta_0}(x, q) = m^*$ for every (x, m^*, q) in the data distribution. Then $\inf_\theta \mathcal{L}^{(0)}_{BTR}(\theta) = 0$, and any minimizer θ^* satisfies $f_{\theta^*}(x, q) = m^*$ for every prompt q in every prompt group — i.e., prompt invariance and correctness simultaneously.*

Proof. We show (i) that the infimum of $\mathcal{L}^{(0)}_{\text{BTR}}$ is zero and (ii) that every minimizer satisfies prompt invariance and correctness.

(i) The infimum is zero. By realizability, θ_0 produces $\hat{m}_k = m^*$ for every $k \in \mathcal{G}$. Hence $\hat{m}_{k^\star} = m^*$ and $\hat{m}_k = \hat{m}_{k^\star}$ for all k. Since m^* is binary and ℓ_{seg} is non-negative and vanishes when the prediction exactly matches the binary target,

$$\ell_{\text{seg}}(\hat{m}_{k^\star}, m^*) = 0 \qquad \text{and} \qquad \ell_{\text{seg}}(\hat{m}_k, \text{sg}(\hat{m}_{k^\star})) = 0 \text{ for every } k \neq k^\star, \tag{11}$$

so $\mathcal{L}^{(0)}_{\text{BTR}}(\theta_0) = 0$. Since $\mathcal{L}^{(0)}_{\text{BTR}} \geq 0$ everywhere, θ_0 attains the infimum.

(ii) Any minimizer is prompt-invariant and correct. Suppose θ^* achieves $\mathcal{L}^{(0)}_{\text{BTR}}(\theta^*) = 0$, and write $\hat{m}_k = f_{\theta^*}(x, q_k)$. Since the loss is a non-negative sum of non-negative terms, every term must vanish:

$$\ell_{\text{seg}}(\hat{m}_{k^\star}, m^*) = 0 \qquad \text{and} \qquad \ell_{\text{seg}}(\hat{m}_k, \text{sg}(\hat{m}_{k^\star})) = 0 \text{ for every } k \neq k^\star. \tag{12}$$

Since the GT target m^* is binary and ℓ_{seg} vanishes only when the prediction exactly matches a binary target, the first equation gives $\hat{m}_{k^\star} = m^*$. Therefore, the stop-gradient target in every consistency term is also m^*. The second equation then gives $\hat{m}_k = m^*$ for every $k \neq k^\star$. Combining, $f_{\theta^*}(x, q_k) = m^*$ for every $q_k \in \mathcal{G}$. □

Remarks.

1. *Same minimum as augmented SFT.* The same argument applied to $\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}\big[\frac{1}{|\mathcal{G}|}\sum_k \ell_{\text{seg}}(\hat{m}_k, m^*)\big]$ shows BTR and augmented SFT share the global minimum location. They differ in the loss landscape away from it (Appendix C).
2. *Effect of the reference KL.* Adding $\lambda_{D_{\text{KL}}} > 0$ shifts the global minimum unless θ_0 already matches the pretrained reference, introducing a tradeoff between alignment, correctness, and proximity to the reference.
3. *When realizability fails.* The infimum becomes strictly positive, and the minimizer balances correctness at the best prompt against agreement among $\hat{m}_k$. The distinction from augmented SFT remains the optimization signal (Appendix C).
4. *Local minima.* The proposition characterizes global minimizers; gradient descent is not guaranteed to reach one. Warmup (Section 4.4) addresses this in practice.

E Connection to Self-Distillation Regularization

BTR is a self-distillation method: the teacher is the model's own highest-IoU prediction in a prompt group, under stop-gradient. Mobahi et al. (2020) prove that, in kernel regression, repeated self-distillation shrinks the effective hypothesis class by attenuating low-singular-value components, which acts as an implicit regularizer beyond the explicit supervised loss. Their result is for previous-iterate distillation, so the formal statement does not transfer literally to BTR's best-prompt teaching. The qualitative claim – that consistency-targeted self-distillation contributes implicit regularization on top of the supervised signal – is part of the rationale for the BTR consistency term, and is consistent with the held-out transfer reported in Section 5.2.

