PERSEG: LEARNING DISCRIMINATIVE PROMPTS FOR MASK-FREE PERSONALIZED SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt spaces in foundation models support semantic generalization, but can be insufficient for personalization, where a particular subject must be distinguished from visually similar instances of the same category. In this work, we study *mask-free one-shot personalized segmentation*: given a single RGB reference image of an unseen subject, without a reference mask or per-identity optimization, the goal is to segment that subject in a query image or video. We introduce **PerSeg**, a framework that predicts personalized prompts from visual evidence while explicitly addressing fine-grained negative distractors. First, **PerSeg-DPT** learns a personalized prompt teacher using positive supervision together with hard negatives, including same-class distractors. Second, **PerSegDelta** amortizes this teacher by predicting an identity-specific offset from a semantic base prompt directly from the RGB reference image, enabling feed-forward personalization of unseen subjects. Third, **PerSeg-Adapt** predicts similarity-gated prompt updates to accommodate target appearance changes over time in video. To systematically evaluate identity discrimination, we introduce the **PODS/DreamBooth composite** benchmark, a controlled 2×2 protocol in which the target remains fixed while distractors progress from different-category instances to visually similar and near-identity subjects. On near-identity distractors, negative-aware PerSeg-DPT improves mIoU from 39.8% to 75.8% over standard prompt tuning, while mask-free PerSegDelta achieves 70.4%, approaching the teacher ceiling. PerSegDelta further obtains gains of $15.3/1.9$ points over the previous best method on the PODS and DreamBooth composite datasets. On ILIAS personalized retrieval, its prompt embeddings achieve a best mAP of 64.9, outperforming class-only and DINOv2 baselines. In video, PerSeg-Adapt improves performance by 2.4 points over PerSegDelta on held-out YouTube-VOS validation. These results show that personalized prompts for unseen subjects can be predicted from a single RGB reference, and that effective personalization depends on explicit supervision against hard negatives.

1 INTRODUCTION

Over the last decade, foundation models (Radford et al., 2021; Carion et al., 2025; Oquab et al., 2024; Rombach et al., 2022) have demonstrated remarkable zero-shot performance for tasks such as recognition, segmentation, retrieval, and image or video generation. Given a semantic description such as *a dog*, *a person*, or *a mug*, modern models can recognize, retrieve, generate, or segment content corresponding to the requested concept, often without task-specific training. A central reason for this flexibility is that these models organize visual and language concepts in structured representation spaces: related concepts occupy related regions, and prompts provide a mechanism for navigating this space. Prompts are also an efficient means of adaptation: optimizing a few input tokens or prompt embeddings on a small set of examples can improve task performance while the foundation model remains frozen.

The structure of prompt spaces is particularly appealing for personalization. Rather than learning an entirely new model for every subject, a personalized concept can be modeled as a refinement of a semantic concept the model already represents. For example, the representation of *my dog* should remain related to the broader concept *a dog*, while incorporating a smaller set of visual attributes that distinguish one individual from other dogs. If identity variation within a category is sufficiently

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.

structured, a model could learn to predict such a refinement directly from visual evidence, mapping a single RGB reference image of an unseen subject to an identity-specific prompt. This would enable *mask-free one-shot personalization*, in which the personalized representation is inferred in a single forward pass rather than optimized separately for each identity.

However, the same structure that supports semantic generalization also exposes a limitation for personalization. Foundation-model prompt spaces are primarily shaped to organize broad semantic concepts, where grouping visually diverse examples of the same category is desirable. Personalization instead requires separating nearby instances within the same semantic neighborhood: not merely *a dog*, but *my dog*; not merely *a mug*, but *my mug*. While examples of a target provide strong evidence about what the desired instance *is*, they provide only weak evidence about the fine-grained boundaries that distinguish it from everything it *is not*. In particular, the most informative negatives are often same-class distractors, which may be visually similar, near-identical, or absent from the images used during personalization. Thus, a prompt that accurately captures the semantic category of a target can still fail to encode the identity-specific distinctions needed to reject nearby distractors.

Personalized segmentation provides an especially demanding setting for studying this problem. Unlike conventional semantic or instance segmentation, the goal is to segment one specified subject while rejecting all other objects, including visually similar instances from the same category that may appear in the same scene. Existing personalized segmentation methods typically rely on a reference image together with a mask, that explicitly localizes the target. We instead study *mask-free one-shot personalized segmentation*: given a single RGB reference image of an unseen subject, without a reference mask or per-identity optimization, the goal is to segment that subject in a query image or video. This setting makes the tension between semantic generalization and identity discrimination explicit: the representation must generalize to new subjects through the structure of the prompt space, yet remain sufficiently discriminative to avoid collapsing onto their semantic category.

Video further amplifies this challenge. A personalized representation must preserve the target identity while visual appearance changes due to pose, viewpoint, illumination, scale, and occlusion. A static prompt can therefore fail in two opposing ways: it may be too broad and respond to same-class distractors, or too specific and fail when the target appearance changes over time. Effective personalization consequently requires representations that are both *discriminative*, separating the target from fine-grained and potentially unseen negatives, and *adaptive*, incorporating new visual evidence without losing target identity.

In this work, we study personalized prompt representations using segmentation as our primary experimental setting and retrieval as a complementary test of fine-grained identity discrimination. Our central hypothesis is that semantic prompt spaces provide a transferable base for zero-shot generalization to unseen subjects, but semantic prompts alone are insufficient for fine-grained personalization. We therefore represent a personalized prompt as the composition of a semantic base prompt and an identity-dependent offset. The semantic component anchors the representation in the structured concept space of the foundation model, while the identity offset captures visual distinctions needed to separate the target from nearby instances. This decomposition makes it possible to learn discriminative personalized prompts from supervision and, crucially, to predict such prompts directly from a reference image for previously unseen subjects.

We first investigate the role of negative information in personalized prompt learning. Using prompt tuning (learnable prompt embeddings) with SAM3, we find that positive-only personalization can perform well under conventional evaluation while remaining vulnerable to same-category distractors. We therefore introduce explicit hard-negative supervision, including same-class distractors. These negatives shape the local decision boundary around a personalized prompt, encouraging it not only to match the target but also to reject visually plausible alternatives. Negative supervision produces substantially larger benefits when evaluation requires fine-grained identity discrimination, supporting the hypothesis that conventional evaluation with predominantly easy negatives can conceal important limitations of personalized prompting.

Explicit negative supervision alone, however, cannot provide examples of every possible distractor that may be encountered at test time. We therefore investigate whether the identity-specific component of a personalized prompt can be inferred directly from visual evidence. Building on the structured geometry of the learned prompt space, we learn image-to-prompt predictors that map a target reference image to an offset from a semantic base prompt. This feed-forward formulation

amortizes personalized prompt learning across many training identities and enables zero-shot generalization: at test time, the model constructs an identity-aware prompt for a new subject from RGB alone, without a mask and without per-subject optimization.

Finally, we consider personalization in video, where the visual evidence needed to discriminate an identity changes over time. Rather than treating personalization as a one-time operation, we introduce dynamic prompting, in which the personalized prompt is updated online using current visual evidence. A lightweight policy conditions on static identity information and current-frame features to predict residual prompt updates, while a similarity gate limits updates under ambiguous observations. The resulting representation balances two competing requirements: adapting to changes in target appearance while remaining anchored to the original identity, thereby reducing drift toward same-class distractors.

Our experiments study these questions across PODS (Sundaram et al., 2025), DreamBooth (Ruiz et al., 2023), DogFace (Mougeot et al., 2019), CUB (Wah et al., 2011) and YouTube-VOS (Xu et al., 2018) subjects for segmentation, and ILIAS (Kordopatis-Zilos et al., 2025) for personalized retrieval. Beyond aggregate performance, we investigate how personalized prompts behave as the distinction between targets and negatives becomes increasingly fine-grained. The results show that the structured prompt space of a foundation model can support feed-forward, mask-free personalization of unseen subjects, but that this capability depends on explicitly learning identity-discriminative offsets rather than relying on semantic prompting alone. Our main contributions are:

- We propose PerSeg, a personalized prompting framework that represents an identity as a semantic base prompt plus an identity-specific offset. Its first stage, PerSeg-DPT, learns discriminative prompts for a frozen SAM3 using hard negatives in both image space (same-class distractor images) and prompt space (the prompts of other same-class instances), together with an identity classifier.
- We introduce PerSegDelta, which amortizes PerSeg-DPT by predicting the identity offset within a low-dimensional subspace of teacher offsets from a single RGB reference, enabling personalization of unseen subjects without a reference mask.
- We extend the framework to video with PerSeg-Adapt, which updates the predicted prompt online through similarity-gated residual updates that follow appearance changes while remaining anchored to the target identity.
- We introduce a controlled negative-difficulty benchmark on PODS and DreamBooth, in which the target is fixed while distractors progress from different-category to near-identical same-category instances, and evaluate PerSeg on this benchmark, on standard personalized segmentation, on ILIAS personalized retrieval, and on video segmentation.

2 Related Work

Foundation models for segmentation. The Segment Anything Model (SAM) established a general-purpose framework for promptable segmentation across diverse images (Kirillov et al., 2023). Subsequent models extend promptable segmentation to video and improve temporal propagation and object tracking, including SAM 2 (Ravi et al., 2025) and SAM 3 (Carion et al., 2025). These models provide strong generic segmentation capabilities, but their standard prompts are not designed to distinguish a particular subject from same-class distractors. Our contribution is orthogonal to the design of the segmentation backbone: we focus on learning and predicting an identity-specific prompt that makes the target distinguishable under a mask-free reference protocol.

