# Make the Pertinent Salient: Task-Relevant Reconstruction for Visual Control with Distraction

**Kyungmin Kim, Charless Fowlkes, and Roy Fox**
Department of Computer Science,
University of California, Irvine, USA

## Abstract

Model-Based Reinforcement Learning (MBRL) has been a powerful tool for visual control tasks. Despite improved data efficiency, it remains challenging to use MBRL to train agents with generalizable perception. Training with visual distractions is particularly difficult due to the high variation they introduce to representation learning. Building on Dreamer, a popular MBRL method, we propose a simple yet effective auxiliary task – to reconstruct task-relevant components only. Our method, Segmentation Dreamer (SD), works either with ground-truth masks or by leveraging potentially error-prone segmentation foundation models. In DeepMind Control suite tasks with distraction, SD achieves significantly better sample efficiency and greater final performance than comparable methods. SD is especially helpful in a sparse reward task otherwise unsolvable by prior work, enabling the training of a visually robust agent without the need for extensive reward engineering.

## 1 Introduction

Among recent advances in MBRL (Sutton, 1991; Ha & Schmidhuber, 2018; Hansen et al., 2022), the Dreamer family (Hafner et al., 2020; 2021; 2023) has shown great promise in diverse visual control tasks, achieving high sample efficiency. This is accomplished by close cooperation between a world model and an actor–critic agent. The world model learns to emulate the environment's forward dynamics and reward function in a latent state space, and the agent is trained by interacting with this world model in place of the original environment.

Dreamer employs image reconstruction as an auxiliary task in world model training to facilitate representation learning (Fig. 1a). In environments with little distraction, image reconstruction works effectively by delivering rich learning signals. In the presence of distractions, however, the image reconstruction task encourages the encoder to keep all image information regardless of task relevance, which wastes model capacity (Fu et al., 2021) and degrades sample efficiency.

Prior approaches (Zhang et al., 2021; Nguyen et al., 2021; Deng et al., 2022; Fu et al., 2021; Bharadhwaj et al., 2022) work around the noisy reconstruction problem by devising reconstruction-free auxiliary tasks. However, many of them suffer from sample inefficiency, requiring many trajectories to isolate the task-relevant information. Moreover, training with these methods becomes more challenging in sparse reward environments where the signal for task relevance is very weak.

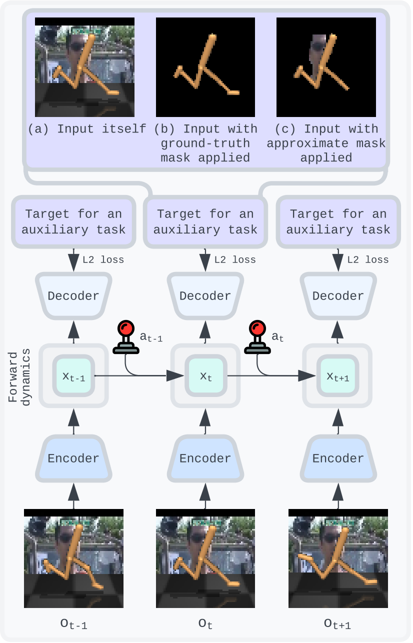

(a) Input itself
(b) Input with ground-truth mask applied
(c) Input with approximate mask applied
Target for an auxiliary task
Target for an auxiliary task
Target for an auxiliary task
L2 loss
L2 loss
L2 loss
Decoder
Decoder
Decoder
$a_{t-1}$
$a_t$
Forward dynamics
$x_{t-1}$
$x_t$
$x_{t+1}$
Encoder
Encoder
Encoder
$o_{t-1}$
$o_t$
$o_{t+1}$

Figure 1: World model learning with 3 choices of an auxiliary task target.

Our proposed solution takes advantage of the observation that identifying task-relevant components within images is often straightforward with a bit of domain knowledge. Given this assumption, we introduce a simple yet effective auxiliary task to reconstruct only the task-related components of image observations (Fig. 1b). We achieve this by using segmentation masks of task-related objects, which are readily available in simulations. By doing this, the world model can now learn features from a rich pixel-reconstruction loss signal without being hindered by the noise of visual distractions.

To make our auxiliary task more practical, we present a way of using it with segmentation estimates. This is made possible by the recent advances in segmentation foundation models (Kirillov et al., 2023; Zhang et al., 2023). Specifically, we leverage PerSAM (Zhang et al., 2023) finetuned with just **a single pair of training data** and use it to generate pseudo-labels (Fig. 1c). We further enhance robustness to prediction errors by identifying pixels where pseudo-labels may be wrong but the world model decoder is correct, ignoring $L_2$ loss for such pixels to avoid providing wrong signals.