F Prompt Variant Generation

F.1 Noun-Phrase Generation Prompt

The full prompt passed to Qwen3-VL-8B for noun-phrase generation is shown in Figure 4. `{category_name}` is replaced with the COCO-style category name of the target instance.

```
You are given TWO images of the SAME target object "{category_name}":
- Image 1: the full scene. Everything OUTSIDE the target has
been replaced by a gray/white CHECKERBOARD pattern. Ignore the
checkerboard -- it contains no information.
- Image 2: a zoomed-in crop of the target, with the same
checkerboard background.

Generate exactly 4 short text descriptions that ALL refer to the
WHOLE target object (the non-checkerboard pixels). Do NOT describe a
sub-part, the checkerboard, or anything else.

Hard constraints:
- All 4 descriptions must refer to the SAME whole {category_name}.
- Each description must be GROUNDED in what you actually see -- do
not invent attributes.
- Each axis answers a DIFFERENT QUESTION about the object. Do not
return paraphrases of the same answer.

Generate one description per axis, in this exact order:

1. APPEARANCE (2-4 words). Question: "What does it look like?"
Visible attributes (color, material, shape, size) + the object noun.
Example for a pencil: "yellow wooden pencil"
Example for a sofa: "gray leather sofa"
Example for a basket: "small woven basket"

2. ALTERNATIVE_NOUN (1-2 words). Question: "What else can you call
it?"
A different noun for the SAME whole object. Not a part. Not a
related thing.
May be a synonym, hypernym, or specific subcategory -- anything that
names
the whole object differently.
Example for a pencil: "writing tool"
Example for a sofa: "couch"
Example for a basket: "container"

3. NO_NAME_DESCRIPTION (3-5 words). Question: "Describe it without
naming it."
Describe the object using its visible features (shape, material,
parts) but
WITHOUT using the word "{category_name}" or any close synonym of it.
Example for a pencil: "thin wooden cylinder with point"
Example for a sofa: "upholstered seating with cushions"
Example for a basket: "open woven storage holder"

4. FUNCTIONAL_PHRASE (3-5 words). Question: "What is it used for?"
Start with a verb. Describe the function or typical use of the
object,
without including the object's name.
Example for a pencil: "Used for writing on paper"
Example for a sofa: "Provides seating for relaxing"
Example for a basket: "Holds items for storage"

Return ONLY valid JSON, no markdown:
{{"variants": ["appearance", "alternative_noun",
"no_name_description", "functional_phrase"]}}
```

Figure 4: Generation prompt passed to Qwen3-VL-8B for noun-phrase variant generation.

F.2 Verification Prompt

The verification prompt passed to Qwen3-VL-32B-AWQ is shown in Figure 5. `{category_name}` is replaced with the COCO-style category name and `{variant_text}` with the noun-phrase variant being verified.

```
You see TWO images of a target object "{category_name}":
- Image 1: full scene, everything outside the target replaced by
gray/white checkerboard.
- Image 2: zoomed crop of the target, same checkerboard background.

A text description was generated for this object:
"{variant_text}"

Judge: is this description an ACCEPTABLE way to refer to the
highlighted object?

DEFAULT TO PASS. Only fail if you are CONFIDENT the description is
wrong.

PASS if the description:
- Refers to the same object (even loosely)
- Does not contain clearly wrong visual attributes
- Could reasonably be used by someone to point at this object

These are all PASS examples -- do NOT fail descriptions like these:
- "seat" for a chair ✓
- "screen" for a TV ✓
- "container" for a box ✓
- "individual" or "person" for any human ✓
- "used for sitting" for a chair ✓
- "holds items" for a box ✓
- "floor covering" for a rug ✓
- Generic synonyms, functional descriptions, or category-level words
are ALL fine.

FAIL only if:
- wrong_object: describes a COMPLETELY different object (e.g., "soap
bar" for a blanket)
- hallucinated_attribute: mentions a SPECIFIC color/shape that is
clearly wrong (e.g., "red car" when it's blue)

Do NOT fail for: being generic, using a synonym, describing
function, or naming a closely-related concept.

Reply with ONLY valid JSON:
If PASS: {{"verdict": "PASS"}}
If FAIL: {{"verdict": "FAIL",
          "reason": "wrong_object|hallucinated_attribute",
          "corrected": "your corrected 1-5 word description"}}
```

Figure 5: Verification prompt passed to Qwen3-VL-32B-AWQ for filtering and regenerating noun-phrase variants.

F.3 Manual Inspection Protocol

We sample 100 instances at random from the post-training set after generation, yielding 400 noun-phrase variants (4 per instance), and inspect each manually. A phrase is marked as an error if it is inaccurate or ambiguous as a referring expression for the target instance, or if it is too long for as SAM 3 is designed to consume short noun phrases rather than full sentences. We trigger the verification stage of Section 4.2 when the error rate on the inspection sample exceeds 3%. We use 3% rather than zero because prompt optimization alone cannot eliminate residual VLM errors. We empirically observe that the additional verification stage significantly reduces the error rate.

F.4 Bounding-Box Variants

Train-time generation. For prompt-group construction at training time, we generate bounding-box variants from each ground-truth mask using five operations applied to the tight ground-truth box.

1. the tight ground-truth box (no perturbation).
2. translate the box in x and y by $\pm 10\%$ of width/height, preserving size.

3. independently expand each of the four sides outward by 0–20% of width/height per side.
4. shrink uniformly around the box center to 80–100% of the original size.
5. independently scale the x and y axes around the center to 80–120% per axis.