Personalized segmentation. Personalized segmentation aims to segment a user-provided subject in new images or videos. Existing methods localize the target by matching a reference exemplar *image-mask pair* to a query image and subsequently use a segmentation model such as SAM to generate its mask. PerSAM (Zhang et al., 2024b) uses SAM features for reference-query correspondence, while subsequent methods Matcher (Liu et al., 2024), GF-SAM (Zhang et al., 2024a), PDM (Samuel et al., 2024) leverage DINOv2 or diffusion features to improve semantic and instance-level matching. FoundYou (Trivigno et al., 2026) instead adapts SAM 2 to unified personalized segmentation and retrieval, constructing an instance memory from a reference image and spatial prompt. Personalized OVSS (Park et al., 2025) learns a personal concept through per-concept prompt tuning from image-mask pairs together with a negative-mask proposal to suppress false positives. Beyond

segmentation, personalized representation learning has used synthetic images generated from a few target examples to contrastively fine-tune instance-specific visual features for recognition, retrieval, detection, and segmentation (Sundaram et al., 2025). In contrast, we study *mask-free one-shot personalized segmentation*: given a single RGB reference image of an unseen subject, PerSeg receives no reference mask and performs no per-identity optimization at test time. Our composite distractor benchmark further measures whether personalization remains reliable as negatives become increasingly similar to the target.

Prompt learning. Prompt learning or tuning (PT) adapts frozen vision–language and segmentation foundation models by optimizing a small set of prompt parameters while retaining the pretrained backbone. CoOp and CoCoOp learn task-specific or image-conditioned prompt representations for recognition (Zhou et al., 2022b;a). Textual inversion and DreamBooth likewise demonstrate that prompts learned from a few images of a specific visual concept can encode its appearance for image generation (Gal et al., 2023; Ruiz et al., 2023). However, while generation-oriented personalization aims to reproduce a subject's visual identity, personalized segmentation requires discriminating the target from other, potentially near-identical, instances in the same image. Positive examples alone can suffice to learn category-level or appearance-level representations without sufficiently defining the fine-grained boundary between the target and same-class distractors. PerSeg aims to overcome this problem by learning prompts with explicit hard-negative supervision, predicting an identity-specific prompt offset from a single unmasked RGB reference without test-time optimization, and using gated updates to maintain identity under video appearance changes.

3 METHOD

PerSeg seeks to refine a semantic prompt for category c into a personalized prompt that identifies a specific target instance. To emphasize this, the prompt is decomposed into

$$\mathbf{p}^{(s)} = \mathbf{p}_{\text{base}}(c) + \Delta^{(s)}, \tag{1}$$

where $\mathbf{p}_{\text{base}}(c)$ is a semantic prompt for category c and $\Delta^{(s)}$ an identity-specific offset. In the static setting, the offset is inferred from a single RGB reference image. In the video setting, it is further updated online as the target appearance changes.

The proposed method has three stages. First, **PerSeg-DPT** learns a discriminative prompt for each training instance by prompt tuning against a frozen SAM3, using hard negatives in both image and prompt space and an identity classifier. This prompt serves as a learned teacher. Second, **PerSegDelta** amortizes the teacher by learning an image-to-prompt predictor that receives a single RGB reference image without a reference mask. Third, **PerSeg-Adapt** updates the predicted prompt online in video using gated residual updates based on current-frame visual evidence. Figure 2 illustrates the framework and how they share the frozen SAM3 backbone and the semantic base prompt $\mathbf{p}_{\text{base}}(c)$.

3.1 NEGATIVE-AWARE DISCRIMINATIVE PROMPT TUNING

The first stage of PerSeg learns a personalization prompt for each training instance, e.g., a specific mug, that conditions the frozen SAM3 to segment that instance and nothing else. This is achieved by 1) tuning a discriminative prompt for the instance, using a few images of the mug as positive examples and a small set of same-class distractors (other mugs) as negative examples, and 2) augmenting SAM3 with an identity classifier that determines whether the instance specified by a prompt is present in an image. The prompts of all instances and the classifier are learned jointly, and the classifier's supervision also shapes the prompts.

Architecture. SAM3 takes as input an image I and a semantic prompt, such as "photo of a dog," specifying the concept c ("dog") to be segmented. We denote the portion of this prompt encoding the concept by $\mathbf{p}_{\text{base}}(c)$. A set of pre-learned DETR-style object queries attends to the image features, fused with the encoded prompt, to identify candidate objects. Each object query q produces a segmentation mask $\hat{M}_q$ and a scalar logit $\hat{s}_q$ indicating whether the candidate object matches the prompt. SAM3 also produces a global target-presence logit $\hat{p}$ indicating whether the concept specified by the prompt is present in I. In the losses below, $\hat{M}$ and $\hat{s}$ denote the mask and matching logit of the highest-scoring query.

While SAM3 is designed to match semantic concepts, personalization requires selectivity between individual instances of the same concept. PerSeg introduces two mechanisms for this: a learnable, instance-specific *personalization prompt* $\mathbf{p}^{(s)}$, which conditions SAM3 on the target instance s, and an *identity classifier*, learned across the training instances, which enhances model selectivity for fine-grained instance discrimination. The classifier predicts whether the identity specified by the prompt is present in the image.

The concept prompt $\mathbf{p}_{\text{base}}(c) \in \mathbb{R}^{K\times D}$ consists of the embeddings of the last $K = 4$ tokens, of dimension $D = 1024$, extracted from the phrase "photo of a c", so that every class maps to exactly K tokens. For each instance s of class c, these tokens are replaced by a learnable personalization prompt $\mathbf{p}^{(s)} \in \mathbb{R}^{K\times D}$, initialized as $\mathbf{p}^{(s)} = \mathbf{p}_{\text{base}}(c)$ and fed to the input of SAM3's text encoder. This ensures that the personalization prompt remains subject to the encoder's language prior, which regularizes the learned tokens.

The identity classifier operates on the object-query and prompt representations from SAM3. Let $\{\mathbf{h}_q\}_{q=1}^{Q}$ denote the object-query representations from the final SAM3 decoder layer and $\{\mathbf{e}_k\}_{k=1}^{K}$ the encoded prompt tokens. These are first mean-pooled, $\bar{\mathbf{h}} = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{h}_q, \bar{\mathbf{e}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_k$, and then combined into a joint image-prompt feature vector $\mathbf{f} = \left[\bar{\mathbf{h}}; \bar{\mathbf{e}}; \bar{\mathbf{h}} \odot \bar{\mathbf{e}}; |\bar{\mathbf{h}} - \bar{\mathbf{e}}|\right]$, where $\odot$ is the Hadamard (element-wise) product. A linear classifier then estimates the probability that the identity specified by the personalization prompt $\mathbf{p}^{(s)}$ is present in I with $\hat{y} = \sigma(\mathbf{w}^T\mathbf{f} + b)$, where σ is the sigmoid function and $(\mathbf{w}, b)$ are learned parameters.

Training Data. To encourage fine-grained identity discrimination, PerSeg learns from positive and negative examples. For each target instance s, e.g., a specific dog, the positive examples consist of images $\{I_i^{(s)}\}_{i=1}^{N_+}$ of the instance and their ground-truth segmentation masks $\{M_i^{(s)}\}_{i=1}^{N_+}$, with identity-presence label $y = 1$,

$$x_{+,i}^{(s)} = \left(I_i^{(s)}, M_i^{(s)}, 1\right), \qquad i = 1, \ldots, N_+. \tag{2}$$

The negative examples are same-class distractors, i.e., images $\{D_j^{(s)}\}_{j=1}^{N_-}$ of other instances of the same semantic class (other dogs). Since the target instance s is absent from these images, their ground-truth target masks are empty and their identity-presence labels are $y = 0$,

$$x_{-,j}^{(s)} = \left(D_j^{(s)}, \mathbf{0}, 0\right), \qquad j = 1, \ldots, N_-. \tag{3}$$

The training set for instance s is therefore $\mathcal{D}^{(s)} = \left\{x_{+,i}^{(s)}\right\}_{i=1}^{N_+} \cup \left\{x_{-,j}^{(s)}\right\}_{j=1}^{N_-}$, leading to the dataset $\mathcal{D} = \bigcup_{s=1}^{N}\mathcal{D}^{(s)}$.

Prompt Distractors. The same-class image distractors in $\mathcal{D}$ provide negative supervision in image space: the image varies while the target prompt stays fixed. A complementary form of negative supervision would operate in prompt space: the image stays fixed while its prompt is replaced by that of another instance of the same class. Unlike image distractors, however, such prompts are not available in advance, since the personalization prompts are themselves being learned. PerSeg therefore obtains prompt distractors from the prompts being optimized. Let $\mathcal{S}(c)$ denote the set of training instances of class c, and c_s the class of instance s. For a positive image $I_i^{(s)}$, an instance r is sampled uniformly from $\mathcal{S}(c_s) \setminus \{s\}$, and $I_i^{(s)}$ is paired with the current value of its prompt $\mathbf{p}^{(r)}$. Since instance r is absent from $I_i^{(s)}$, the prompt-distractor pair $(I_i^{(s)}, \mathbf{p}^{(r)})$ has an empty target mask and identity-presence label $y = 0$, exactly like an image distractor.