We demonstrate the effectiveness of our method on six tasks in DeepMind Control Suite (DMC-1M) (Tassa et al., 2018), perturbed with visual distraction. Training with ground-truth masks, $\text{SD}_{\text{GT}}$, in the presence of distraction reaches the performance of training in standard environment with little distraction. Training with approximate masks, $\text{SD}_{\text{approx.}}$, also shows impressive performance, often matching $\text{SD}_{\text{GT}}$, with the help of the selective $L_2$ loss. Our experiment shows that $\text{SD}_{\text{approx.}}$ achieves higher sample efficiency than previous approaches and higher or comparable with those w.r.t. final performance.

## 2 Method

**Task-Relevant Reconstruction as an Auxiliary Task.** Under the assumption that task-relevant parts are easily identifiable within images with domain knowledge, we propose a new image reconstruction-based auxiliary task that spotlights task-relevant regions. Specifically, we employ a task-relevant segmentation mask applied RGB image (Fig. 1b) as a target to reconstruct. Since the reconstruction target only contains parts that are salient to a downstream task, learned latent representations would also only focus on important regions concerning the task. By explicitly avoiding capturing task-irrelevant parts, the latent dynamics can also become much simpler and easier to learn than the original over-complicated dynamics, allowing more sample-efficient training. We term the variant of Dreamer trained with this new auxiliary task Segmentation Dreamer (SD).

**Task-Relevant Reconstruction with Approximations.** To make our auxiliary task more useful in practice, e.g. when no GT mask is available during training, we integrate a segmentation foundation model into the pipeline. Among off-the-shelf segmentation models, we choose PerSAM-F (Zhang et al., 2023) as it can obtain a personalized segmentation model by finetuning on a *single* in-domain data, consisting of an RGB image and a segmentation mask. For the RGB, we obtain an image observation corresponding to a state sampled from the initial state distribution and manipulate it to keep the RGB values of task-relevant pixels and fill the rest with zeros, which in effect makes task-irrelevant pixels black-colored. For the segmentation mask, the regions of interest are filled with one, otherwise zero. Once finetuning is complete, we incorporate the PerSAM-F model into the SD pipeline to create pseudo-labels for the auxiliary task.

**A Strategy to Improve Error Robustness.** Although PerSAM provides decent mask predictions, it is inevitable to have some errors in the prediction, as illustrated in Fig. 1c. For example, PerSAM on videos has flickering effects since each frame is handled independently. Missing information, i.e. false negatives, would be particularly detrimental when combined with the naive $L_2$ loss on image reconstruction. With occasionally missing task-relevant parts in auxiliary targets, the encoder may be trained to encode a complete agent embodiment or to drop some task-related information. This would lead to large variances in latent representations and confuse the forward dynamics learning in the world model. However, SD can still produce correct results despite noisy targets, as long as most of the data is accurately labeled during training. When this happens, it is not desirable to flow gradients from those regions where the pseudo-label is wrong but the SD prediction is correct.

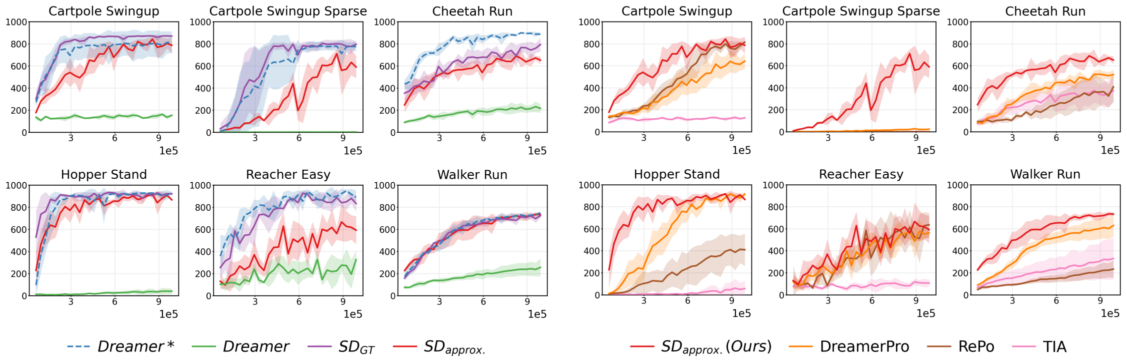
If we did, the incorrect target would provide a misleading signal to the model. We address this by masking out pixels in which to omit the $L_2$ loss computation, which we referred to as a selective $L_2$ loss. To do so, we present a heuristic method of estimating which pixels should not be included in the $L_2$ loss. Details, including formal descriptions, can be found in the Appendix A.