The number of boxes per group, num_boxes, controls how the operations are sampled:

- num_boxes $= 1$: *tight* only.
- num_boxes ≤ 5: *tight* plus num_boxes $- 1$ operations sampled without replacement from the remaining four.
- num_boxes > 5: *tight* plus all four operations, with additional boxes sampled with replacement from the four perturbations.

Evaluation-time perturbation. For evaluation we use a fixed set of $T_{\text{box}} = 11$ box perturbations per instance: the tight ground-truth box, a 20% shrink and a 20% expansion around the box center, and translations in eight directions by 10% of width/height (up, down, left, right, and the four diagonals).

F.5 Detailed RClicks Evaluation

Table 7 provides the full per-dataset comparison under real-user clicks and simulated-click perturbations using User@1, Sample@1, and NSR@1.

F.6 Detailed RClicks Evaluation

Table 7 provides the full per-dataset comparison under real-user clicks and simulated-click perturbations using User@1, Sample@1, and NSR@1.

G Additional Analysis

G.1 Wrong-but-Consistent Predictions

To analyze wrong-but-consistent predictions, we evaluate 5,858 EntitySeg instances under the same 15 clicks across model states. Stability is the mean pairwise IoU among 15 click-conditioned masks, and accuracy is their mean GT IoU. Using thresholds of 0.90 for stability and 0.80 for correctness, respectively, we categorize predictions as stable-correct (SC), stable-but-wrong (SW), unstable-correct (UC), or unstable-wrong (UW), where "correct" and "wrong" denote above and below the selected threshold.

Checkpoint 1 marks the end of warmup. During post-warmup training, SC increases from 60.12% to 68.93%, while UW decreases from 23.78% to 14.46%. Overall, 453 instances (7.73%) transition from below to above the accuracy threshold, compared with 92 (1.57%) in the opposite direction. Moreover, 97.02% of existing SC instances remain SC.

Although 328 instances (5.60%) newly enter SW by checkpoint 5, 281 of them (85.7%) originate from UW and were therefore already below the accuracy threshold. Only 47 instances (0.80%) transition from an above-threshold state to SW, including 41 SC-to-SW cases (0.70%). Thus, post-warmup BTR predominantly improves accuracy and stability, while a smaller but non-negligible wrong-but-consistent population remains at checkpoint 5 (14.36%).

G.2 Category-wise Analysis

We conduct a category-wise analysis of click-prompt performance over the 49 categories with at least 15 instances in the EntitySeg evaluation set. Table 10 reports representative thin or structurally complex categories.

H Additional Discussion

Complementary to architectural prompt-robustness methods. A line of recent work addresses SAM's prompt sensitivity through architectural modifications: deformable-attention plugins (Fan

Table 7: Detailed RClicks evaluation under real-user clicks and simulated-click perturbations. User@1 and Sample@1 are reported as mean±standard deviation, and lower NSR@1 is better.

