Personalize Segment Anything Model with One Shot

Renrui Zhang[1,2], Zhengkai Jiang[3], Ziyu Guo[1], Shilin Yan[1], Junting Pan[2]
Hao Dong[4], Peng Gao[1], Hongsheng Li[2]

[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]Tencent Youtu Lab [4]CFCS, School of CS, Peking University
{zhangrenrui, gaopeng}@pjlab.org.cn, zhengkjiang@tencent.com

Abstract

Driven by large-data pre-training, Segment Anything Model (SAM) has been demonstrated as a powerful and promptable framework, revolutionizing the segmentation models. Despite the generality, customizing SAM for specific visual concepts without man-powered prompting is under explored, e.g., automatically segmenting your pet dog in different images. In this paper, we propose a training-free ***Per****sonalization approach for SAM, termed as* ***PerSAM****. Given only a single image with a reference mask, PerSAM first localizes the target concept by a location prior, and segments it within other images or videos via three techniques: target-guided attention, target-semantic prompting, and cascaded post-refinement. In this way, we effectively adapt SAM for private use without any training. To further alleviate the mask ambiguity, we present an efficient one-shot fine-tuning variant,* ***PerSAM-F****. Freezing the entire SAM, we introduce two learnable weights for multi-scale masks, only training* ***2 parameters*** *within* ***10 seconds*** *for improved performance. To demonstrate our efficacy, we construct a new segmentation dataset, PerSeg, for personalized evaluation, and test our methods on video object segmentation with competitive performance. Besides, our approach can also enhance DreamBooth to personalize Stable Diffusion for text-to-image generation, which discards the background disturbance for better target appearance learning. Code is released at* https://github.com/ZrrSkywalker/Personalize-SAM.

1. Introduction

Foundations models in vision [11, 30, 51, 63], language [4, 10, 43, 50], and multi-modality [21, 31, 41] have gained unprecedented prevalence, ascribed to the considerable availability of pre-training data and computational resources. They demonstrate extraordinary generalization capacity in zero-shot scenarios, and display versatile interactivity incorporating human feedback. Inspired by the achievements of large language models, Segment Anything (SAM) [27] develops a delicate data engine for collecting 11M image-mask data, and subsequently trains a powerful segmentation foundation model, known as SAM. It firstly defines a novel promptable segmentation paradigm, i.e., taking as input a handcrafted prompt and returning the expected mask. The acceptable prompt of SAM is generic enough, including points, boxes, masks and free-form texts, which allows for segmenting anything in visual contexts.

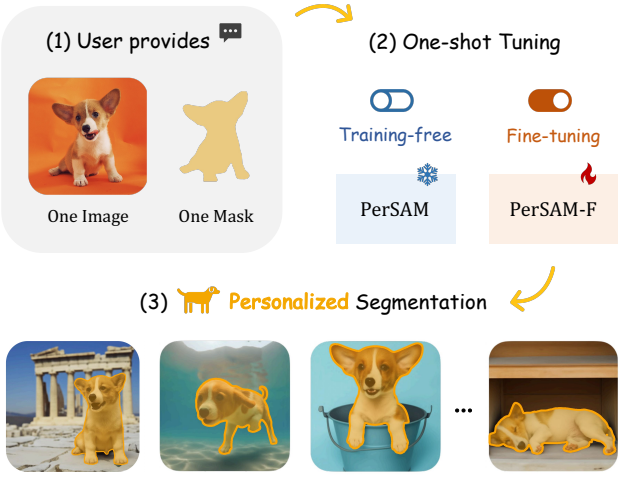

Figure 1. **Personalization of Segment Anything Model.** We customize Segment Anything Model (SAM) [27] for specific visual concepts, e.g., your pet dog. With only one-shot data, we introduce two efficient solutions: a training-free PerSAM, and a fine-tuning PerSAM-F. Image examples are from DreamBooth [45].

However, SAM inherently loses the capability to segment specific visual concepts. Imagine intending to crop your lovely pet dog in a photo album, or find the missing clock from a picture of your bedroom. Utilizing the vanilla SAM model would be both labor-intensive and time-

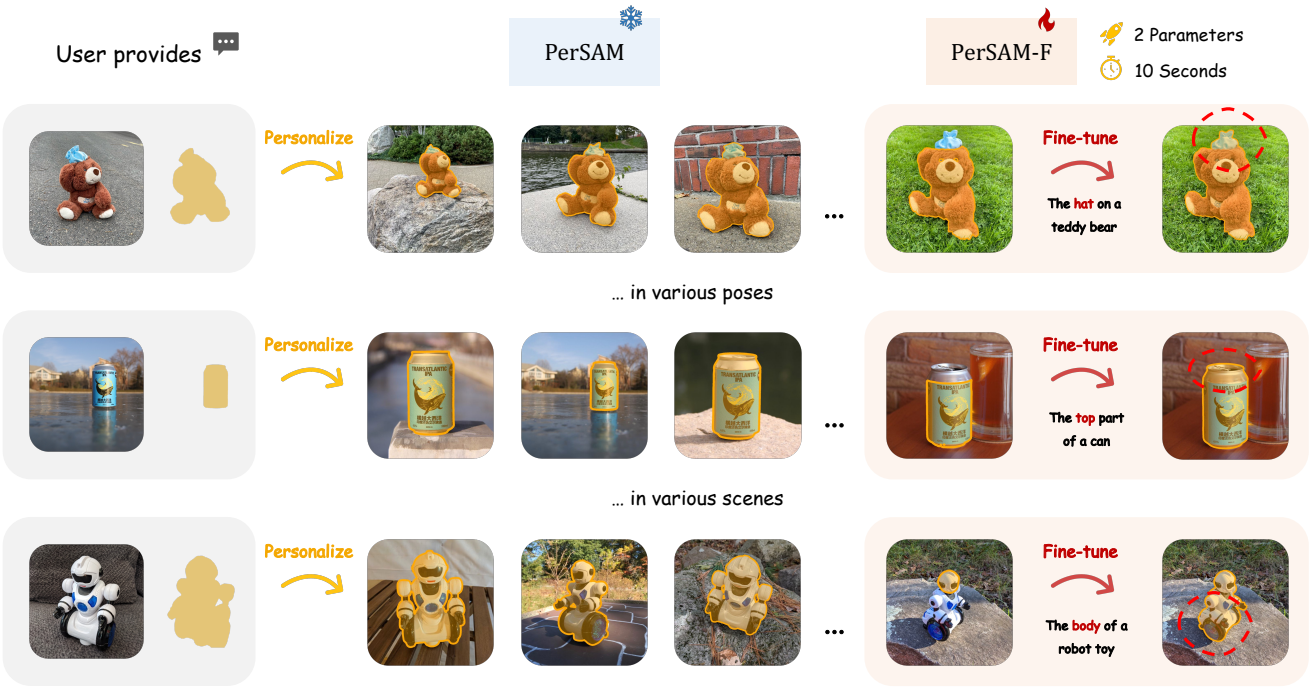

Figure 2. **Personlization Examples of Our Approach.** The training-free PerSAM (Left) customizes SAM [27] to segment user-provided objects in any poses or scenes with favorable performance. On top of this, PerSAM-F (Right) further enhances the segmentation accuracy by efficiently fine-tuning only 2 parameters within 10 seconds. Examples are from our annotated dataset, PerSeg.

consuming. For each image, you are required to locate the target object in different poses or contexts, and then activate SAM with precise prompt for segmentation. Therefore, we ask: *Can we personalize SAM to automatically segment unique visual concepts in a simple and efficient manner?*

To this end, we propose **PerSAM**, a training-free personalization approach for Segment Anything Model. As shown in Figure 1, our method efficiently customizes SAM using only one-shot data, i.e., a user-provided image and a rough mask designating the personal concept. Specifically, we first utilize SAM's image encoder and the given mask to encode the embedding of the target object in the reference image. Then, we calculate the feature similarity between the object and all the pixels on the new test image. On top of this, two points are selected as the positive-negative pair, which are encoded as prompt tokens and serve as a location prior for SAM. Within SAM's decoder processing the test image, we introduce three techniques to unleash its personalizetion potential without parameter tuning.