Discriminative Prompt Tuning. PerSeg jointly learns the personalization prompts and the identity classifier from the three types of image-prompt pairs: positive pairs $(I_i^{(s)}, \mathbf{p}^{(s)})$, image-distractor pairs $(D_j^{(s)}, \mathbf{p}^{(s)})$, and prompt-distractor pairs $(I_i^{(s)}, \mathbf{p}^{(r)})$. For a pair $(I, \mathbf{p})$, let $\hat{M}$, $\hat{s}$, $\hat{p}$, and $\hat{y}$ denote the predicted mask, SAM3 matching logit, SAM3 target-presence logit, and identity-classifier output. Segmentation is supervised by

$$\mathcal{L}_{\text{seg}}(\hat{M}, M) = \lambda_{\text{bce}}\mathcal{L}_{\text{BCE}}(\hat{M}, M) + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}(\hat{M}, M), \tag{4}$$

where M is the ground-truth mask. Responses to negative pairs are further suppressed by the distractor response suppression (DRS) loss

$$\mathcal{L}_{\text{DRS}}(\hat{s}, \hat{p}) = \lambda_{\text{score}}\mathcal{L}_{\text{BCE}}(\hat{s}, 0) + \lambda_{\text{pres}}\mathcal{L}_{\text{BCE}}(\hat{p}, 0), \tag{5}$$

and the identity classifier is supervised using a binary cross-entropy loss $\mathcal{L}_{\mathrm{id}}(\hat{y}, y) = \mathcal{L}_{\mathrm{BCE}}(\hat{y}, y)$. All three pair types are supervised by the same pair loss

$$\ell(I, \mathbf{p}, M, y) = \mathcal{L}_{\mathrm{seg}}(\hat{M}, M) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}}(\hat{y}, y) + (1 - y)\,\lambda_{\mathrm{DRS}}\mathcal{L}_{\mathrm{DRS}}(\hat{s}, \hat{p}), \tag{6}$$

where all predictions are computed from $(I, \mathbf{p})$. Positive pairs use $(M, y) = (M_i^{(s)}, 1)$, while image- and prompt-distractor pairs use $(M, y) = (\mathbf{0}, 0)$, so that the DRS term is active only for negatives. The two kinds of negatives are thus supervised identically and differ only in which element of the pair is replaced: the image for image distractors, the prompt for prompt distractors.

Joint optimization. Let $\mathcal{D}_+^{(s)}$ and $\mathcal{D}_-^{(s)}$ denote the positive and image-distractor examples of instance s. The prompts $\{\mathbf{p}^{(s)}\}_{s=1}^N$ and the classifier $(\mathbf{w}, b)$ are learned jointly, with all SAM3 parameters frozen, by minimizing

$$\begin{aligned}\mathcal{L}_{\mathrm{DPT}} = \mathbb{E}_s\Big[\, \tfrac{1}{2}\,\mathbb{E}_{(I,M,1)\sim\mathcal{D}_+^{(s)}}\Big(\ell(I, \mathbf{p}^{(s)}, M, 1) + \mathbb{E}_{r\sim\mathcal{S}(c_s)\setminus\{s\}}\,\ell(I, \mathbf{p}^{(r)}, \mathbf{0}, 0)\Big) \\ + \tfrac{1}{2}\,\mathbb{E}_{(D,\mathbf{0},0)\sim\mathcal{D}_-^{(s)}}\,\ell(D, \mathbf{p}^{(s)}, \mathbf{0}, 0)\,\Big]\end{aligned} \tag{7}$$

This objective is minimized stochastically. At each iteration, an instance s is sampled uniformly and, with equal probability, either a positive example or an image distractor of s is drawn. For a positive example, an instance $r \neq s$ of the same class is also sampled, and the same image is paired with the current value of $\mathbf{p}^{(r)}$ to form a prompt distractor. A gradient step is then taken on $\mathbf{p}^{(s)}$, $(\mathbf{w}, b)$, and, when a prompt distractor is present, $\mathbf{p}^{(r)}$. We train for 4,000 iterations per instance on average, with $\lambda_{\mathrm{bce}} = \lambda_{\mathrm{dice}} = \lambda_{\mathrm{score}} = \lambda_{\mathrm{pres}} = \lambda_{\mathrm{DRS}} = 1$ and $\lambda_{\mathrm{id}} = 0.1$. We refer to the learned prompt $\mathbf{p}_{\mathrm{DPT}}^{(s)}$ as the **discriminative prompt-tuned (DPT)** prompt of instance s.

3.2 PerSegDelta: Image-to-Prompt Prediction

Per-identity prompt tuning is computationally expensive and requires optimization for every new identity. We therefore leverage the structured prompt space learned by PerSeg-DPT to train a feed-forward predictor to infer the identity-specific prompt offset directly from a single RGB reference image R_s.

Teacher offsets. For each training instance s of class c_s, the teacher offset is the difference between its DPT prompt and the semantic prompt,

$$\Delta^{*(s)} = \mathbf{p}_{\mathrm{DPT}}^{(s)} - \mathbf{p}_{\mathrm{base}}(c_s) \in \mathbb{R}^{K\times D}. \tag{8}$$

Let $\boldsymbol{\delta}^{*(s)} = \mathrm{vec}(\Delta^{*(s)}) \in \mathbb{R}^{KD}$ denote its vectorization ($KD = 4{,}096$).

Offset subspace. Directly regressing the KD-dimensional offset overfits the training identities and generalizes poorly to unseen subjects and so, prediction is constrained to the principal subspace of the teacher offsets. Let $\bar{\boldsymbol{\delta}} = \frac{1}{N}\sum_s \boldsymbol{\delta}^{*(s)}$ be the mean offset and $\mathbf{W} \in \mathbb{R}^{KD\times r}$ the matrix whose orthonormal columns are the top r principal directions of $\{\boldsymbol{\delta}^{*(s)} - \bar{\boldsymbol{\delta}}\}_{s=1}^N$. Each teacher offset is then represented by its coefficient vector

$$\mathbf{z}^{*(s)} = \mathbf{W}^T\left(\boldsymbol{\delta}^{*(s)} - \bar{\boldsymbol{\delta}}\right) \in \mathbb{R}^r, \qquad \boldsymbol{\delta}^{*(s)} \approx \bar{\boldsymbol{\delta}} + \mathbf{W}\mathbf{z}^{*(s)}. \tag{9}$$

The mean $\bar{\boldsymbol{\delta}}$ and basis $\mathbf{W}$ are computed once on the training instances and kept fixed.

Multiscale reference features. To learn the identity cues ranging from low-level appearance (color, texture, markings) to higher-level shape and part structure, features encoded at different depths of a vision transformer are used. The reference image is therefore described by the class tokens of a frozen DINOv2 encoder at multiple depths. Let $\mathbf{c}_\ell(R) \in \mathbb{R}^d$ denote the class token at the output of transformer block ℓ (indexed from zero) and $\mathrm{LN_f}$ the pretrained final layer normalization of DINOv2. The reference descriptor is $\mathbf{u}(R) = [\,\mathrm{LN_f}(\mathbf{c}_\ell(R))\,]_{\ell\in\mathcal{L}} \in \mathbb{R}^{|\mathcal{L}|d}$, $\mathcal{L} = \{2, 5, 8, 11\}$, which samples every third block of the 12-block ViT-B/14 encoder, up to and including the last.

Coefficient predictor. The predictor g_ϕ is a multilayer perceptron with a single hidden layer and GELU activation,

$$\hat{\mathbf{z}} = g_\phi\left(\mathbf{u}(R)\right) = \mathbf{A}_2\,\mathrm{GELU}(\mathbf{A}_1\mathbf{u}(R) + \mathbf{a}_1) + \mathbf{a}_2, \tag{10}$$

where $\phi = \{\mathbf{A}_1, \mathbf{a}_1, \mathbf{A}_2, \mathbf{a}_2\}$. The predicted offset and personalized prompt for the instance s are reconstructed as

$$\hat{\Delta}^{(s)} = \mathrm{vec}^{-1}\left(\bar{\boldsymbol{\delta}} + \mathbf{W}\hat{\mathbf{z}}^{(s)}\right), \qquad \hat{\mathbf{p}}^{(s)} = \mathbf{p}_{\mathrm{base}}(c) + \hat{\Delta}^{(s)}. \tag{11}$$

Training. Each training instance s uses all the positive images $\{I_i^{(s)}\}_{i=1}^{N_+}$ from the first stage as reference images, sharing the target $\mathbf{z}^{*(s)}$. The predictor is learned by minimizing

$$\mathcal{L}_{\mathrm{Delta}} = \frac{1}{N}\sum_{s=1}^{N}\frac{1}{N_+}\sum_{i=1}^{N_+}\left\| g_\phi\left(\mathbf{u}(I_i^{(s)})\right) - \mathbf{z}^{*(s)}\right\|_2^2, \tag{12}$$

with DINOv2, SAM3, $\bar{\boldsymbol{\delta}}$, and $\mathbf{W}$ frozen. Because the columns of $\mathbf{W}$ are orthonormal, $\|\hat{\mathbf{z}} - \mathbf{z}^{*(s)}\|_2 = \|\mathbf{W}\hat{\mathbf{z}} - \mathbf{W}\mathbf{z}^{*(s)}\|_2$, so this objective equals the Euclidean distance, in prompt space, between the predicted offset and the projection of the teacher offset onto the principal subspace.