Method	Model	Data	GrabCut			Berkeley			DAVIS			COCO-MVal			TETRIS		
			User@1 (±std)	Sample@1 (±std)	NSR@1↓	User@1 (±std)	Sample@1 (±std)	NSR@1↓	User@1 (±std)	Sample@1 (±std)	NSR@1↓	User@1 (±std)	Sample@1 (±std)	NSR@1↓	User@1 (±std)	Sample@1 (±std)	NSR@1↓
GPCIS	RN50	C+L	83.41±8.67	83.49±7.68	11.63	78.79±8.30	79.15±8.04	12.24	72.53±9.71	73.66±7.39	16.30	64.04±9.06	64.43±7.95	18.21	71.99±10.13	71.75±9.73	17.24
CDNet	RN34	C+L	87.85±10.22	88.27±8.19	12.50	81.92±10.42	82.47±9.45	14.69	71.97±11.50	73.41±8.24	19.72	76.05±10.28	76.80±8.00	16.32	79.56±11.38	79.61±10.84	17.32
CDNet	RN34	SBD	85.27±10.06	84.88±9.36	13.39	81.04±9.06	81.58±7.75	13.07	69.16±11.18	70.50±7.68	19.96	65.20±9.74	65.72±8.23	19.07	73.45±10.97	73.22±10.69	18.51
RITM	HR18	C+L	86.10±7.57	86.11±7.34	10.53	82.55±6.09	82.33±5.75	9.37	67.30±8.63	67.80±7.99	20.88	73.81±8.34	74.09±6.86	17.44	76.99±7.11	76.99±7.09	12.89
RITM	HR18s-IT	C+L	83.60±10.15	83.12±9.83	13.68	77.51±9.11	77.55±8.45	13.57	66.50±11.70	67.86±8.66	24.22	71.98±10.05	72.37±8.03	19.00	73.25±9.88	73.29±9.19	16.66
RITM	HR18-IT	C+L	85.85±7.59	85.25±7.28	10.18	82.08±6.48	82.43±5.20	9.42	68.72±10.38	69.69±8.44	21.86	75.55±9.02	76.04±7.23	17.00	77.13±7.13	77.08±6.79	11.96
RITM	HR32-IT	C+L	86.20±7.99	86.12±7.16	10.79	82.01±7.11	82.38±6.13	10.66	69.13±10.96	70.35±8.23	22.71	74.91±10.20	75.34±7.97	18.40	76.98±8.18	76.95±7.66	13.93
RITM	HR18-IT	SBD	79.40±3.86	79.28±4.38	7.35	76.34±4.90	76.50±4.06	8.34	64.47±9.05	65.14±7.61	21.64	61.59±7.02	61.72±6.26	15.19	71.19±4.85	71.17±4.60	9.92
AdaptClick	ViT-B	C+L	89.26±4.47	89.19±4.87	6.05	83.45±5.56	84.10±4.16	8.87	74.92±7.09	76.18±5.94	13.45	77.17±6.49	77.53±5.26	12.87	82.16±6.15	82.18±6.07	10.82
AdaptClick	ViT-B	SBD	90.31±2.38	90.41±2.95	3.61	83.73±3.56	84.28±2.94	5.41	76.57±6.21	77.91±4.91	11.28	61.85±6.86	62.03±6.41	16.10	76.49±4.36	76.43±4.32	8.98
SimpleClick	ViT-B	C+L	89.20±3.78	89.12±4.55	4.89	84.13±5.73	85.02±4.22	8.50	72.08±7.34	73.11±6.28	14.67	76.48±6.98	76.68±5.98	13.23	81.03±6.58	80.96±6.42	11.36
SimpleClick	ViT-L	C+L	88.51±4.48	88.79±4.78	6.34	83.66±4.88	84.25±3.76	7.72	74.35±6.71	75.16±5.59	14.34	80.34±6.16	80.62±5.17	11.74	83.47±6.98	83.48±6.83	11.16
SimpleClick	ViT-H	C+L	87.91±4.72	88.49±4.25	7.30	83.16±4.71	83.75±3.52	7.49	72.44±7.56	73.22±6.19	16.76	80.48±5.98	80.74±4.65	10.85	83.88±6.01	83.79±5.78	10.47
SimpleClick	ViT-XT	SBD	78.01±7.50	77.48±8.28	12.15	69.47±8.93	69.26±8.88	16.24	52.56±10.30	53.08±9.59	24.75	50.43±9.13	50.21±8.74	22.41	58.03±10.06	57.71±10.01	21.58
SimpleClick	ViT-B	SBD	88.84±2.67	88.94±2.86	3.74	83.10±4.24	83.67±3.87	6.50	74.22±6.11	75.56±4.87	11.20	60.56±6.59	60.52±6.08	15.56	75.05±4.45	74.99±4.45	8.69
SimpleClick	ViT-L	SBD	88.48±3.56	88.49±3.75	6.12	83.41±3.29	84.01±2.98	5.18	76.25±6.21	77.67±4.45	11.59	64.97±6.71	64.96±5.96	15.57	77.76±4.14	77.65±4.18	8.85
SimpleClick	ViT-H	SBD	88.54±2.09	88.32±2.47	3.01	82.46±3.08	83.03±2.42	4.92	74.96±7.10	76.44±5.64	13.14	65.52±6.39	65.71±6.01	14.60	77.72±3.11	77.72±3.16	6.69
CFR-ICL	ViT-H	C+L	88.95±4.72	89.33±4.86	7.29	85.05±5.88	85.75±4.59	9.37	75.48±8.23	76.43±6.71	17.09	81.18±6.55	81.47±5.28	12.34	85.92±7.02	85.86±6.68	11.47