- **Target-guided Attention.** We guide every token-to-image cross-attention layer in SAM's decoder by the calculated feature similarity. This compels the prompt tokens to mainly concentrate on foreground target regions for effective feature interaction.
- **Target-semantic Prompting.** To better provide SAM with high-level target semantics, we fuse the original low-level prompt tokens with the embedding of target object, which provides the decoder with more sufficient visual cues for personalized segmentation.
- **Cascaded Post-refinement.** For finer segmentation results, we adopt a two-step post-refinement strategy. We utilize SAM to progressively refine its generated mask. This process only costs an extra 100ms.

With the aforementioned designs, PerSAM exerts favorable personalized segmentation performance for unique subject in a variety of poses or contexts, as visualized in Figure 2. Nevertheless, there might be occasional failure cases, where the subject comprises hierarchical structures to be segmented, e.g., a hat on top of a teddy bear, the head of a robot toy, or the top part of a can. Such ambiguity casts a challenge for PerSAM in determining the appropriate scale of mask as the segmentation output, since both the local part and the global shape can be regarded as valid masks by SAM from the pixel level.

To alleviate this, we further introduce a fine-tuning variant of our approach, **PerSAM-F**. We freeze the entire SAM to preserve its pre-trained knowledge, and only fine-tunes ***2 parameters*** within ***10 seconds***. In detail, we enable SAM to produce multiple segmentation results with different mask scales. To adaptively select the best scale for varying ob-

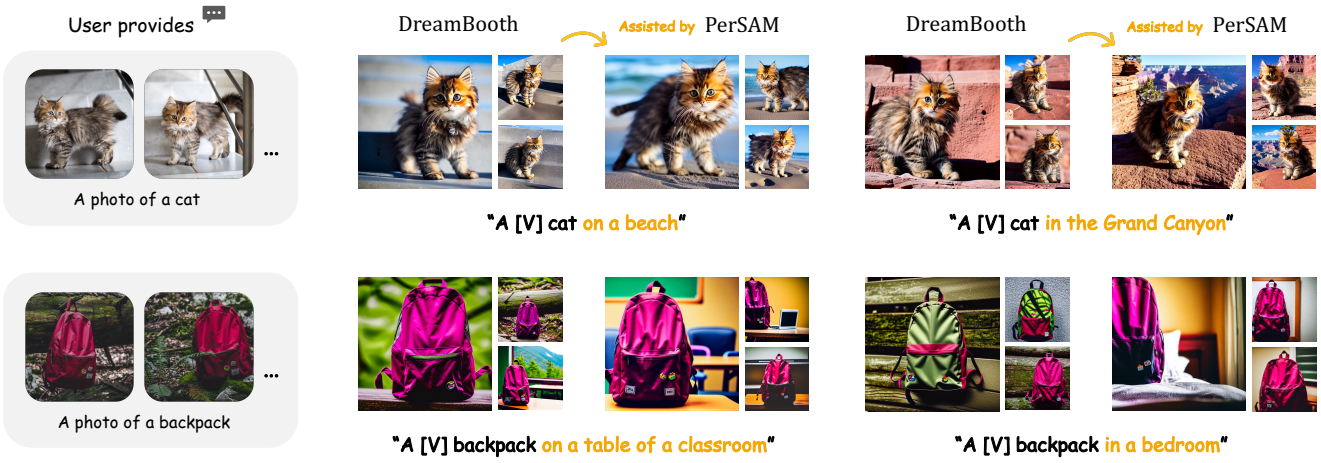

Figure 3. **Better Personalization of Stable Diffusion.** Our approach can be utilized to assist DreamBooth [45] in fine-tuning better Stable Diffusion [44] for personalized image synthesis. We adopt PerSAM to segment the target object in the user-provided few-shot images, which eliminates the background disturbance and benefits the target representation learning. More visualization is in Figure 10.

jects, we employ learnable relative weights for each scale, and conduct a weighted summation as the final mask output. By such efficient one-shot training, PerSAM-T exhibits better segmentation accuracy shown in Figure 2 (Right). Instead of using prompt tuning [29] or adapters [19], the ambiguity issue can be effectively restrained by efficiently weighing multi-scale masks.

Moreover, we observe that our approach can also assist DreamBooth [45] to better fine-tune Stable Diffusion [44] for personalized text-to-image generation, as shown in Figure 3. Given a few images containing a specific visual concept, e.g., you pet cat, DreamBooth and its other works [28] convert these images into an identifier [V] in the word embedding space, which is then utilized to represent the target object in the sentence. However, the identifier simultaneously include the visual information of backgrounds in the given images, e.g., stairs. This would not only override the new backgrounds in the generated images, but also disturb the representation learning of the target object. Therefore, we propose to leverage our PerSAM to efficiently segment the target object, and only supervise Stable Diffusion by the foreground area in the few-shot images, enabling more diverse and higher-fidelity synthesis.

We summarize the contributions of our paper as follows:

- **Personalized Segmentation Task.** From a new standpoint, we investigate how to customize segmentation foundation models into personalized scenarios with minimal expense, i.e., from general to private purpose.
- **Efficient Adaption of SAM.** For the first time, we research on adapting SAM to downstream applications by only fine-tuning 2 parameters, and propose two lightweight solutions: PerSAM and PerSAM-F.
- **Personalization Evaluation.** We annotate a new segmentation dataset, PerSeg, containing various categories in diverse contexts. We also test our approach on video object segmentation with competitive results.
- **Better Personalization of Stable Diffusion.** By segmenting the target object in the few-shot images, we alleviate the disturbance of backgrounds and improve the personalized generation of DreamBooth.