Inference. Given a single RGB reference image R of an unseen instance of class c, PerSegDelta computes $\hat{\mathbf{p}} = \mathbf{p}_{\mathrm{base}}(c) + \mathrm{vec}^{-1}(\bar{\boldsymbol{\delta}} + \mathbf{W}g_\phi(\mathbf{u}(R)))$ in a single forward pass and uses it to prompt SAM3. No reference mask, spatial annotation, or per-identity optimization is required. The identity classifier score is used as a confidence score for the predicted masks. Since the DPT prompts are optimized with per-identity supervision, PerSeg-DPT provides a reference ceiling, for PerSegDelta.

3.3 PerSeg-Adapt for Video

In video, the target may undergo changes in pose, viewpoint, illumination, scale, and occlusion, so a prompt inferred from the initial reference alone may become insufficient over the sequence. PerSeg-Adapt therefore adds a learned residual to the static PerSegDelta coefficients, $\mathbf{z}_t = \mathbf{z}_0 + \delta\mathbf{z}_t$, which is recomputed every T frames and decoded into the prompt $\mathbf{p}_t$ through the same offset basis as in Eq. 11. The residual is predicted by a lightweight MLP policy h_ψ from three DINOv2 features: the first-frame reference, which anchors the target identity; a crop around the previous prediction, which captures the target's current appearance; and the current full frame, which provides scene context. To prevent drift after tracking failures or when a same-class distractor enters the crop, the crop descriptor is refreshed only if its cosine similarity to the reference exceeds a threshold τ_{sim}. With all other components frozen, h_ψ is trained with the segmentation loss together with inter- and intra-class contrastive terms that keep the adapted prompt within its semantic class and closer to the target's static prompt than to those of other same-class instances. Appendix B provides more details on the dynamic prompt prediction, similarity-gated visual evidence, the residual policy, and its training objective.

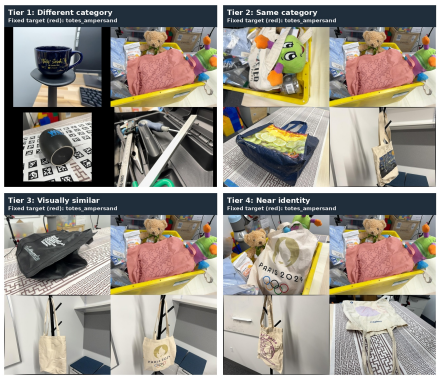
4 Experimental Setup and Results

Datasets. For personalized segmentation, we use PODS (Sundaram et al., 2025) and DreamBooth (Ruiz et al., 2023) standard and composite test images; for retrieval on ILIAS (Kordopatis-Zilos et al., 2025), and for video, DAVIS (Perazzi et al., 2016) and YouTube-VOS (Xu et al., 2018).

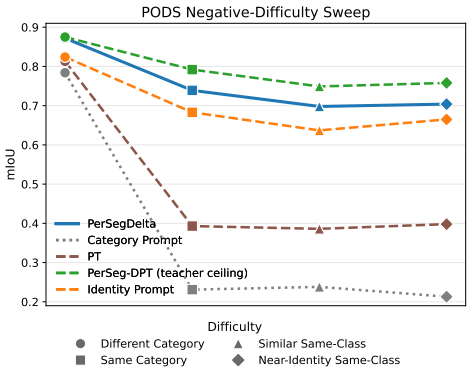
Controlled negative-difficulty composites. To isolate identity discrimination, we construct four matched 2×2 composite datasets for PODS and DreamBooth subjects. Each example contains one target crop and three distractor crops; evaluation uses only the target mask. The target is fixed across tiers while distractors are drawn from (i) a different category, (ii) the same category, (iii) the top 35% most visually similar same-category identities, or (iv) the three nearest same-category identities, using cosine similarity of 64×64 masked RGB crops. Crops are centered on the target box at $1.8\times$ scale. PODS has 1,050 test composites (70 disjoint identities) and Dreambooth has 128 test composites (30 identities); see Fig. 1(a).

Personalized retrieval. ILIAS contains 1,232 queries and 4,715 gallery images from 1,000 identities. We flatten each four-token SAM3 prompt into a 4,096-d vector and rank by cosine similarity. The query prompt is the SAM3 text embedding of a Qwen3-VL-predicted (Bai et al., 2025) class name plus the predicted PCA delta and we compare it with the class-only embedding and DINOv2.

Metrics and Baselines. We report mIoU, boundary F1 (BF1), precision and recall for image segmentation; mAP, MRR, R@K and median first-positive rank for retrieval; and Dice or J&F for



(a) Negative-difficulty composites.



(b) PerSegDelta sweep.

Figure 1: (a) Controlled PODS negative-difficulty composites for a fixed held-out tote bag target (red overlay). Each 2×2 composite contains the target plus three distractors. From left to right and top to bottom, distractors progress from different-category to same-category, visually similar, and nearest same-category identities. (b) Negative-difficulty sweep: absolute mIoU as negatives become increasingly similar to the target, with curves shown for Category, Identity, standard PT, PerSegDelta, and the PerSeg-DPT teacher ceiling.

Table 1: Composite personalized segmentation on PODS and DreamBooth 2 × 2 evaluations. "Mask" indicates whether the method receives a reference mask at test time: GF-SAM, FoundYou, PerSAM/PerSAM-F and PDM use a one-shot image–mask reference, while PerSegDelta is mask-free.

		PODS				DreamBooth			
Method	Mask	mIoU	BF1	Prec.	Rec.	mIoU	BF1	Prec.	Rec.
GF-SAM	Yes	32.6	28.7	33.7	**90.1**	79.6	44.8	80.7	**97.9**
FoundYou	Yes	59.0	45.0	62.7	65.3	81.6	51.3	82.6	90.0
PerSAM	Yes	29.8	22.8	32.4	32.0	41.4	35.4	43.5	47.0
PerSAM-F	Yes	28.0	22.2	29.6	29.8	42.7	37.4	42.9	45.1
PDM	Yes	25.4	20.8	26.5	47.0	29.5	19.6	29.7	65.3
PerSegDelta	**No**	**74.3**	**57.5**	**79.9**	75.7	**83.5**	**75.6**	**85.0**	84.4

video. Segmentation baselines are SAM3 category and Qwen3-VL identity-description prompts, positive-only PT, and mask-assisted GF-SAM, FoundYou, PerSAM, PerSAM-F, and PDM. Implementation details are provided in Appendix G.

4.1 Image Segmentation

Comparison with mask-assisted methods. Table 1 compares PerSegDelta, which receives only an RGB reference, with methods given a reference image–mask pair. PerSegDelta achieves the best mIoU and BF1 on both PODS and DreamBooth composites; GF-SAM attains the highest recall but low precision. A reference mask alone thus does not resolve identity discrimination: PDM, PerSAM, and PerSAM-F remain far below PerSegDelta despite mask-conditioned correspondence. Notably, DreamBooth was not used to train PerSegDelta or to estimate its offset subspace, so these results are cross-dataset and indicate that the predicted offsets transfer beyond the training distribution.

Negative-difficulty sweep. Fig. 1(b) reports the four PODS tiers results. Table 5 in Appendix shows the full results. *"Category"* prompts SAM3 with the class name (e.g., "mug"); *Identity* prompts SAM3 using a Qwen3-VL description (e.g., "blue mug with white patch"). PT refers to learned prompts without using negatives. As distractors become more similar, category prompting and positive-only PT collapse (PT: $81.3 \rightarrow 39.8$) and identity prompting degrades ($82.4 \rightarrow 66.5$). PerSeg-DPT remains strong on all same-category tiers, and PerSegDelta closely follows it, retaining 80.7% of its different-category performance ($87.2 \rightarrow 70.4$). On the near-identity tier, PerSegDelta exceeds identity prompting by 3.9 and PT by 30.6 mIoU, within 5.4 of the teacher, supporting the proposed decomposition. Qualitative results for PerSegDelta are shown in Appendix F.

Table 2: ILIAS personalized retrieval (cosine similarity; 1,232 queries, 4,715 gallery images). Class-only uses the Qwen-predicted class embedding without the PerSegDelta identity delta.

Method	mAP	MRR	R@1	R@5	R@10	Med.
DINOv2 ViT-B/14	27.8	56.5	48.5	65.3	71.3	2
Qwen3-VL only	55.0	55.1	46.3	63.9	76.2	2
FoundYou	59.5	**84.1**	**81.3**	**88.5**	90.8	**1**
Qwen3-VL + PerSegDelta	**64.9**	77.7	69.4	87.7	**94.0**	**1**

Table 3: Ablations on PODS composite 2×2 evaluation (%). Negative-image supervision and auxiliary identity classifier are toggled independently.

$\mathcal{L}_{\mathrm{DRS}}$	$\mathcal{L}_{\mathrm{id}}$	$\mathcal{PD}$	mIoU	BF1	Prec.	Rec.
×	×	×	39.26	34.11	40.94	40.53
✓	×	×	60.16	48.08	64.27	61.21
×	✓	✓	70.28	57.55	74.63	71.34
✓	✓	✓	**79.18**	**62.99**	**85.12**	**79.83**

Table 4: PerSeg-Adapt on held-out validation videos from Youtube-VOS, video-averaged Dice.