MobileSAM	ViT-Tiny	SA-1B	76.61±14.67	77.02±13.49	24.45	75.30±15.97	76.03±14.91	27.28	58.17±16.01	59.48±14.55	40.95	66.65±15.01	66.85±13.09	32.55	66.81±18.77	67.12±18.60	36.21
SAM	ViT-B	SA-1B	64.65±21.07	63.52±20.47	42.61	66.02±20.18	66.81±18.58	41.66	45.27±17.48	46.55±16.41	57.06	60.90±18.20	61.20±15.88	44.16	55.05±21.59	54.99±21.25	51.19
SAM	ViT-L	SA-1B	75.06±17.92	74.27±18.13	31.16	75.48±17.33	76.87±15.75	31.29	57.47±16.86	58.75±15.67	45.10	66.90±16.85	66.80±14.75	36.12	66.98±20.15	67.18±19.78	41.32
SAM	ViT-H	SA-1B	75.19±16.43	72.94±17.31	28.46	74.13±17.30	75.28±15.82	32.63	54.88±16.88	56.15±15.34	47.70	67.26±17.00	67.46±14.94	35.88	63.88±20.39	64.34±20.03	42.55
SAM-HQ	ViT-B	SA-1B	61.04±22.55	60.62±21.47	47.48	61.41±21.19	62.46±20.22	49.09	42.99±16.76	44.77±15.99	56.99	58.21±18.68	58.25±16.35	48.36	48.03±20.88	47.80±20.58	58.47
SAM-HQ	ViT-L	SA-1B	70.29±21.94	69.26±20.11	38.31	68.20±18.76	69.62±18.09	40.26	50.73±17.63	52.01±16.82	50.86	62.03±18.38	62.05±15.94	42.80	58.30±21.43	58.48±21.04	49.68
SAM-HQ	ViT-H	SA-1B	67.97±19.63	66.56±17.44	35.33	66.21±17.92	67.74±17.24	39.68	48.68±16.28	50.30±15.05	50.54	63.07±17.92	63.00±15.59	40.89	55.19±19.26	55.30±18.94	46.68
SAM 2	Hiera-T	SA-V	76.80±19.32	77.24±17.90	30.87	76.13±15.15	77.66±14.17	26.23	66.32±16.48	67.67±14.70	33.39	66.48±15.87	66.80±13.43	34.98	45.90±18.61	45.88±18.32	55.51
SAM 2	Hiera-S	SA-V	72.22±16.88	72.62±15.92	29.73	73.59±16.20	74.50±14.81	29.32	61.96±17.75	63.84±16.90	39.49	66.57±16.38	66.91±14.26	35.31	45.56±18.56	45.52±18.21	56.13
SAM 2	Hiera-B+	SA-V	74.60±21.37	75.02±19.14	34.70	76.38±17.20	77.44±15.72	28.62	62.17±17.84	63.69±16.55	39.86	65.92±17.78	65.99±15.24	38.82	49.40±19.08	49.31±18.63	52.91
SAM 2	Hiera-L	SA-V	83.08±16.58	83.78±14.15	24.71	80.77±15.25	81.45±13.05	23.24	73.25±14.62	74.69±12.89	25.44	74.42±13.61	74.83±11.33	25.93	50.56±19.00	50.29±18.68	51.62
SAM 2.1	Hiera-T	SA-V	74.13±20.60	73.79±19.29	34.54	74.69±16.71	76.19±15.20	30.45	63.52±18.10	64.65±16.67	38.59	64.93±16.35	65.14±14.07	37.46	45.98±18.83	45.82±18.48	56.18
SAM 2.1	Hiera-S	SA-V	72.85±19.23	72.91±17.96	32.05	75.00±16.24	76.03±15.15	28.94	66.15±17.13	67.78±15.84	35.37	66.58±16.49	66.87±14.13	36.02	45.80±18.70	45.71±18.37	56.39
SAM 2.1	Hiera-B+	SA-V	72.19±22.57	72.17±21.18	38.18	74.46±18.27	75.30±16.85	31.82	57.69±19.10	59.03±17.75	46.23	65.27±17.93	65.16±15.43	40.14	48.43±18.59	48.32±18.18	52.70
SAM 2.1	Hiera-L	SA-V	78.15±20.94	78.07±18.45	32.10	77.44±16.78	78.49±14.85	28.02	70.58±17.09	72.40±14.91	31.05	73.23±14.75	73.54±12.87	27.93	48.50±18.76	48.17±18.41	53.51
SAM (ViT-H) + BTR	ViT-H	EntitySeg(Finetuning)	83.43±15.63	82.87±16.49	22.46	80.19±16.48	81.96±15.28	25.74	69.37±15.87	70.28±14.33	29.77	71.40±18.13	71.26±16.80	31.90	78.12±20.14	77.86±20.39	31.30
SAM 3	–	–	74.94±16.26	75.55±15.56	29.28	80.45±12.26	81.07±11.16	21.03	70.37±12.98	71.98±11.24	25.44	72.87±15.26	73.33±13.21	30.07	47.89±19.17	47.80±18.79	55.24
SAM 3 + BTR	–	EntitySeg(Finetuning)	**89.28±3.38**	88.75±4.35	5.51	**86.59±3.32**	87.01±3.06	5.63	**76.68±7.41**	77.67±5.92	14.34	**79.42±6.72**	79.59±5.52	12.81	**88.54±5.99**	88.34±5.91	9.27