2. Related Work

Segmentation in Vision. As a fundamental task in computer vision, segmentation [23, 24, 34, 36, 55, 60] requires a pixel-level comprehension of a given image. Multiple segmentation-related tasks have been explored, such as semantic segmentation, which classifies each pixel into a predefined set of classes [1, 5, 7, 47, 54, 61]; instance segmentation, focusing on the identification of individual object instances [18, 49, 52]; panoptic segmentation, combining semantic and instance segmentation tasks by assigning both class labels and instance identification [26, 32]; and interactive segmentation, involving human intervention during the segmentation process for refinement [6, 16]. Recently, Segment Anything Model (SAM) [27] designs a promptable segmentation task and achieves strong zero-shot generalization on numerous image distributions. The concurrent SegGPT [53] and SEEM [63] also present general frameworks for a diversity of segmentation scenarios. In this study, we introduce a new task termed personalized segmentation, aiming to segment user-provided objects in any unseen poses or scenes. We propose two approaches, PerSAM and PerSAM-F, to efficiently customize SAM for personalized segmentation.

Foundation Models. With powerful generalization capacity, the pre-trained foundation models can be adapted for various downstream tasks with promising performance. In the field of natural language processing, BERT [10, 38], GPT series [4, 39, 42, 43], and LLaMA [58] have demonstrated remarkable in-context learning abilities. These models can be transferred to new language tasks without training, requiring only a few task-specific prompts during inference. Similarly, CLIP [41] and ALIGN [21], which are trained on web-scale image-text pairs using contrastive loss, exhibit exceptional performance in zero-shot visual learning tasks. Painter [51] introduces a vision model that unifies architectures and prompts to automatically accomplish diverse vision tasks without the necessity for task-specific heads. CaFo [59] cascades different foundation models and collaborates their pre-trained knowledge for zero-shot image classification. SAM [27] presents the first foundation model for image segmentation, which is pre-trained on 1 billion masks and conditioned on a variety of input prompts, e.g., point, bounding box, mask, and text. From another perspective, we propose to personalize the foundation segmentation model, i.e., SAM, for specific visual concepts, which adapts a generalist into a specialist with only one shot. Our method can also assist the personalization of text-to-image foundation models, i.e., Stable Diffusion [44] and Imagen [46], which improves the generation quality by segmenting the target objects from the background area.

Parameter-efficient Fine-tuning. Directly tuning the entire foundation models on downstream tasks can be computationally expensive and memory-intensive, posing challenges for resource-constrained applications. To address this issue, recent works have focused on developing parameter-efficient methods [15, 17, 48, 57] to freeze the weights of foundation models and append small-scale modules for fine-tuning. Prompt Tuning [13, 22, 29, 62] suggests using learnable soft prompts alongside frozen models to perform specific downstream tasks, achieving more competitive performance with scale and robust domain transfer compared to full model tuning. Low-Rank Adaption (LoRA) [9, 20] injects trainable rank decomposition matrices concurrently to each pre-trained weights, significantly reducing the number of learnable parameters required for downstream tasks. Adapters [19, 56] are designed to be inserted between layers of the original transformer, which introduce lightweight MLPs for fine-tuning. LLaMA-Adapter [14, 58] proposes a zero-init attention to progressive incorporate new knowledge into foundation models, stabilizing the early-stage training. Different from existing works, we adopt a more efficient adaption method for SAM by either the training-free PerSAM, or PerSAM-F fine-tuning only 2 parameters. This effectively avoids the over-fitting on one-shot data with satisfactory performance.

3. Method

In Section 3.1, we first revisit Segment Anything Model (SAM) [27] and introduce the task definition of personalized segmentation. Then, we illustrate the methodology of our training-free PerSAM and its fine-tuned variant, PerSAM-F, in Section 3.2 and Section 3.3, respectively. Finally in Section 3.4, we utilize our approach to assist DreamBooth [45] in better personalizing Stable Diffusion [44] for text-to-image generation.

3.1. Preliminary

A Revisit of Segment Anything. SAM defines a new promptable segmentation task, the goal of which is to return a segmentation mask for any given prompt. Using a data engine with model-in-the-loop annotation, SAM is fully pre-trained by 1 billion masks on 11M images, enabling powerful generalization capacity. SAM consists of three main components, a prompt encoder, an image encoder, and a lightweight mask decoder, which we respectively denote as Enc_P, Enc_I, and Dec_M. As a promptable framework, SAM takes as input an image I, and a set of prompts P, e.g., foreground or background points, bounding boxes, or a coarse mask to be refined. SAM first utilizes Enc_I to obtain the input image feature and Enc_P to encode the human-given prompts into c-dimensional tokens as

$$F_I = \text{Enc}_I(I), \quad T_P = \text{Enc}_P(P), \tag{1}$$

where $F_I \in \mathbb{R}^{h \times w \times c}$ and $T_P \in \mathbb{R}^{k \times c}$, with h, w denoting the resolution of the image feature and k denoting the prompt length. After that, the encoded image and prompts are fed into the decoder Dec_M for attention-based feature interaction. SAM constructs the input tokens of the decoder by concatenating several learnable tokens T_M as prefix to the prompt tokens. These mask tokens are responsible for generating the final mask output. We formulate the decoding process as

$$M = \text{Dec}_M\Big(F_I,\ \text{Concat}(T_M, T_P)\Big), \tag{2}$$

where M denotes the zero-shot mask prediction by SAM.

Personalized Segmentation Task. In spite that, SAM is generalized to segment anything that prompted by users, it lacks the ability to segment specific subject instances. To this end, we define a new task for personalized segmentation. The user provides only a single reference image, along with a mask indicating the target visual concept. The given mask can either be an accurate segmentation, or a rough sketch drawn by users online. Our goal is to customize SAM to segment the designated subject within new images or videos, without man-powered prompting. For model

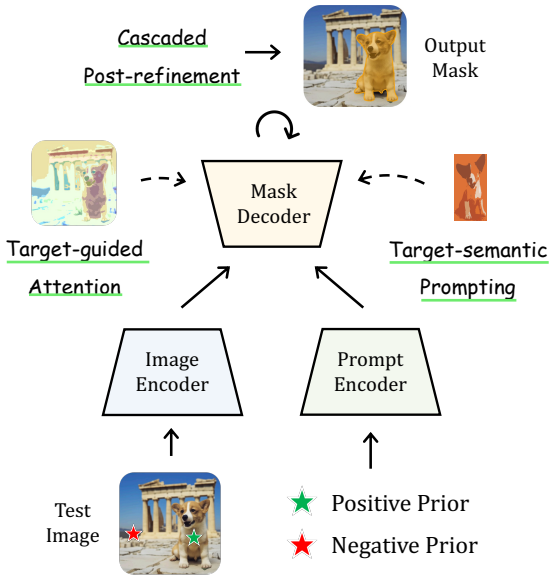

Figure 4. **Overall Pipeline of PerSAM.** Prompted by positive and negative location priors, PerSAM achieves training-free personalized segmentation with three techniques: target-guided attention, target-semantic prompting, and cascaded post-refinement.

evaluation, we annotate a new dataset for personalized segmentation, named PerSeg. The raw images are taken from the works for subject-driven diffusion models [12, 28, 45], containing various categories of visual concepts in different poses or scenes. In this paper, we propose two efficient solutions for this task, PerSAM and PerSAM-F, which we specifically illustrate as follows.