## 3 Experiments

We evaluate our method on six tasks in DeepMind Control Suite 1M (DMC-1M) (Tassa et al., 2018). The standard DMC environment comes with a simple background with little distraction. To create distracting environments, the background is replaced with color videos randomly sampled from the 'driving car' class in the Kinetics 400 dataset (Kay et al., 2017), similar to the setups in past work (Zhang et al., 2021; Nguyen et al., 2021; Deng et al., 2022).

**Comparison with Dreamer.** We first study how our methods, **SD$_{\mathbf{GT}}$** and **SD$_{\mathbf{approx.}}$**, the latter trained with a selective $L_2$ loss, compare to the standard Dreamer. **Dreamer*** is the standard Dreamer trained in *standard* DMC whereas **Dreamer** is trained in distracting DMC. In effect, we would expect Dreamer* to be an upper bound of all the Dreamer-based methods (including SD) trained on distracting environments, to the extent that the original environments are distraction-free. Fig. 2a shows evaluation returns during training. As expected, Dreamer struggles in all tasks because task-irrelevant information in reconstruction targets makes forward dynamics training difficult and confuses the agent training. On the other hand, SD$_{\mathrm{GT}}$ is on par with Dreamer* in most tasks. Notably, it achieves a somewhat better curve in a few tasks, e.g. Cartpole Swingup (Sparse) and Hopper Stand, which tells us that even a little but non-zero distraction in the standard DMC (e.g. little moving dots in the background) can slow down standard Dreamer training. We find that SD$_{\mathrm{approx.}}$ reaches the final performance of SD$_{\mathrm{GT}}$ in most tasks. Certainly, it takes longer to converge due to noisy targets, but in a few tasks such as Walker Run, the curve seems very similar to its GT counterpart. One failure case is Reacher Easy, where the goal is to move a two-jointed robot arm's end effector close to a target. SD$_{\mathrm{approx.}}$ struggles in this task because the task-relevant objects (a robot arm and a target) are small, making it challenging to predict correct segmentation.

**Comparison with Baselines.** We compare SD$_{\mathrm{approx.}}$ with the state-of-the-art methods, which include: 1) **DreamerPro** (Deng et al., 2022) which uses prototypical representation learning (Caron et al., 2020) in the Dreamer framework; 2) **RePo** (Zhu et al., 2023) which minimizes mutual information between the observation and the representation, while maximizing it between the representation and all future rewards, to only keep predictable information; and 3) **TIA** (Fu et al., 2021) which learns separate representations for task-relevant and task-irrelevant parts which are then combined to reconstruct the original, distracting image. There are other reconstruction-free



(a) Dreamer vs. SD

(b) Our method vs. Baselines

Figure 2: Evaluation return during training on DMC-1M. X-axis is the number of environment steps and y-axis is evaluation return. All curves show the mean over 4 seeds with the standard deviation shaded. Best viewed in color.

model-based RL methods, such as TD-MPC (Hansen et al., 2022; 2023), but it is shown that they have difficulty training in the presence of distractions (Zhu et al., 2023).

The results in Fig. 2b suggest that our final performance is always better or on par with prior methods. Our method also achieves higher sample efficiency in all tasks, except comparable in Reacher Easy. TIA appears to underperform the most in many tasks. Since it has to infer what the task-relevant parts are during training, it not only requires much data but also is very sensitive to hyper-parameters used to balance the task-relevant and -irrelevant branches. Even with the best hyper-parameters, it sometimes ends up in a degenerate solution where a single branch takes all information. In contrast, our method can effectively focus only on task-relevant parts without any additional hyper-parameter tuning, empowered by the off-the-shelf segmentation model and prior knowledge. RePo shows comparable performance to our method in Cartpole Swingup but underperforms significantly in other tasks and converges very slowly. Again, it requires many trajectories to infer which perceptual features are predictable. Also, backgrounds can sometimes be predictable yet distracting, in which case RePo would count them as task-relevant. Among these methods, DreamerPro performs most competitively, which demonstrates the effectiveness of the prototypical representation learning in learning useful features for control. However, it still needs more environment interactions for training in many cases and converges to lower performance. Most importantly, none of the baselines are able to train an agent in a sparse reward since it becomes extremely challenging to infer task-relevance when the signal hinting at task-relevance is very weak. Nevertheless, our method achieves compelling performance, being the first method that is able to train an agent with sparse rewards under distraction.