Table 8: Distribution of prediction states across model states.

Model state	SC (%)	SW (%)	UC (%)	UW (%)
Frozen SAM 3	34.35	34.14	3.09	28.42
BTR checkpoint 1	60.12	11.20	4.90	23.78
BTR checkpoint 5	68.93	14.36	2.25	14.46

Table 9: Transition matrix from BTR checkpoint 1 to checkpoint 5. Percentages are computed over all 5,858 evaluation instances.

Checkpoint 1 ↓ / Checkpoint 5 →	SC	SW	UC	UW
SC	3,417 (58.33%)	41 (0.70%)	30 (0.51%)	34 (0.58%)
SW	85 (1.45%)	513 (8.76%)	10 (0.17%)	48 (0.82%)
UC	227 (3.88%)	6 (0.10%)	43 (0.73%)	11 (0.19%)
UW	309 (5.27%)	281 (4.80%)	49 (0.84%)	754 (12.87%)

et al., 2025), variance-insensitive refinement modules (Fang et al., 2024), perturbed-prompt fine-tuning (Rahman et al., 2024), noise-tolerant click formulations (Zhang et al., 2025), and probabilistic prompt encoders (Wang et al., 2025). BTR is a loss, not an architecture, and is therefore orthogonal to these methods: the BTR consistency term could be applied on top of any of the above without changing their architectural contributions. We leave such combinations to future work.

Pretraining provides the capacity; the post-training objective induces invariance. BTR raises EntitySeg click $\mathrm{IoU}_{\mathrm{avg}}$ from 47.84 to 81.97 on SAM 3 (Table 1) with no architectural changes — only a change of loss. Matched augmented SFT recovers most of this improvement over the frozen model, showing that exposure to diverse prompts is itself a strong source of robustness. These results suggest that SAM 3 already contains substantial capacity for prompt-robust prediction, while the post-training objective determines how consistently that capacity is expressed across prompt variations. This perspective motivates BTR-style consistency objectives as a general post-training mechanism for promoting prompt robustness without modifying the underlying architecture.

I Additional Qualitative Results

Table 10: Category-wise click-prompt performance on representative thin or structurally complex EntitySeg categories.

Category	Frozen SAM 3	BTR (click)	Gain
electric_wire	0.1790	0.3322	+0.1532
rope	0.3622	0.5198	+0.1576
cable	0.4087	0.5329	+0.1242
branch	0.4004	0.6350	+0.2346
leaf	0.3204	0.6743	+0.3539