3.2. Training-free PerSAM

Positive-negative Location Prior. Figure 4 presents the overall pipeline of our training-free PerSAM. Firstly, conditioned on the user-provided image I_R and mask M_R, PerSAM obtains a location prior of the target object on the new test image I using SAM. In detail, as shown in Figure 5, we apply SAM's pre-trained image encoder to extract the visual features of both I and I_R as

$$F_I = \text{Enc}_I(I), \quad F_R = \text{Enc}_I(I_R), \tag{3}$$

where $F_I, F_R \in \mathbb{R}^{h\times w\times c}$. Then, we utilize the reference mask $M_R \in \mathbb{R}^{h\times w\times 1}$ to derive the features of pixels within the target visual concept from F_R, and adopts an average pooling to aggregate its global visual embedding $T_R \in \mathbb{R}^{1\times c}$ as

$$T_R = \text{Pooling}(M_R \circ F_R), \tag{4}$$

where $\circ$ denotes spatial-wise multiplication. With the target embedding T_R, we can acquire a location confidence map by calculating the cosine similarity S between the T_R and test image feature F_I as

$$S = F_I T_R^T \in \mathbb{R}^{h\times w}, \tag{5}$$

where F_I and T_R are pixel-wisely L2-normalized. After this, to provide SAM with a location prior on the test image, we select two pixel coordinates with the highest and lowest similarity values from S, denoted as P_h and P_l, respectively. The former represents the most likely foreground position of the target object, while the latter inversely indicates the background. Then, they are regarded as the positive and negative point pair, and fed into the prompt encoder as

$$T_P = \text{Enc}_P(P_h, P_l), \tag{6}$$

where $T_P \in \mathbb{R}^{2\times c}$ serves as the prompt tokens for SAM's decoder. In this way, SAM would tend to segment the contiguous region surrounding the positive point, while discarding the negative one's on the test image.

Target-guided Attention. Although the positive-negative prior has been employed, we further propose a more explicit guidance to the cross-attention mechanisms in SAM's decoder, which concentrates feature aggregation within foreground target regions. As shown in Figure 6, the calculated similarity map S in Equation 5 can clearly indicate the pixels within the target visual concept on the test image. Given this, we utilize S to modulate the attention map in every token-to-image cross-attention layer. We denote the attention map after the Softmax function as $A \in \mathbb{R}^{h\times w}$, and guide its distribution by

$$A^g = \text{Softmax}\Big(A + \alpha \cdot \text{Softmax}(S)\Big), \tag{7}$$

where α denotes a balance factor. By the attention bias, the tokens are compelled to capture more visual semantics associating with the target subject, other than the unimportant background. This contributes to more effective feature interaction in attention layers, and enhances the final segmentation accuracy of PerSAM in a training-free manner.

Target-semantic Prompting. The vanilla SAM only receives the prompt carrying low-level positional information, such as the coordinates of points or boxes. For incorporating more personalized cues, we propose to additionally utilize the visual embedding T_R of the target concept as a high-level semantic prompt for PerSAM. Specifically, we element-wisely add the target embedding with all the input tokens in Equation 2, before feeding into every decoder block shown in Figure 6, formulated as

$$\text{Repeat}(T_R) + \text{Concat}(T_M, T_P), \tag{8}$$

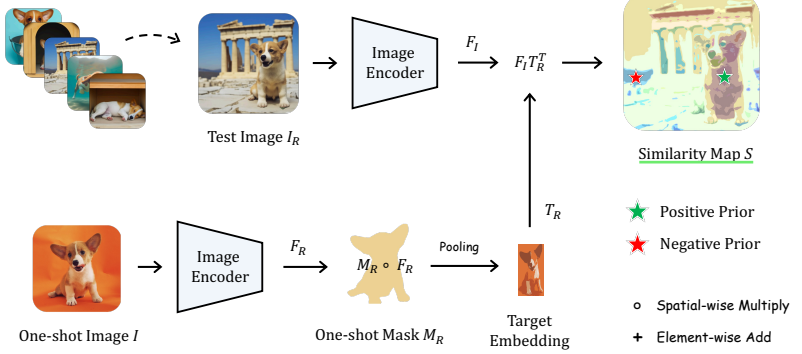

Figure 5. **Positive-negative Location Prior.** To obtain a location prior on the test image, we adopt SAM's [27] image encoder to extract visual features, and calculate a similarity map for positive-negative point selection, which provides PerSAM with foreground and background cues without human prompting.

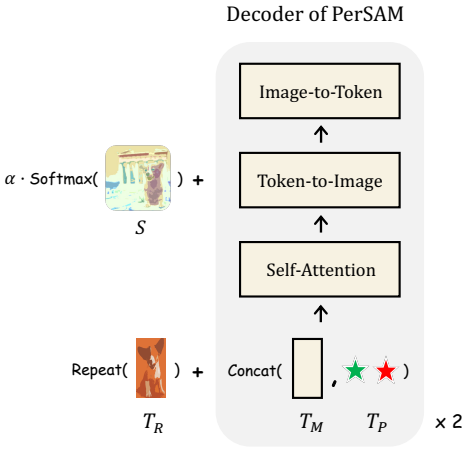

Figure 6. **Target-guided Attention & Target-semantic Prompting.** In PerSAM's decoder, we incorporate the semantics of target object for attention guidance and high-level prompting.

where the Repeat operation is performed alone the token dimension. Aided by the simple token incorporation, PerSAM is not only prompted by low-level location prior, but also high-level target semantics with auxiliary visual cues.

Cascaded Post-refinement. Via the above techniques, we obtain an initial segmentation mask on the test image from SAM's decoder, which however, might include some rough edges and isolated noises in the background. For further refinement, we iteratively feed the mask back into SAM's decoder for a two-step post-processing. In the first step, we prompt SAM's decoder by the initial mask along with the previous positive-negative location prior. Then, for the second step, we calculate a bounding box of the mask from the first step, and prompt the decoder additionally with this box for more accurate object localization. As we only requires the lightweight decoder for iterative refinement without the large-scale image encoder, the post-processing is efficient and only costs an additional 100ms.

3.3. Fine-tuning of PerSAM-F

Ambiguity of Mask Scales. The training-free PerSAM can tackle most cases with satisfactory segmentation accuracy. However, some target objects contain hierarchical structures, which leads to several masks of different scales to be segmented. As shown in Figure 7, the teapot on top of a platform is comprised of two parts: a lid and a body. If the positive prior (denoted by a green star) is located at the body, while the negative prior (denoted by a red star) does not exclude the platform in a similar color, PerSAM would be ambiguous for segmentation. Such issue is also discussed in SAM [27], where it proposes an alternative to simultaneously generate multiple masks of three scales, respectively corresponding to the whole, part, and subpart of an object. Then, the user is required to manually select one mask out of three, which is effective but consumes extra manpower. In contrast, our personalized task aims to customize SAM for automatic object segmentation without the need for human prompting. This motivates us to develop a scale-aware personalization approach for SAM by efficiently fine-tuning only a few parameters.