	SAM3 Box Tracking	PerSegDelta	PerSegDelta-updates	PerSeg-Adapt
Dice (video-averaged)	71.2	70.0	69.1	**72.4**

Ablation on Hard-negative supervision and auxiliary identity classifier We ablate the components of PerSeg-DPT training: image-distractor supervision ($\mathcal{L}_{\mathrm{DRS}}$ together with empty-mask segmentation on distractor images), the identity classifier ($\mathcal{L}_{\mathrm{id}}$), and prompt distractors ($\mathcal{PD}$). Since prompt distractors are supervised through the identity classifier, $\mathcal{PD}$ is enabled together with $\mathcal{L}_{\mathrm{id}}$. All variants use four learned prompt tokens, 4,000 updates, batch size one, and the same PODS composite 2×2 evaluation protocol. Table 3 reports the results. Both forms of negative supervision improve identity-discriminative segmentation over standard prompt tuning. Image distractors alone raise mIoU from 39.26% to 60.16%, and the identity classifier with prompt distractors alone raises it to 70.28%, even without any distractor images. Combining them yields 79.18% (+39.92), showing that negatives in image space and in prompt space are complementary. Additional ablation studies are reported in Appendix C.

4.2 Retrieval

Table 2 tests whether predicted prompts encode identity beyond the segmentation decoder using a retrieval setup despite not being trained for it. PerSegDelta prompts achieve the best mAP (64.9) and R@10 (94.0). Removing the identity delta drops mAP to 55.0 and R@1 from 69.4 to 46.3, showing that the delta carries identity information absent from the class prompt. DINOv2 features retrieve some near neighbors (48.5 R@1) but rank same-identity images far less consistently (27.8 mAP). PerSegDelta even outperforms FoundYou by **+5.4** mAP that is explicitly trained for retrieval while FoundYou is stronger at the top ranks (R@1, MRR) indicating that PerSegDelta ranks the full set of same-identity images more consistently than FoundYou. We further analyze the prompt-space geometry on CelebA and CUB datasets in the Appendix D.

4.3 Video Segmentation

We evaluate PerSeg-Adapt on held-out 11,292 YouTube-VOS frames, including crowded multi-instance scenes (bird flock, ice hockey, multi-basketball, multi-soccer, pedestrians, skiing). PerSeg-Adapt raises video-averaged Dice from 70.0 (PerSegDelta) or SAM3 Box tracking (71.2) to 72.4, and outperforms the gated PerSegDelta-updates baseline (69.1), which re-predicts the prompt from each accepted crop by 3.3 points, showing that the learned policy incorporates reliable evidence where naive re-prediction causes drift (Fig. 3). Additional results are reported in Appendix B.1.

5 Conclusion

We presented **PerSeg**, a framework for mask-free one-shot personalized segmentation. PerSeg-DPT learns an identity-discriminative prompt teacher with explicit hard negatives; PerSegDelta amortizes it by predicting an identity offset from a single unmasked RGB reference without test-time optimization; and PerSeg-Adapt extends it to video with gated updates that follow appearance changes while limiting drift. Results on controlled distractor composites, ILIAS retrieval, and video segmentation show that structured prompt spaces can support personalization of unseen subjects, provided they are trained against hard negative distractors and adapted in a controlled way.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that our results can be reproduced. All three stages of PerSeg are fully specified in Section 3: the discriminative prompt tuning objective and its sampling procedure (Eqs. 4–7), the offset subspace and coefficient predictor of PerSegDelta (Eqs. 8–12), and the dynamic prompt, similarity gate, residual policy, and training objective of PerSeg-Adapt (Appendix B, Eqs. 13–18). Appendix G lists all hyperparameters, including the number of prompt tokens ($K{=}4$) and their dimension ($D{=}1024$), loss weights, number of training iterations, batch sizes, the DINOv2 layers used for the reference descriptor, predictor architecture, number of PCA components ($r{=}256$), update interval ($T{=}5$), similarity threshold ($\tau_{\text{sim}}{=}0.6$), and learning rates. All experiments use publicly released frozen SAM3, DINOv2 ViT-B/14, and Qwen3-VL checkpoints. We will release the code upon acceptance of the paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021. doi: 10.1109/TPAMI.2019.2957464.

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*, pp. 4015–4026, 2023.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Šuma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiří Matas, Ondřej Chum, and Giorgos Tolias. ILIAS: Instance-level image retrieval at scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *International Conference on Learning Representations*, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Guillaume Mougeot, Dewei Li, and Shuai Jia. A deep learning approach for dog face verification and recognition. In Abhaya C. Nayak and Alok Sharma (eds.), *PRICAI 2019: Trends in Artificial Intelligence*, pp. 418–430, Cham, 2019. Springer International Publishing. ISBN 978-3-030-29894-4.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Sunghyun Park, Jungsoo Lee, Shubhankar Borse, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Understanding personal concept in open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19957–19966, October 2025.

F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Dvir Samuel, Rami Ben-Ari, Matan Levy, Nir Darshan, and Gal Chechik. Where's waldo: Diffusion features for personalized segmentation and retrieval. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 128160–128181. Curran Associates, Inc., 2024. doi: 10.52202/079017-4070. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e73f4935d2ef5e23997de852e8a52661-Paper-Conference.pdf.

Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. Personalized representation from personalized generation. In *International Conference on Learning Representations*, volume 2025, pp. 92572–92606, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e76eb4cf820bfd93abd9cf46b74aef35-Paper-Conference.pdf.

Gabriele Trivigno, Marcos Alfaro, Claudia Cuttano, Gabriele Berton, Luis Payá, and Carlo Masone. FoundYou: A unified model for personalized segmentation and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Computation & Neural Systems Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, CA, July 2011. URL https://resolver.caltech.edu/CaltechAUTHORS:20111026-120541847.

Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision*, pp. 585–601, 2018.

Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, 2024a.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Yu Qiao, Gao Peng, and Hongsheng Li. Personalize segment anything model with one shot. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18250–18279, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4f44e5c82dc7b0ba4472083e5dedb08e-Paper-Conference.pdf.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022b.

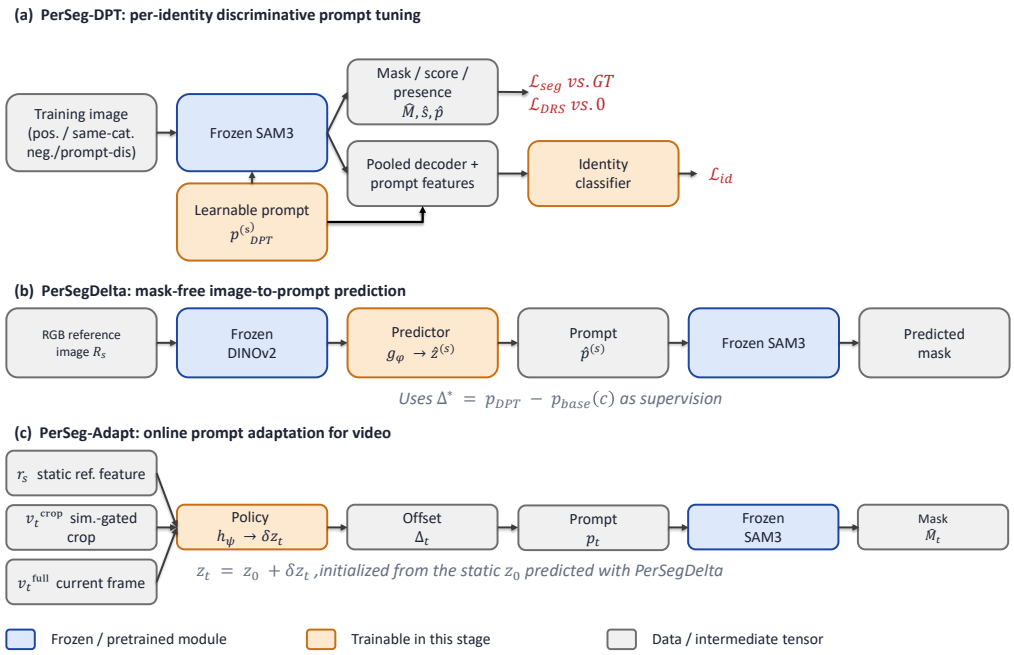


Figure 2: Overview of the PerSeg framework. **(a)** PerSeg-DPT tunes a personalized prompt against a frozen SAM3 using positive supervision and same-category hard negatives. **(b)** PerSegDelta distills the PerSeg-DPT prompt offset $\Delta^*_{\mathrm{DPT}} = \mathbf{p}^{(s)}_{\mathrm{DPT}} - \mathbf{p}_{\mathrm{base}}(c)$ into a feed-forward predictor that maps a single mask-free RGB reference image to a low-dimensional PCA coefficient vector $\hat{\mathbf{z}}^{(s)}$, reconstructing $\hat{\Delta}^{(s)} = \bar{\Delta} + \mathbf{W}\hat{\mathbf{z}}^{(s)}$ and the personalized prompt $\hat{\mathbf{p}}^{(s)} = \mathbf{p}_{\mathrm{base}}(c) + \hat{\Delta}^{(s)}$. **(c)** PerSeg-Adapt updates the coefficients online in video, $\mathbf{z}_t = \mathbf{z}_0 + \delta\mathbf{z}_t$ with $\delta\mathbf{z}_t = h_\psi(\mathbf{r}_s, \mathbf{v}_t^{\mathrm{crop}}, \mathbf{v}_t^{\mathrm{full}})$, using a lightweight policy over similarity-gated visual features. All three stages share the same frozen SAM3 backbone and semantic base prompt $\mathbf{p}_{\mathrm{base}}(c)$; blue-bordered modules are frozen and pretrained, orange-bordered modules are trainable in that stage, and gray modules denote data or intermediate tensors.