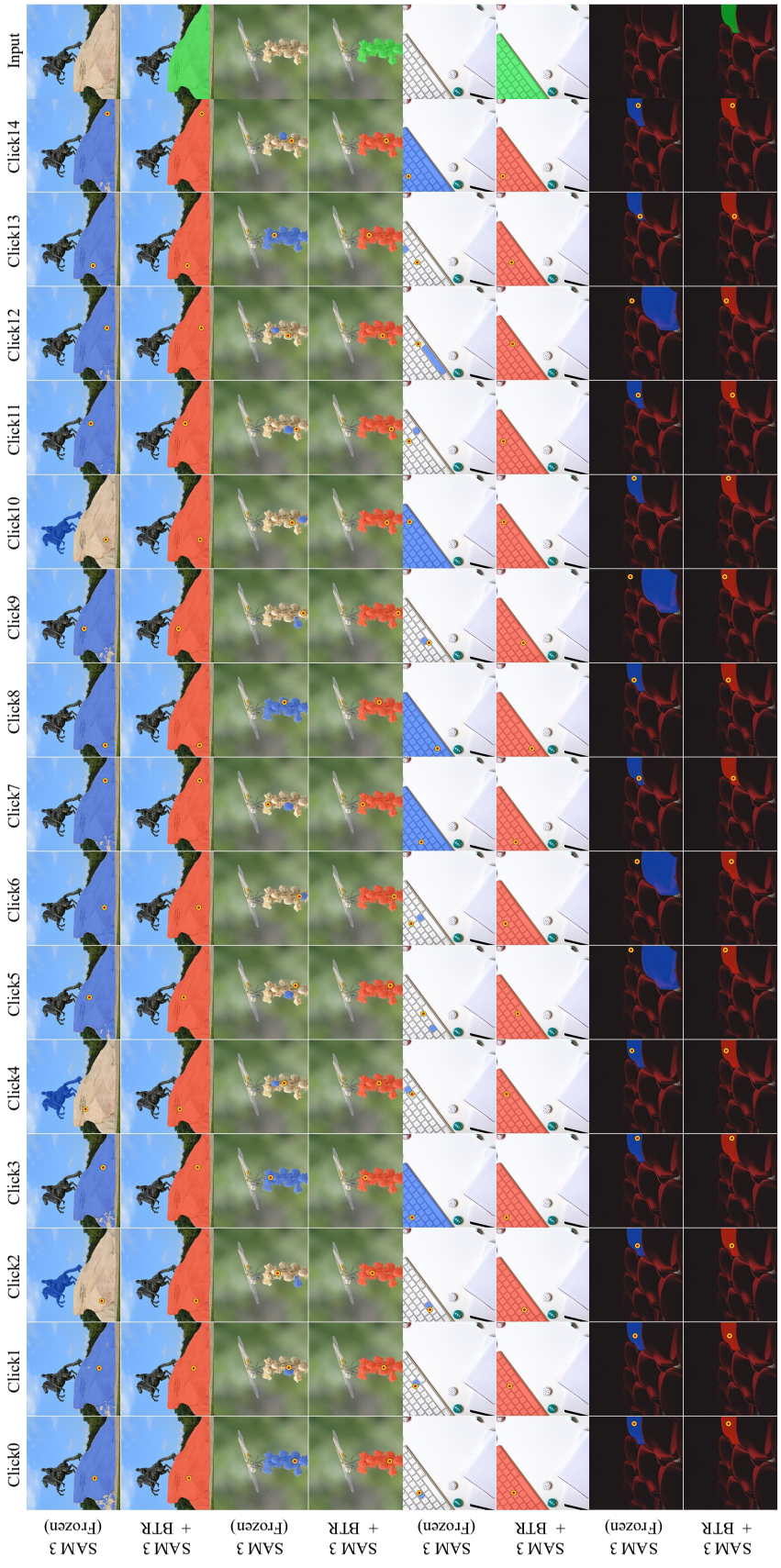


Figure 6: Qualitative results under click perturbations. Each column corresponds to a different click prompt referring to the same target object. Frozen SAM 3 often produces inconsistent masks across click locations, while SAM 3 + BTR yields substantially more stable predictions across the prompt group.