Learnable Scale Weights. For adaptive segmentation with appropriate mask scale, we introduce a fine-tuning variant, PerSAM-F. Unlike the training-free model only producing one mask, PerSAM-F first refers to SAM's solution to output three-scale masks, denoted as M_1, M_2, and M_3, respectively. On top of this, we adopt two learnable mask weights, w_1, w_2, and calculate the final mask output by a weighted summation as

$$M = w_1 \cdot M_1 + w_2 \cdot M_2 + (1 - w_1 - w_2) \cdot M_3, \quad (9)$$

where w_1, w_2 are both initialized as $1/3$. To learn the optimal weights, we conduct one-shot fine-tuning on the reference image, and regard the given mask as the ground truth. Note that, we freeze the entire SAM model to preserve its pre-trained knowledge, and only fine-tunes the ***2 parameters*** of w_1, w_2 within ***10 seconds***. We do not adopt any learnable prompts or adapter modules to avoid overfitting on the one-shot data. In this way, our PerSAM-F efficiently learns the best mask scale for different visual concepts, and exhibits stronger segmentation performance than the training-free PerSAM.

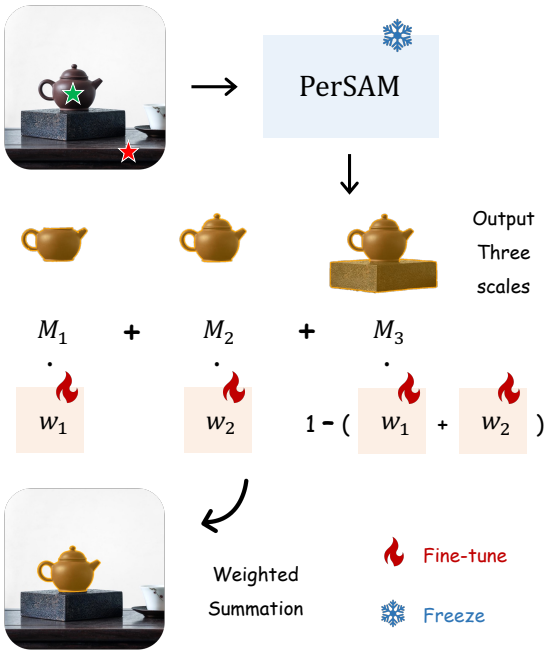

Figure 7. **Efficient Fine-tuning of PerSAM-F.** To alleviate the ambiguity of segmentation scales, PerSAM-F adopts two learnable weights to adaptively aggregate the output masks of three scales by efficient one-shot fine-tuning.

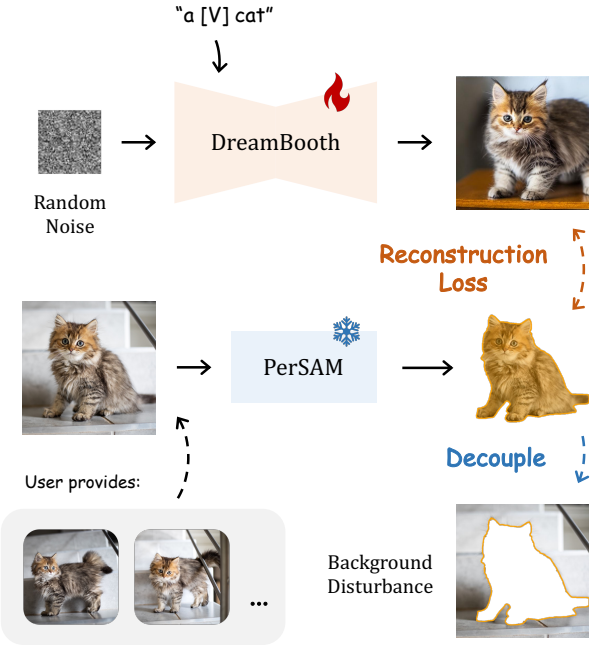

Figure 8. **Better Personalization of Stable Diffusion.** We utilize PerSAM to decouple the target objects and the background disturbance for DreamBooth [45] fine-tuning. This contributes to better personalized text-to-image generation of Stable Diffusion [44].

3.4. Better Personalization of Stable Diffusion

A Revisit of DreamBooth. Similar to personalized segmentation, Textual Inversion [12], DreamBooth [45] and follow-up works [28] fine-tune the pre-trained text-to-image models, e.g., Stable Diffusion [44] and Imagen [46], to synthesize the images of specific visual concepts indicated by users. As an example, given 3∼5 ground-truth photos of a cat, DreamBooth conducts few-shot training and learns to generate that cat by taking as input a textual prompt, "a [V] cat". Therein, [V] serves as a unique identifier to represent the specific cat in the word embedding space. After training, the personalized DreamBooth is able to synthesize novel renditions of the cat in different contexts, such as "a [V] cat on a beach." or "a [V] cat in the Grand Canyon.". However, DreamBooth calculates an L2 loss between the entire reconstructed images and ground-truth photos. As shown in Figure 3, this would inject the redundant background information of the few-shot images into the identifier [V], which overrides the newly generated background and disturb the representation learning of the target object.

PerSAM-assisted DreamBooth. In Figure 8, we introduce a strategy to alleviate the background disturbance in DreamBooth. If the user additionally provides an object mask for any of the few-shot images, we could leverage our PerSAM or PerSAM-F to segment all the foreground targets, and discard the gradient back-propagation for pixels within background area. Then, the Stable Diffusion is only fine-tuned to memorize the visual appearances of the target object, and no supervision is imposed to the background for preserving its diversity. After this, the PerSAM-assisted DreamBooth not only synthesizes the subject instance with better visual correspondence, but also increases the variability to the new contexts guided by the textual prompt.

4. Experiment

We first evaluate our approach for personalized segmentation in Section 4.1, and report the results on video object segmentation in Section 4.2. Then in Section 4.3, we show the improved text-to-image generation of DreamBooth [45] aided by our background masking. Finally, we conduct ablation study to investigate the effectiveness of each our component in Section 4.4.

4.1. Personalized Evaluation

PerSeg Dataset. To test the personalization capacity, we construct a new segmentation dataset, termed as PerSeg. The raw images are collect from the training data of subject-driven diffusion models: DreamBooth [45], Textual Inversion [12], and Custom Diffusion [28]. PerSeg contains 40 objects of various categories in total, including daily neces-

Method	mIoU	Param.	Can	Barn	Clock	Cat	Back-pack	Teddy Bear	Duck Toy	Thin Bird	Red Cartoon	Robot Toy
Existing Methods												
Painter [51]	56.35	354M	19.06	3.21	42.89	94.06	88.05	93.04	33.27	20.92	98.19	64.99
Visual Prompting [2]	65.88	383M	61.23	58.55	59.23	76.60	66.67	79.75	89.93	67.35	81.03	72.37
SEEM [63]	80.50	157M	88.80	74.34	53.93	94.53	90.92	96.72	98.30	70.64	97.16	89.58
SegGPT [53]	94.26	354M	96.62	63.79	92.56	94.13	94.40	93.67	97.15	92.60	97.33	96.19
Our Approach												
PerSAM	89.32	**0**	96.17	38.91	96.19	90.70	95.39	94.64	97.31	93.73	96.96	60.56
PerSAM-F	**95.33**	2	96.72	97.50	96.10	92.27	95.52	95.19	97.31	93.96	97.11	96.67
Improvement	+6.01		+0.55	+58.59	-0.09	+1.57	+0.13	+0.55	+0.0	+0.23	+0.15	+36.11

Table 1. **Personalized Segmentation on PerSeg Dataset**. We compare the overall mIoU (%) and learnable parameters of different methods for personalization. We report the results of 10 selected objects and highlight the improvement from PerSAM to PerSAM-F in blue.