Table 5: Negative-difficulty sweep mIoU by method and distractor tier. PerSeg-DPT supplies the PerSegDelta ground-truth prompt targets and is the corresponding static teacher ceiling. All methods use SAM3 backbone

Tier	Category Prompt	Identity Prompt	PT	PerSeg-DPT	PerSegDelta
Different-category	78.4	82.4	81.3	**87.5**	87.2
Same-category	23.1	68.3	39.3	**79.2**	74.3
Visually similar same-category	23.8	63.7	38.6	**74.9**	69.8
Near-identity same-category	21.3	66.5	39.8	**75.8**	70.4

A Appendix

Figure 2 shows the overview of PerSeg framework and Table 5 shows the detailed results of the negative difficulty sweep experiment.

B PerSeg-Adapt for Video

Dynamic prompt. Let R_0 be the reference crop of the target in the first frame, provided with the sequence, and $\mathbf{z}_0 = g_\phi(\mathbf{u}(R_0))$ the static coefficients predicted by PerSegDelta, with the offset basis $(\bar{\boldsymbol{\delta}}, \mathbf{W})$ of PerSegDelta. At frame t, the coefficients are the static coefficients plus a learned residual,

$$\mathbf{z}_t = \mathbf{z}_0 + \delta\mathbf{z}_t, \qquad \mathbf{p}_t = \mathbf{p}_{\mathrm{base}}(c) + \mathrm{vec}^{-1}\left(\bar{\boldsymbol{\delta}} + \mathbf{W}\mathbf{z}_t\right), \tag{13}$$

and $\mathbf{p}_t$ prompts SAM3 on frame F_t. The residual is recomputed every T frames and held fixed in between, i.e., $\delta\mathbf{z}_t = \delta\mathbf{z}_{t_k}$ for $t_k \le t < t_k + T$, where $t_k = kT$.

Visual evidence and gating. At each update frame t_k, three descriptors are computed with the multiscale DINOv2 encoder $\mathbf{u}(\cdot)$ of PerSegDelta: the reference descriptor $\mathbf{r} = \mathbf{u}(R_0)$, a full-frame descriptor $\mathbf{v}_{t_k}^{\text{full}} = \mathbf{u}(F_{t_k})$, and a candidate crop descriptor $\tilde{\mathbf{v}}_{t_k}^{\text{crop}} = \mathbf{u}(C_{t_k})$, where C_{t_k} is the crop of F_{t_k} around the target predicted at the previous frame. A poorly localized crop, e.g., after a tracking failure or when a same-class distractor enters the box, would corrupt the update. The crop descriptor is therefore refreshed only if it remains similar to the reference,

$$\mathbf{v}_{t_k}^{\text{crop}} = \begin{cases} \tilde{\mathbf{v}}_{t_k}^{\text{crop}}, & \cos\left(\mathbf{r}, \tilde{\mathbf{v}}_{t_k}^{\text{crop}}\right) > \tau_{\text{sim}}, \\ \mathbf{v}_{t_{k-1}}^{\text{crop}}, & \text{otherwise}, \end{cases} \tag{14}$$

with $\mathbf{v}_{t_0}^{\text{crop}} = \mathbf{r}$ and $\tau_{\text{sim}} = 0.6$.

Residual policy. A residual policy h_ψ, implemented as a lightweight MLP, predicts the coefficient update from the concatenated descriptors,

$$\delta \mathbf{z}_{t_k} = h_\psi\big([\mathbf{r};\, \mathbf{v}_{t_k}^{\text{crop}};\, \mathbf{v}_{t_k}^{\text{full}}]\big). \tag{15}$$

The reference descriptor anchors the update to the target identity, the crop descriptor provides current target appearance, and the full-frame descriptor provides scene context, including the presence of potential distractors.

Training objective. The policy is trained with DINOv2, SAM3, g_ϕ, and $(\bar{\boldsymbol{\delta}}, \mathbf{W})$ are frozen, so that only ψ is learned directly with the mask supervision from groundtruth. For a training video of instance s of class c_s with ground-truth masks $\{M_t\}$, the segmentation term is $\mathcal{L}_{\text{seg}}\big(\hat{M}(F_t, \mathbf{p}_t), M_t\big)$, with $\mathcal{L}_{\text{seg}}$ as in PerSeg-DPT. Two contrastive terms regularize the dynamic prompt embedding $\mathbf{m}_t = \text{vec}(\mathbf{p}_t)$. Let $\mathbf{a}_c := \mathbf{p}_{\text{base}}(c)$ denote an anchor embedding of semantic class c and $\mathbf{m}_0^{(s')}$ the static PerSegDelta prompt embedding of instance s'. The inter-class term classifies $\mathbf{m}_t$ among the class anchors,

$$\mathcal{L}_{\text{inter}} = -\log \frac{\exp\big(\cos(\mathbf{m}_t, \mathbf{a}_{c_s})/\tau\big)}{\sum_c \exp\big(\cos(\mathbf{m}_t, \mathbf{a}_c)/\tau\big)}, \tag{16}$$

and the intra-class term classifies it among the static prompt embeddings of the instances $\mathcal{S}(c_s)$ of the same class,

$$\mathcal{L}_{\text{intra}} = -\log \frac{\exp\big(\cos(\mathbf{m}_t, \mathbf{m}_0^{(s)})/\tau\big)}{\sum_{s' \in \mathcal{S}(c_s)} \exp\big(\cos(\mathbf{m}_t, \mathbf{m}_0^{(s')})/\tau\big)}, \tag{17}$$

where τ is a temperature. The inter-class term keeps the adapted prompt within its semantic class, while the intra-class term keeps it closer to the target's own static prompt than to those of other same-class instances, preventing drift toward distractors. The policy is learned by minimizing

$$\mathcal{L}_{\text{Adapt}} = \mathbb{E}_t\Big[\mathcal{L}_{\text{seg}}\big(\hat{M}(F_t, \mathbf{p}_t), M_t\big) + \lambda_{\text{inter}}\mathcal{L}_{\text{inter}} + \lambda_{\text{intra}}\mathcal{L}_{\text{intra}}\Big], \tag{18}$$

where the expectation is over training videos and frames. Gradients of the segmentation term reach ψ through the frozen SAM3 and the linear map from $\mathbf{z}_t$ to $\mathbf{p}_t$.

Implementation. YouTube-VOS videos are split 80/20 into training and held-out validation sets. We use an update interval of $T = 5$ frames and $\lambda_{\text{inter}} = \lambda_{\text{intra}} = 0.2$, and train the policy for 5 epochs with learning rate 3×10^{-4}.

PerSegDelta-updates baseline. To isolate the effect of the learned policy, we also consider a baseline that, at each update frame whose candidate crop passes the same similarity gate, replaces the coefficients with a fresh PerSegDelta prediction from the crop, $\mathbf{z}_{t_k} = g_\phi(\tilde{\mathbf{v}}_{t_k}^{\text{crop}})$, without the reference anchor $\mathbf{z}_0$.

B.1 Results

On DAVIS-person (Table 6), all methods condition only on frame 0: PerSeg-DPT tunes a per-video prompt from three sampled frames, PerSegDelta and PerSeg-Adapt use only the frame-0 RGB region (PerSeg-Adapt re-predicting every five frames), and FoundYou receives the frame-0 mask-centroid point. PerSeg-DPT is strongest (69.30 F), while mask-free PerSegDelta and PerSeg-Adapt reach 58.87/60.51 F, far above FoundYou (25.84 F).



Figure 3: Qualitative examples from two videos. Left: ice hockey sweep video; Right: pedestrian crossing video where the static prompt drifts off the reference target from both the videos, while PerSeg-Adapt's dynamic policy re-anchors to it on the same target in both videos.

Table 6: Comparison on the DAVIS-person subset (39 sequences, 2,707 frames). All values are percentages, frame-averaged then macro-averaged over sequences, including the frame-0 reference frame in every method's scored frames. J and F are region (Jaccard) and boundary similarity, and J&F is their average, following standard DAVIS evaluation protocol.

Method	Reference	J	F	J&F	Precision	Recall
FoundYou	frame-0 mask centroid point	30.19	25.84	28.02	81.83	31.47
PerSegDelta	frame-0 RGB (mask-free)	61.36	58.87	60.12	92.64	60.03
PerSeg-Adapt	frame-0 RGB (mask-free) + online updates	61.37	60.51	60.94	93.86	58.19
PerSeg-DPT	Learned	75.77	69.30	72.54	91.95	77.65

C Ablation Studies

Predictor parameterization PerSegDelta predicts the identity-specific offset in the principal subspace of the teacher offsets (Sec. 3.2) rather than regressing the full KD-dimensional offset directly. Table 7 compares the two parameterizations against the PerSeg-DPT teacher, whose per-identity prompts serve as the regression targets. Both predictors use the same multiscale DINOv2 descriptor, training identities, and evaluation protocol, and differ only in the output space: the PCA variant predicts $r = 256$ coefficients, while the direct variant outputs all $KD = 4{,}096$ offset entries.