An interesting perspective on our method is that behind its strength is the power of segmentation foundation models. As the foundation model had been trained on web-scale data, its fine-tuned version with a one-shot data can generalize well, e.g. to different poses. Our method effectively addresses the difficulty of training agents with distraction by offloading the task of identifying task-relevant regions to the out-of-the-box segmentation model, achieving high sample efficiency and generalization ability. In contrast, previous work has faced difficulty in training with highly noise-susceptible RL algorithms and learning robust representations at the same time.

**How does the selective $L_2$ loss help overcome noisy auxiliary targets originating from segmentation prediction errors?** Tab. 1 shows that $\text{SD}_{\text{approx.}}$ consistently outperforms $\text{SD}_{L_2}$ across all tasks and tends to have lower variance overall. This trend is particularly discernible in complex locomotion tasks such as Cheetah Run and Walker Run, where cooperation of the joints is crucial in achieving high rewards.

Table 1: Final performance of SD with selective and naive $L_2$ loss. Mean scores over 4 seeds with standard deviations are presented.

| Task | $\text{SD}_{\text{approx.}}$ | $\text{SD}_{L_2}$ |
|---|---|---|
| Cartpole Swingup | **730 ± 129** | 719 ± 108 |
| Cartpole Swingup Sparse | **521 ± 160** | 408 ± 198 |
| Cheetah Run | **619 ± 61** | 486 ± 101 |
| Hopper Stand | **846 ± 47** | 790 ± 88 |
| Reacher Easy | **597 ± 168** | 415 ± 87 |
| Walker Run | **730 ± 22** | 557 ± 89 |

## 4 Conclusion

In this paper, we propose SD, a simple yet effective way to learn task-relevant features in MBRL framework using segmentation masks. A variant trained with the ground-truth masks achieves near-oracle performance with a great sample efficiency on distracting environments given good prior knowledge. The main method, trained with the estimates leveraging the off-the-shelf segmentation model with a single pair of training data and using a modified $L_2$ loss, also reaches a decent performance and outperforms baselines. It is particularly notable that our approach is the first method that is able to train an agent in a sparse reward environment under distraction, enabling agent training robust to distractions without extensive reward engineering. This work also furthers the combining of computer vision and RL approaches by presenting a novel way of leveraging the recent advances in segmentation for addressing difficulties in visual control tasks. The proposed method also provides interface with human to indicate task relevance effectively. This enables practitioners to readily train an agent for their own purpose without extensive reward engineering.

### Acknowledgments

This work was funded in part by the National Science Foundation (Award #2321786) and by the HPI Research Center in Machine Learning and Data Science at UC Irvine.

## References

Homanga Bharadhwaj, Mohammad Babaeizadeh, Dumitru Erhan, and Sergey Levine. Information prioritization through empowerment in visual model-based rl. In *International Conference on Learning Representations*, 2022.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi (eds.), *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://aclanthology.org/W14-4012.

Fei Deng, Ingook Jang, and Sungjin Ahn. Dreamerpro: Reconstruction-free model-based reinforcement learning with prototypical representations. In *International Conference on Machine Learning*, pp. 4956–4975. PMLR, 2022.

Xiang Fu, Ge Yang, Pulkit Agrawal, and Tommi Jaakkola. Learning task informed abstractions. In *International Conference on Machine Learning*, pp. 3480–3491. PMLR, 2021.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Tung D Nguyen, Rui Shu, Tuan Pham, Hung Bui, and Stefano Ermon. Temporal predictive coding for model-based planning in latent space. In *International Conference on Machine Learning*, pp. 8130–8139. PMLR, 2021.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023.