Method	J&F	J	F
with video data			
AGSS [33]	67.4	64.9	69.9
AGAME [25]	70.0	67.2	72.7
SWEM [35]	84.3	81.2	87.4
XMem [8]	87.7	84.0	91.4
without video data			
Painter [51]	34.6	28.5	40.8
SegGPT [53]	70.0	66.4	73.7
PerSAM	60.3	56.6	63.9
PerSAM-F	**71.9**	69.0	74.8

Table 2. **Video Object Segmentation (%) on DAVIS 2017 [40]**. The methods of 'with video data' are specially developed for video segmentation and trained by in-domain videos, while 'without video data' involves no video training data. We report the results of SegGPT without the ensemble strategy for fair comparison.

Variant	mIoU	Gain
Only Positive Prior	69.11	-
+ Negative Prior	72.47	+3.63
+ Post-refinement	83.91	+11.44
+ Target. Attention	85.82	+1.91
+ Target. Prompting	89.32	+3.50
+ Fine-tuning	95.33	+6.01

Table 3. **Ablation Study (%) of PerSAM and PerSAM-F**. We progressively add our proposed components on top of the baseline model in the first row. The last two rows denote the performance of PerSAM and PerSAM-F, respectively.

sities, animals, and buildings. Contextualized in different poses or scenes, each object is related with 5∼7 images with our annotated masks. As default, we regard the first image as the user-provided one-shot data and evaluate the models by the metric of mean Intersection over Union (mIoU).

Experimental Details. We adopt pre-trained SAM [27] with a ViT-H [11] image encoder as the segmentation foundation model. For PerSAM, we apply the proposed target-guided attention and target-semantic prompting to all the three transformer blocks in SAM's decoder, i.e., two regular blocks and one final block. The balance factor α in Equation 7 is simply set as 1. For PerSAM-F, we conduct one-shot training for 1,000 epochs with a batch size 1. We set the initial learning rate as 10^{-3}, and adopt the AdamW [37] optimizer with a cosine scheduler. Note that we do not apply the target-guided attention and target-semantic prompting in PerSAM-F to better reveal the effectiveness of fine-tuning. Not data augmentation is utilized during training.

Performance. In Table 1, we report the segmentation results of our approach and other existing methods on PerSeg dataset. As shown, the fine-tuned PerSAM-F achieves the best performance, and effectively enhances PerSAM on most visual concepts by +6.01% overall mIoU. Visual Prompting [2], Painter [51], and SegGPT [53] are in-context learners that segment arbitrary objects according to given prompt images. Similar to SAM, the recent SEEM [63] is a large-scale prompt-based model with stronger interactivity and compositionality. They can also be adopted for personalized segmentation by regarding the one-shot data as a prompt. Our training-free PerSAM can outperform Painter, Visual Prompting and SEEM with significant margins. Although SegGPT attains comparable results to PerSAM-F, it contains numerous parameters and is specially trained by extensive data for personalization capability. In contrast, PerSAM-F only fine-tunes 2 learnable weights to efficiently customize off-the-shelf SAM for private use. More visualizations are shown in Figure 11.

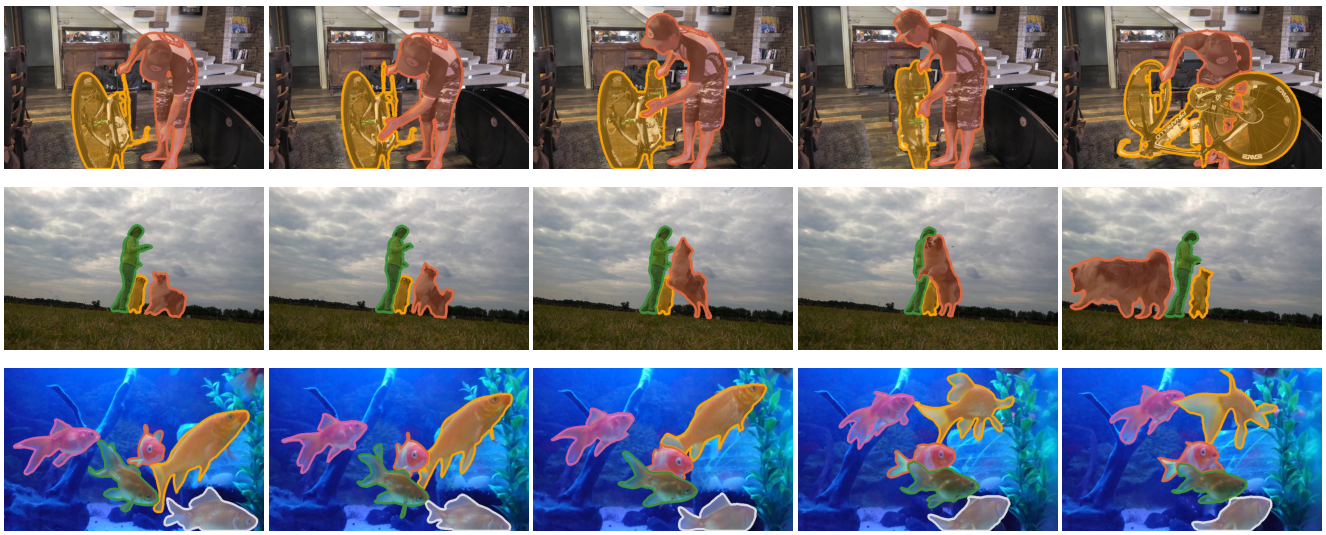

Figure 9. **Visualization of PerSAM-F on DAVIS 2017 [40] dataset.** Given the first-frame masks, our personalization approach performs well for video object tracking and segmentation. We denote the masks of different objects with different colors.

4.2. Video Object Segmentation

Experimental Details. Besides images with only one object, PerSAM and PerSAM-F can also be extended to segmenting multiple objects in video frames. Given the first frame and its object masks, our approach can be personalized to simultaneously segment and track multiple objects in the video. We select the popular DAVIS 2017 [40] dataset for evaluation, and adopt the official J and F scores as metrics. For PerSAM, we regard the top-2 highest-similarity points as the positive location prior, and additionally utilize the bounding boxes from the last frame along with their center point to prompt the decoder. This provides more sufficient temporal cues for object tracking and segmentation. For PerSAM-F, we conduct one-shot fine-tuning on the first frame for 800 epochs with a learning rate 4^{-4}. We follow the personalization experiment for other configurations.

Performance. The video segmentation results on DAVIS 2017 validation set is shown in Table 2. Compared to methods without video data, the training-free PerSAM largely surpasses Painter [51] by +25.7% J&F score, and PerSAM-F achieves +1.9% better performance than SegGPT [53] without the ensemble strategy. Notably, our fine-tuning approach can even outperform AGSS [33] and AGAME [25] by +4.5% and +1.9% J&F scores, both of which are fully trained by extensive video data. The results fully illustrate our strong generalization ability for temporal video data with multiple visual concepts. We visualize the segmentation results of PerSAM-F on three video frames in Figure 9, where our approach shows favorable performance for multi-object tracking and segmentation.