Under PODS 2×2 composite evaluation, the PCA predictor closely approaches the teacher, reaching 74.3 IoU compared with 79.2 for PerSeg-DPT (93.8% of the teacher ceiling), with similarly small gaps in boundary F1, precision, and recall. In contrast, direct prediction of the full offset collapses to 24.9 IoU, with precision and recall both near 25, indicating that the predicted prompts fail to localize the target. Since the direct predictor has access to the same features and supervision, this gap reflects generalization rather than capacity: with far fewer training identities than output dimensions, the unconstrained predictor fits the training offsets but produces offsets for unseen subjects that fall outside the region of prompt space learned by the teacher. Restricting predictions to the span of the teacher offsets keeps them within this region and captures the dominant directions of identity variation, which is what enables feed-forward personalization of unseen subjects.

Number of PCA components. Table 8 varies the dimension r of the teacher-offset subspace in which PerSegDelta predicts coefficients (Sec. 3.3). All variants use the same multiscale DINOv2 descriptor, predictor architecture, training identities, and PODS 2×2 composite evaluation as Table 7, and differ only in r. A small subspace ($r = 64$) has low variance and underfis the identities, while a larger subspace ($r = 512$) overfits and loses generalization similar to the direct prompt prediction in Table 7. We use $r = 256$ in all other experiments since it provides a balance between the two.

PerSeg-Adapt loss objective Table 9 ablates the two consistency terms of Eq. 18. Either term alone makes adaptation harmful (-2.19/-2.10 Dice), no better than segmentation alone, whereas combining them gives $+2.39$ point gain over the static prompt. This indicates that inter-class and intra-class consistency are complementary constraints, they keep updates near the target identity while separating it from competing identities.

D Prompt-space geometry.

We further analyze whether predicted prompts share structure across identities. We extract four-token, 4,096-dimensional PerSegDelta prompts for 10,177 CelebA (Liu et al., 2015) identities and 200 CUB species, using one RGB reference image per identity or species, with semantic base

Table 7: Effect of the offset parameterization in PerSegDelta. PerSeg-DPT is the per-identity teacher and provides the ceiling for feed-forward prediction. PCA predicts $r = 256$ coefficients of the teacher-offset subspace; Direct regresses the full $4{,}096$-dimensional offset.

Method	IoU	BF1	Precision	Recall
PerSeg-DPT (teacher)	79.2	63.0	85.1	79.8
PerSegDelta (PCA)	**74.3**	**57.5**	**79.9**	**75.7**
PerSegDelta (Direct)	24.9	22.3	25.9	25.4

Table 8: Effect of the number of PCA components r used to parameterize the PerSegDelta offset, on PODS 2×2 composite evaluation. "Var." is the fraction of teacher-offset variance captured by the first r components. PerSeg-DPT is the per-identity teacher and provides the ceiling for feed-forward prediction.

Method	r	Var.	IoU	BF1	Prec.	Rec.
PerSeg-DPT (teacher)	–	–	79.2	63.0	85.1	79.8
PerSegDelta (PCA)	64	42.2	66.4	50.3	74.6	70.1
PerSegDelta (PCA)	256	71.8	**74.3**	**57.5**	**79.9**	**75.7**
PerSegDelta (PCA)	512	82.4	29.8	31.4	31.7	30.5

prompts *person* and *bird*, respectively. All analyses use standardized coordinates in the first 64 principal components of the predicted prompts, and all probes and retrieval evaluations use identities held out from direction estimation.

Low-dimensional structure. The first 64 components explain 86.5% of the prompt variance on CelebA and 87.7% on CUB. This is expected, since PerSegDelta predicts prompts within a 256-dimensional offset subspace by construction; the analyses below therefore focus on whether the dominant directions are semantically meaningful.

Attribute probes. We fit linear probes on training identities and evaluate them on held-out identities. We also report a *direction AUC*, obtained by projecting prompts onto the difference between positive and negative class centroids without fitting a classifier, and a *direction consistency*, measured as the cosine similarity between directions estimated on disjoint folds. On CelebA (Table 10), probes recover facial attributes with high accuracy, including *Male* (99.6 AUC), *Wearing Lipstick* (97.7), and *Gray Hair* (95.2). Direction AUCs are within 1 point of probe AUCs for all attributes, and consistency exceeds 98%, indicating that each attribute is captured largely by a single stable linear direction. On CUB (Table 12), where continuous class attributes are binarized at the top and bottom tertiles, probes recover bill shape, size, plumage color, and body shape with 78.3–88.6 AUC. Performance is lower than on CelebA, consistent with CUB's 200 species providing far fewer samples than CelebA's 10,177 identities, but direction consistency remains above 85%.

Prompt arithmetic. To test whether these directions transfer across identities, we take a held-out prompt lacking an attribute, add the attribute direction estimated only from training identities, and measure the fraction of its 10 nearest held-out prompts that have the attribute. We compare against the unmodified prompt (*Baseline*) and an equal-norm random direction. On CelebA (Table 11), the attribute direction raises this rate from 12.8% to 50.4% for *Male* and from 16.0% to 50.0% for *Wearing Lipstick*, with gains of 17.6–35.9 points over the random control across all attributes. On CUB (Table 13), gains over the random control range from 14.5 to 21.4 points, e.g., from 28.2% to 50.5% for an all-purpose bill shape.

Together, these results suggest that PerSegDelta does not merely memorize individual prompt targets: it predicts offsets in a structured space whose directions encode reusable semantic and taxonomic attributes.

Table 9: PerSeg-Adapt loss-objective ablation on a held-out validation split of YouTube-VOS videos. All rows share the same static PerSegDelta prompt; only the consistency terms used to train the online adaptation policy differ. Dice values are frame-averaged percentages.

Loss objective	Static	Dynamic	Gain
Segmentation only	70.05	70.30	$+0.25$
Inter-class only ($\mathcal{L}_{\text{inter}}$)	70.05	67.86	-2.19
Intra-class only ($\mathcal{L}_{\text{intra}}$)	70.05	67.94	-2.10
$\mathcal{L}_{\text{adapt}}$ (full)	70.05	**72.44**	**+2.39**

Table 10: Held-out annotation probes from PerSegDelta predicted prompts. AUC is measured across identities; direction AUC uses a centroid difference without fitting a classifier.

Attribute	Probe AUC	Direction AUC	Direction consistency
Male	99.6	99.5	99.8
Wearing Lipstick	97.7	97.6	99.7
Gray Hair	95.2	94.6	99.3
Heavy Makeup	94.7	94.7	99.7
Wearing Necktie	94.2	93.3	99.4
Bald	93.0	92.3	98.4
Double Chin	91.7	91.3	99.0
No Beard	91.6	90.8	99.3

E Standard (Non-Composite) Single-Reference Segmentation Comparison

Table 1 evaluates all methods on center-cropped 2×2 distractor composites. Here we report the simpler, non-composite counterpart: one (image, mask) reference pair per identity drawn from the train split, evaluated against ordinary query images from the plain PODS and Dogface test splits, with no distractor grid. FoundYou, PerSAM, PerSAM-F, and PDM receive the reference image–mask pair at test time; PerSegDelta receives only the unmasked reference image and predicts its prompt without per-identity optimization. We report PerSegDelta, using PCA coefficients predicted from a shared PCA basis fit jointly over PODS, Dogface, CUB and GOT training identities.

The ranking under this simpler, non-composite protocol differs from the composite 2×2 setting. Without distractors to reject, FoundYou – a mask-conditioned correspondence method that struggled on PODS composites – performs better, leading boundary F1 on both PODS and dogs and recall on dogs. PerSegDelta obtains the best results overall despite using no reference mask, leading mIoU and precision on both PODS and dogs and recall on PODS. PDM and the PerSAM variants, which trailed in the composite comparison, remain the weakest methods here as well. Comparing PerSegDelta's PODS mIoU here (86.9) with its composite PODS row in Table 1 (74.3 mIoU) shows the expected direction: the plain protocol has no distractor grid to reject, so mIoU is higher across the board, and the composite comparison remains the more discriminative test of identity-specific personalization.

F Qualitative Results

This section shows qualitative examples of PerSeg on controlled distractor composites and on natural multi-instance scenes. In all figures, the ground-truth target mask is shown in green and the predicted mask in blue.

Controlled distractor composites. Figs. 4–6 show 2×2 composites from the negative-difficulty benchmark (Sec. 4). In each, the target appears in one cell and three distractors fill the rest; only the target mask is evaluated. In Fig. 4, PerSegDelta segments a held-out dog from a single unmasked RGB reference (IoU 99.1). It does not respond to the three other dogs, two of which share its reddish-brown color which alone cannot identify the target. Figs. 5 and 6 show PerSeg-DPT prompts on two mug identities from PODS. In Fig. 5, the prompt segments the blue mug (IoU 97.1) and rejects other different mugs that are occluded. Fig. 6 is harder: all four cells contain mugs of similar shape, placed on the same stand, and one distractor has the same dark-blue color as the target. The prompt

Table 11: Held-out predicted-prompt vector addition. For a negative-source identity, we add a feature direction estimated only from training identities and retrieve the 10 nearest held-out prompts.

Attribute	Baseline	Added direction	Gain over random
Male	12.8	50.4	+35.9
Wearing Lipstick	16.0	50.0	+31.1
Heavy Makeup	17.5	47.2	+28.3
Gray Hair	3.7	23.6	+19.8
Wearing Necktie	4.4	23.3	+19.3
No Beard	56.5	76.2	+17.6

Table 12: Held-out annotation probes from PerSegDelta predicted CUB prompts, for continuous class-attributes binarized at the top/bottom tertile across the 200 species. AUC is measured across held-out species folds; direction AUC uses a centroid difference without fitting a classifier.