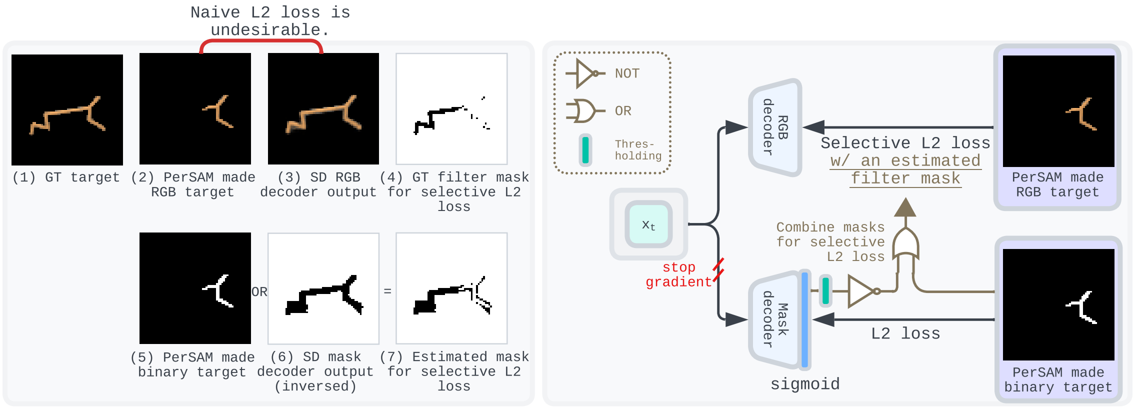
Chuning Zhu, Max Simchowitz, Siri Gadipudi, and Abhishek Gupta. Repo: Resilient model-based reinforcement learning by regularizing posterior predictability. In *Advances in Neural Information Processing Systems*, 2023.

# Appendix

## A Details on Selective $L_2$ Loss

**Selective $L_2$ Loss.** Even when targets are noisy, neural networks can overcome label errors and predict correctly if a large majority of the data is labeled correctly. In addition, since DREAMER's latent dynamics employs GRUs (Cho et al., 2014) as part of its neural architecture, its outcome would inherently tend to be temporally consistent even when the targets are flickering. When this happens, as illustrated in Fig. 3a (2)&(3), it is not desirable to flow gradients from those regions where the pseudo-label is wrong but the SD prediction is correct. Fig. 3a (4) shows a ground-truth filter mask that reveals, as zeroed-out pixels, where PerSAM is wrong and the SD is correct. In these pixels, we should not compute the $L_2$ loss. In many practical settings, of course, the ground-truth filter mask is unavailable even in training. Thus we next describe a selective $L_2$ loss using an estimated mask.

**Selective $L_2$ Loss with Estimated Filter Mask.** We present a heuristic method for estimating a filtering mask for selective $L_2$ loss. Preliminary experiments suggested that binary mask prediction with a sigmoid layer on top, as an auxiliary task, recovers very well from false negative labels. Based on this observation, we devise a world model with two reconstruction tasks (Fig. 3b) — one for RGB with spotlights on task-relevant parts as described in Sec. 2 and the other for binary segmentation mask with stop gradient — and use the binary mask prediction branch for selective $L_2$ loss filtering mask estimates.


Naive L2 loss is undesirable.
(1) GT target
(2) PerSAM made RGB target
(3) SD RGB decoder output
(4) GT filter mask for selective L2 loss
OR
=
(5) PerSAM made binary target
(6) SD mask decoder output (inversed)
(7) Estimated mask for selective L2 loss
NOT
OR
Thres-holding
RGB decoder
Selective L2 loss w/ an estimated filter mask
PerSAM made RGB target
$x_t$
Combine masks for selective L2 loss
stop gradient
Mask decoder
L2 loss
sigmoid
PerSAM made binary target

(a) GT filter mask and its estimate for **selective L2 loss**

(b) Two separate heads for RGB and mask

Figure 3: (a) Components of the selective $L_2$ loss. (b) A world model equipped with two decoders, one for reconstructing task-relevant RGB and the other for binary mask, the targets for which are generated by a segmentation model. Note that latent representations in the world model are subjected to a training signal only from the RGB branch, and the binary branch is only utilized to estimate the filtering masks.

The first component of the estimated filter mask is that we always include in the $L_2$ loss computation pixels within the SAM prediction. This comes with the trade-off of including false positives occasionally, but we designed this rule to provide a strong signal to true positives. The second component is that we compute the loss for pixels which SD predicts are task-*ir*relevant. Formally, we compute the $L_2$ loss on the mask

$$\text{mask}_{estimate} = \text{mask}_{\text{SAM}} \vee \neg\text{mask}_{\text{SD}}, \tag{1}$$

where $\text{mask}_{\text{SD}}$ is obtained by binning the SD binary mask prediction using a threshold of 0.9. Fig. 3a (5-7) describes the components to estimate filter mask. Intuitively, pixels are ruled out from the $L_2$ computation when SD is confident that they are task-relevant but they are excluded from the PerSAM prediction. The estimates are, of course, not the same as the true filtering mask, as in Fig. 3a (4) vs. (7). However, our experiments suggest that this selective loss is effective in overcoming noisy labels from segmentation prediction.