4.3. PerSAM-assisted DreamBooth

Experimental Details. We utilize the pre-trained Stable Diffusion [44] as the base text-to-image model. We follow most model hyperparameters and training configurations in DreamBooth [45], including a 10^{-6} learning rate, a batch size 1, and a 200-image regularization dataset. We fine-tune DreamBooth for 1,000 iterations within 5 minutes on a single NVIDIA A100 GPU. For better accuracy, we adopt PerSAM-F to segment target objects, which conducts one-shot fine-tuning by a given image-mask pair. Note that the training-free PerSAM also achieves similar results, and we name by 'PerSAM-assisted' just for simplicity.

Performance. In addition to Figure 3, we visualize more results of PerSAM-assisted DreamBooth in Figure 10. For the dog lying on a grey sofa, the "jungle" and "snow" by DreamBooth are still the sofa with green and white decorations. Assisted by PerSAM-F, the newly-generated background is totally decoupled with the sofa and well corresponds to the textual prompt. For the other two subjects, the background disturbance of mountains behind the barn and the couch beside the table is also alleviated. The incorrect "orange table" by DreamBooth in the last line also indicates, PerSAM-F can boost the visual appearance learning of the target for better personalizing text-to-image models.

4.4. Ablation Study

In Table 3, we investigate the effectiveness of our proposed components in PerSAM and PerSAM-F on PerSeg dataset. As shown, we first start from a baseline model with 69.11 mIoU, in which only the positive location prior

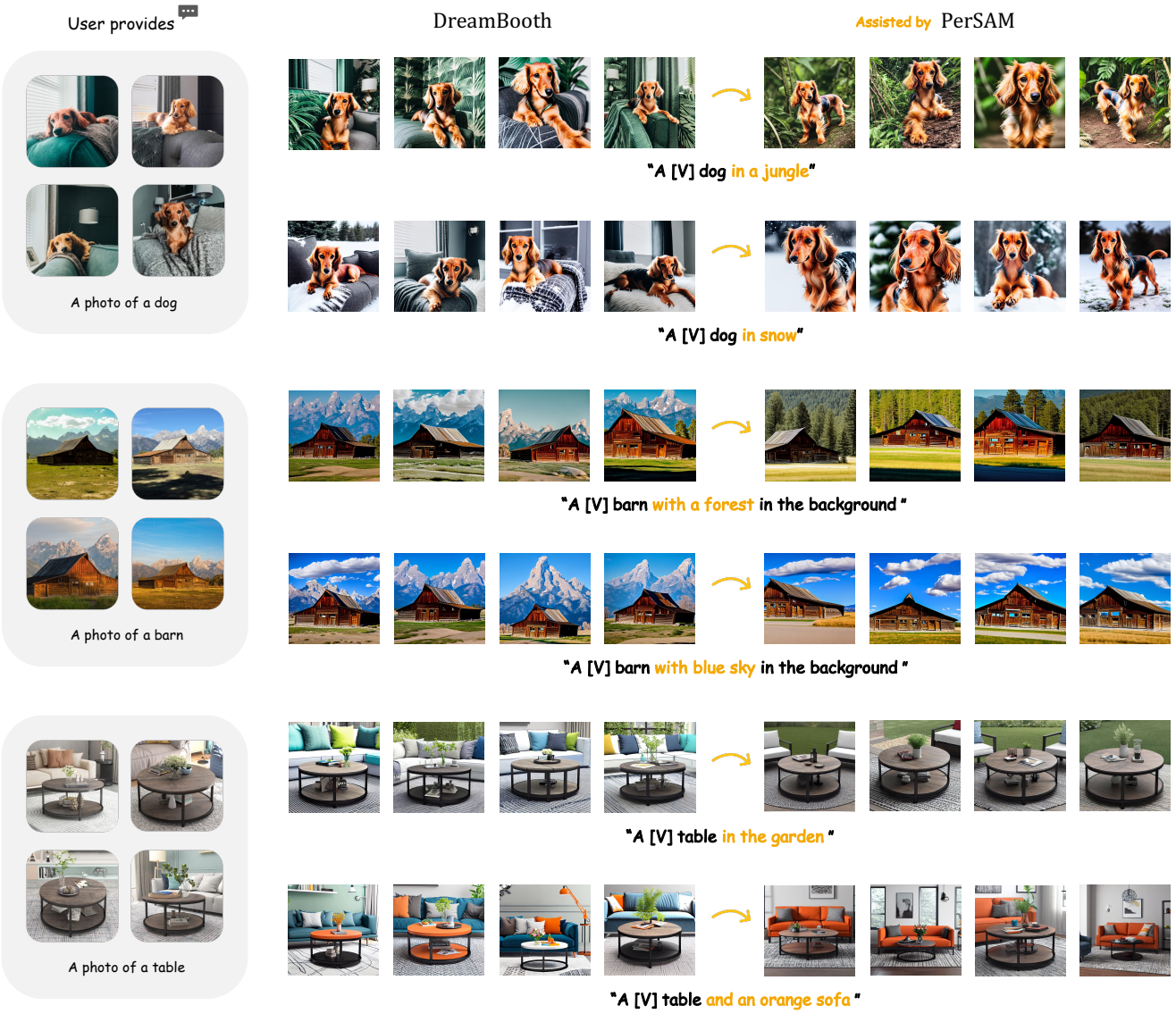

Figure 10. **Visualization of PerSAM-guided DreamBooth.** By our approach decoupling the background information, the improved DreamBooth exhibits stronger personalization capacity and preserves the diversity for synthesizing various contexts.

is utilized to automatically prompt SAM. Then, we respectively add the negative location prior and cascaded post-refinement, enhancing the segmentation accuracy by +3.63% and +11.44%, respectively. This constructs a competitive model with 83.91% mIoU, already stronger than the well pre-trained Painter [51] and SEEM [63]. On top of that, we introduce the high-level semantics of target objects into SAM's decoder to guide the cross-attention and prompting mechanisms. The +1.91% and +3.50% mIoU improvement fully indicate the significance of our designs. Finally, via the efficient one-shot fine-tuning, PerSAM-F boost the score by +6.01% and achieves 95.33% mIoU, demonstrating superior personalization capacity.

5. Discussion

What is the Difference between SegGPT and PerSAM? Painter [51] and the follow-up SegGPT [53] both adopt an in-context learning framework, which redefines the traditional segmentation task into an image coloring problem. Given one-shot prompt, they can also achieve personalized segmentation similar to PerSAM, as compared in Table 1. However, they contain 354M learnable parameters and unify a diverse set of segmentation data for large-scale training. In contrast, our approach is either training-free, or fine-tuning only 2 parameters within 10 seconds. We aim at a more efficient way to customize an off-the-shelf foundation model, i.e., SAM, into private use at the least cost.