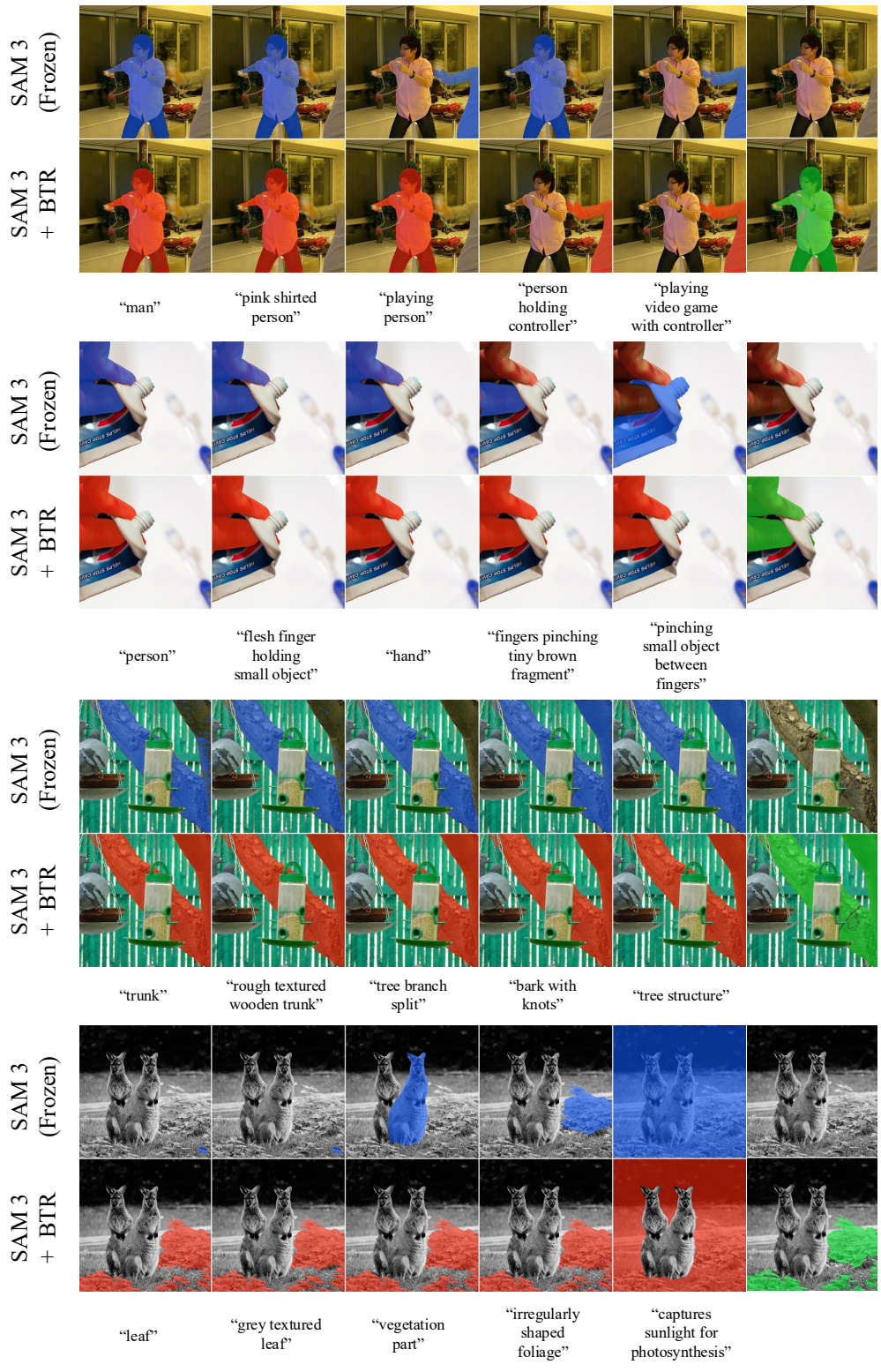


Figure 7: Qualitative results under noun-phrase perturbations. Each column corresponds to a different noun phrase referring to the same target object. Frozen SAM 3 often produces inconsistent or semantically drifted masks under paraphrased prompts, while SAM 3 + BTR yields substantially more stable predictions across semantic prompt variations.

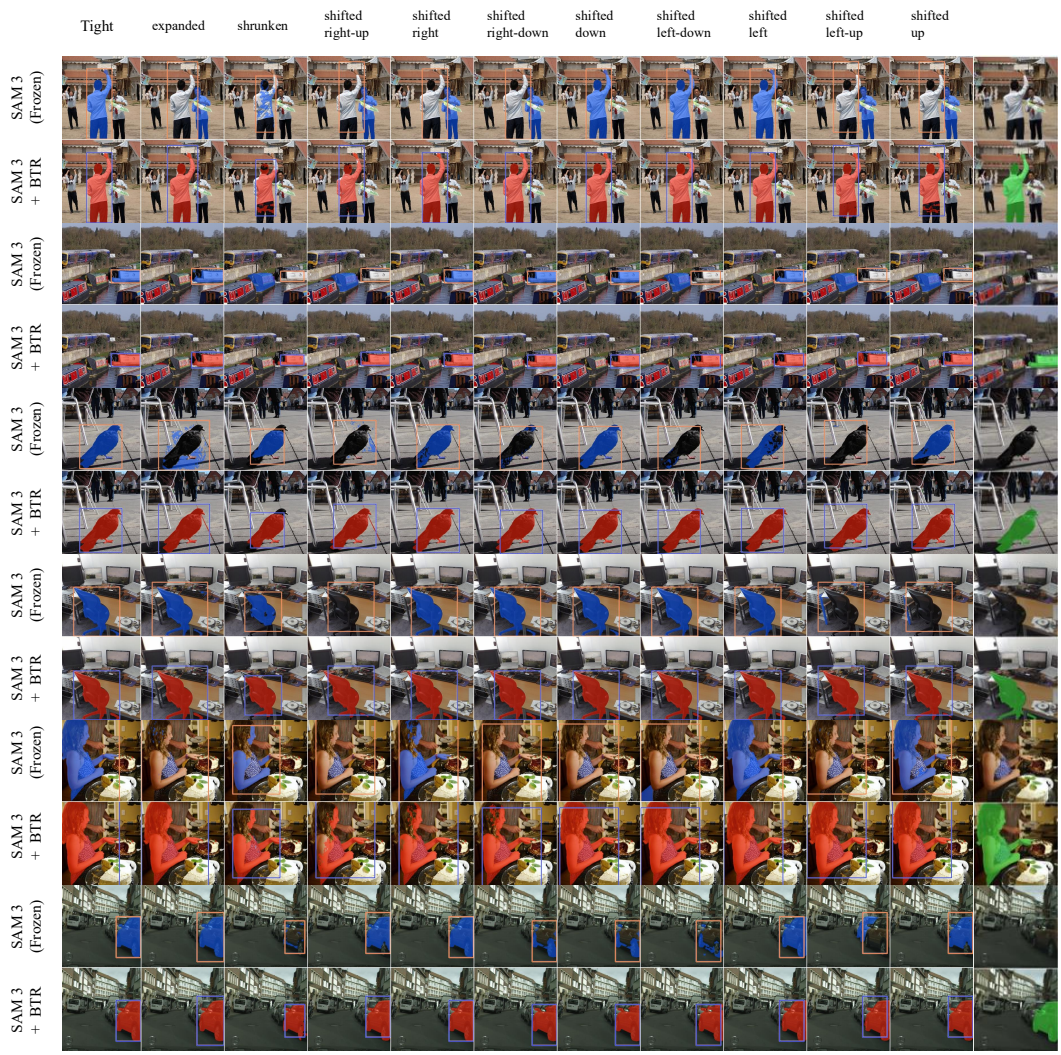


Figure 8: Qualitative results under bounding-box perturbations. Although BTR(click) training does not use box prompts, SAM 3 + BTR produces substantially more stable predictions under shifted, expanded, and shrunken box prompts than the frozen SAM 3 baseline.