Attribute	Probe AUC	Direction AUC	Direction consistency
bill shape: all-purpose	88.6	88.2	87.7
size: very small 3 - 5 in	84.6	85.5	87.5
wing color: brown	83.5	82.8	86.7
under tail color: black	82.3	84.4	86.7
shape: perching-like	80.3	82.6	86.5
primary color: white	79.5	82.1	86.3
underparts color: white	78.7	81.5	87.3
primary color: black	78.3	78.7	85.9

still isolates the target (IoU 95.6), indicating that it encodes cues beyond category, color, and scene context, such as the printed logo.

Natural multi-instance scenes. Figs. 7(a) and 7(b) show PerSegDelta predictions on real images, with the RGB reference on the left and the query on the right. In Fig. 7(a), the reference van is segmented in a busy street scene with several white cars, which share its color but not its shape. Fig. 7(b) shows a harder case: the query contains three puppies of the same breed, in a different pose and setting from the reference. PerSegDelta segments only the puppy whose coat pattern matches the reference and rejects the two tricolor puppies. These examples illustrate the setting that motivates PerSeg: the target must be separated from same-category instances whose differences are small, using a prompt predicted from a single reference without a mask.



Figure 4: PerSegDelta on a held-out dog composite. Left: input composite. Middle: ground-truth target mask (green). Right: PerSegDelta prediction (blue), IoU 99.1. The prompt is predicted from a single unmasked RGB reference, and the three same-category distractors are not segmented.

G Implementation details

All stages use a frozen SAM3 image model. Personalization prompts consist of $K = 4$ token embeddings of dimension $D = 1024$ at the input of the SAM3 text encoder, and the semantic prompt

Table 13: Held-out predicted-prompt vector addition on CUB. For a species without the target binarized attribute, we add a direction estimated only from training species and retrieve the 10 nearest held-out prompts.

Attribute	Baseline	Added direction	Gain over random
bill shape: all-purpose	28.2	50.5	+21.4
under tail color: black	24.8	41.7	+20.3
primary color: white	26.4	43.0	+17.2
size: very small 3 - 5 in	26.9	45.0	+16.2
primary color: black	24.7	40.9	+15.2
head pattern: plain	28.3	41.3	+14.5

Table 14: Non-composite, single-reference personalized segmentation on the standard PODS and DogFace test splits (Section E). "Mask" indicates whether the method receives a reference mask at test time: FoundYou, PerSAM/PerSAM-F and PDM use a one-shot image–mask reference, while PerSegDelta is mask-free. Bold marks the best value in each column.

		PODS				DogFace			
Method	Mask	mIoU	BF1	Prec.	Rec.	mIoU	BF1	Prec.	Rec.
FoundYou	Yes	81.8	**68.2**	87.3	87.2	91.9	**72.3**	95.9	**95.0**
PerSAM	Yes	26.2	30.3	31.4	29.2	54.2	49.3	60.8	56.4
PerSAM-F	Yes	68.5	55.0	75.6	72.0	87.8	68.8	93.1	89.4
PDM	Yes	62.2	61.6	63.8	84.4	75.2	64.8	76.5	89.2
PerSegDelta	**No**	**86.9**	67.2	**95.8**	**88.2**	**93.0**	71.4	**99.5**	93.4

$\mathbf{p}_{\text{base}}(c)$ is the token embedding of the last four tokens of "photo of a c". PerSeg-DPT is trained with batch size one, sampling instances uniformly and positive and same-class negative examples with equal probability, for 4,000 iterations per instance on average, with $\lambda_{\text{bce}} = \lambda_{\text{dice}} = \lambda_{\text{score}} = \lambda_{\text{pres}} = \lambda_{\text{DRS}} = 1$ and $\lambda_{\text{id}} = 0.1$. PerSegDelta is trained on the DPT teacher offsets of 1000 training identities drawn from PODS, Dogs, CUB-200-2011 (Wah et al., 2011), and GOT (Huang et al., 2021) training identities, using the semantic classes "dog", "bird", and "person" for the latter three. Its reference descriptor concatenates the class tokens of blocks $\{2, 5, 8, 11\}$ of a frozen DINOv2 ViT-B/14 (Oquab et al., 2024), and the predictor is an MLP with one hidden layer of width 1,024 and GELU activation that outputs $r = 256$ PCA coefficients. It is trained with the mean squared error on the coefficients for 200 epochs with batch size 32. For video, the PerSegDelta predictor is fine-tuned on YouTube-VOS teacher offsets, and the PerSeg-Adapt policy is trained on 80% of the YouTube-VOS videos, with the remaining 20% held out for validation. The policy updates the prompt every $T = 5$ frames with similarity threshold $\tau_{\text{sim}} = 0.6$ and uses $\lambda_{\text{inter}} = \lambda_{\text{intra}} = 0.2$; it is trained for 5 epochs with learning rate 3×10^{-4} and $\tau = 1$. In all video experiments, the first-frame reference crop of the target is taken from the dataset annotation.

H Limitations and Future Work

First, PerSegDelta predicts the personalized prompt from an object-centric reference image and conditions on the category name c through the semantic base prompt $\mathbf{p}_{\text{base}}(c)$. It therefore does not require a reference mask, but it assumes that the target is the dominant object in the reference and that its category is known or can be predicted, e.g., by a vision-language model as in our retrieval experiments. Relaxing these assumptions, for instance by predicting the offset in a shared vision-language space directly from uncropped scenes, is a natural extension.

Second, PerSegDelta removes per-identity optimization at test time, but its training relies on PerSeg-DPT prompts for a set of training identities. The coverage of these identities determines the offset subspace, and hence the range of subjects to which the predictor generalizes. Training the predictor end-to-end with segmentation and distractor losses, without an intermediate teacher, could reduce this dependence and extend coverage to broader categories.

Another promising direction is to model the distribution of temporal prompts, for example with a generative model whose sampling can be guided away from distractors as they are observed during the video.



Figure 5: PerSeg-DPT on same-category PODS mug composite. Left: input composite. Middle: ground-truth target mask (green). Right: prediction (blue), IoU 97.1. The prompt rejects partially visible same-category distractor mugs of different colors.

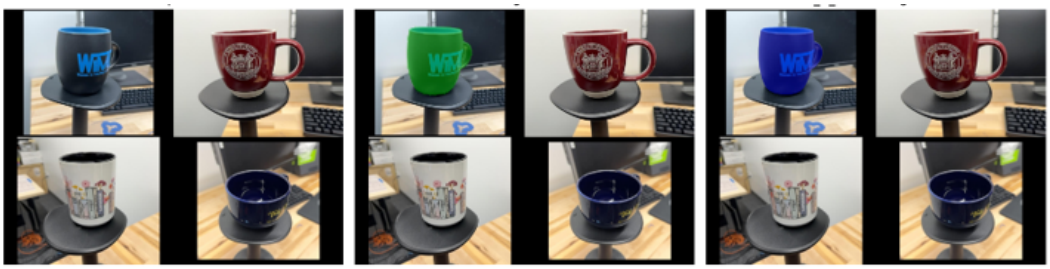


Figure 6: PerSeg-DPT on visually similar PODS mug composite. Left: input composite. Middle: ground-truth target mask (green). Right: prediction (blue), IoU 95.6. The prompt rejects visually similar distractor mug (bottom right) and other instances of the same mug category.

Ethics Statement

This work studies personalized segmentation and retrieval: localizing or retrieving a specific individual subject, rather than a category, from a single reference image. Such capabilities have beneficial uses, including personal photo organization, assistive robotics, and wildlife and pet monitoring. However, they could also be misused for surveillance, tracking, or re-identification of people without their consent, particularly when applied to the *person* category. Our experiments include person identities from GOT-10k, DAVIS, and YouTube-VOS, and an analysis of prompt-space geometry on CelebA. We use these datasets only to measure identity discrimination and attribute structure, not to build or evaluate face-recognition or person-tracking systems. The CelebA attribute analysis (e.g., *Male*, *Wearing Lipstick*) uses the dataset's existing annotations and is reported to characterize the learned representation; we do not endorse inferring sensitive or demographic attributes from images, and such annotations may encode social biases.

All datasets used in this work (PODS, DreamBooth, DogFace, CUB-200-2011, CelebA, GOT-10k, DAVIS, YouTube-VOS, and ILIAS) are publicly available and were used in accordance with their licenses and intended research use. We did not collect new images or annotations of people; our composite benchmarks are constructed solely by recombining crops from existing datasets. Upon release, we will distribute code and benchmark construction scripts under terms that restrict use to research, and will document the intended use and known limitations of the models, including the risk of misuse for person re-identification.

AI Use Statement

In this work, we used generative AI tools for polishing the writing of the paper, including improving grammar, clarity, and conciseness of text drafted by the authors. We have not used generative AI tools for generating research ideas, designing experiments, analyzing data, or producing results, figures, or tables. Additionally, we did not use generative AI tools for literature search or for drafting the technical content of the paper. We have reviewed all AI-assisted work: every suggested edit was



(a) Street scene (b) Same-breed distractors

Figure 7: PerSegDelta on natural multi-instance scenes. In each panel, the left image is the RGB reference and the right image shows the prediction (blue) on the query image. (a) The van is segmented while the surrounding white cars are rejected. (b) Only the puppy matching the reference color pattern is segmented; the other two tricolor puppies are rejected.

read and checked by the authors to ensure that it did not alter the technical meaning, claims, or reported numbers.