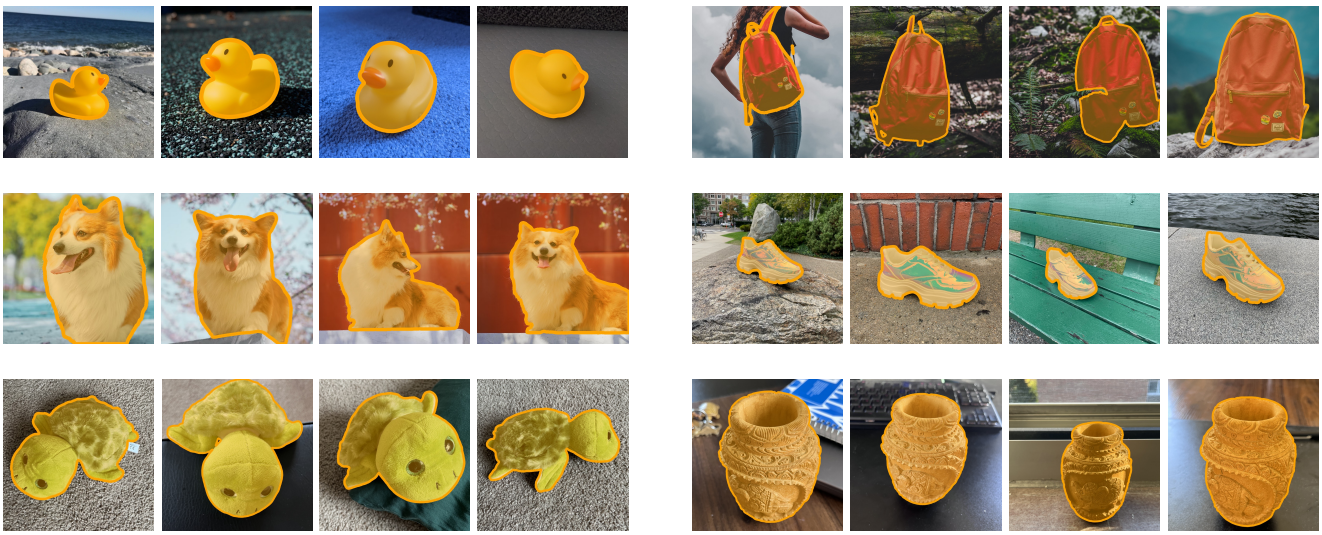

Figure 11. **Visualization of PerSAM-F on PerSeg dataset.** With only one-shot user-provided data, our personalization approach can effectively customize SAM [27] to segment specific visual concepts in any poses or scenes.

Can PerSAM Tackle Multi-object Scenarios? Yes. As visualized in Figure 9, the video object segmentation task of Table 2 requires to segment and track more than one object across the frames, e.g., a man and his bike. For multiple visual concepts, we respectively encode and store their target embeddings within the first frame. Then, for subsequent frames, we only run the image encoder once to extract the visual feature, and independently prompt the mask decoder for different objects. In this way, our PerSAM and PerSAM-F can be efficiently personalized to segment multiple visual concepts the user designates.

Robustness to Quality of the One-shot Mask? For more robust interactivity with humans, we investigate how PerSAM and PerSAM-F perform when the given one-shot mask is of low quality. In Table 4, we respectively shrink and enlarge the area of the reference mask and compare the segmentation results on PerSeg dataset. When the mask is smaller than the size of target object (shrink), the fine-tuned PerSAM-F exhibits stronger robustness to SegGPT and PerSAM. This is because the internal points around the object center can not comprehensively represent all its visual characters, which harms the obtained target embedding, weakening the effectiveness of target-guided attention and target-semantic prompting. When the mask is larger than the object (enlarge), the inaccurate mask size would mislead the one-shot training of PerSAM-F. Instead, despite of some background noises, the target embedding can incorporate complete visual appearances of the object, which brings little influence to the training-free techniques in PerSAM. Overall, our PerSAM-F indicates better robustness to quality of the given mask than SegGPT.

Method	Shrink↓↓	Shrink↓	Enlarge↑	Enlarge↑↑
SegGPT [53]	80.39	81.79	83.22	76.43
PerSAM	78.48	81.10	**89.32**	**88.92**
PerSAM-F	**85.16**	**88.28**	83.19	81.19

Table 4. **Robustness to Quality of the One-shot Mask**. We respectively shrink and enlarge the area of the reference mask using the erode and dilate functions in OpenCV [3]. We evaluate on PerSeg dataset and adopt kernel sizes of two levels, 75 and 95.

6. Conclusion

In this paper, we propose to personalize Segment Anything Model (SAM) for specific visual concepts with only one-shot data. Firstly, we introduce a training-free approach, PerSAM, which calculates a location prior on the test image, and adopts three personalization techniques: target-guided attention, target-semantic prompting, and cascaded post-refinement. On top of this, we further present a 10-second fine-tuning variant, PerSAM-F. By only 2 learnable parameters, PerSAM-F effectively alleviates the ambiguity of mask scales and achieves leading performance on our annotated PerSeg dataset. Besides, we also evaluate our approach on video object segmentation, and verify its efficacy to assist DreamBooth in fine-tuning text-to-image diffusion models. We hope our work can motivate future works for personalizing segmentation foundation models by parameter-efficient methods.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[2] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

[3] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[6] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7345–7354, 2021.

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[9] Pedro Cuenca and Sayak Paul. Using lora for efficient stable diffusion fine-tuning. https://huggingface.co/blog/lora, January 2023.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[13] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. *AAAI 2023 Best Student Paper*, 2022.

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.

[15] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022.

[16] Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1551–1560, 2021.

[17] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2022.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

[23] Zhengkai Jiang, Zhangxuan Gu, Jinlong Peng, Hang Zhou, Liang Liu, Yabiao Wang, Ying Tai, Chengjie Wang, and Liqing Zhang. Stc: spatio-temporal contrastive learning for video instance segmentation. In *European Conference on Computer Vision Workshops*, pages 539–556. Springer, 2023.

[24] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 36–54. Springer, 2022.

[25] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8953–8962, 2019.

[26] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9404–9413, 2019.

[27] 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.

[28] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[30] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. *arXiv preprint arXiv:2211.09808*, 2022.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[32] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019.

[33] Huaijia Lin, Xiaojuan Qi, and Jiaya Jia. Agss-vos: Attention guided single-shot video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3949–3957, 2019.

[34] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022.

[35] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1362–1372, 2022.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[38] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019.

[39] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[40] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[42] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[44] 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*, pages 10684–10695, 2022.

[45] 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. *arXiv preprint arXiv:2208.12242*, 2022.

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[47] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Xiangyu Zhang, Hongbin Sun, Jian Sun, and Nanning Zheng. Rethinking learnable tree filter for generic feature transform. *Advances in Neural Information Processing Systems*, 33:3991–4002, 2020.

[48] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.

[49] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European Conference on Computer Vision*, pages 282–298. Springer, 2020.

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[51] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. *arXiv preprint arXiv:2212.02499*, 2022.

[52] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

[53] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[54] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and

efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[55] Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3056–3064, 2021.

[56] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[57] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022.

[58] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[59] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023.

[60] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[61] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[63] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.