PERSONALIZE SEGMENT ANYTHING MODEL WITH ONE SHOT

Anonymous authors
Paper under double-blind review

A OVERVIEW

- Section B: Related work.
- Section C: Experimental details and visualization.
- Section D: Additional experiments and analysis.
- Section E: Additional discussion.

B RELATED WORK

Foundation Models. With powerful generalization capacity, pre-trained foundation models can be adapted for various downstream scenarios and attain promising performance. In natural language processing, BERT (Devlin et al., 2018; Lu et al., 2019), GPT series (Brown et al., 2020; OpenAI, 2023; Radford & Narasimhan, 2018; Radford et al., 2019), and LLaMA (Zhang et al., 2023c) have demonstrated remarkable in-context learning abilities, and can be transferred to new tasks by domain-specific prompts. Similarly, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), which conduct contrastive learning on image-text pairs, exhibit exceptional accuracy in zero-shot visual recognition. Painter (Wang et al., 2022) introduces a vision model that unifies network architectures and in-context prompts to accomplish diverse vision tasks, without downstream fine-tuning. CaFo (Zhang et al., 2023d) cascades different foundation models and collaborates their pre-trained knowledge for robust low-data image classification. SAM (Kirillov et al., 2023) presents a foundation model for image segmentation, which is pre-trained by 1 billion masks and conducts prompt-based segmentation. There are some concurrent works extending SAM for high-quality segmentation (Ke et al., 2023), faster inference speed (Zhao et al., 2023; Zhang et al., 2023a), all-purpose matching (Liu et al., 2023b), 3D reconstruction (Cen et al., 2023), object tracking (Yang et al., 2023), medical (Ma & Wang, 2023; Huang et al., 2023) image processing. From another perspective, we propose to personalize the segmentation foundation 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 (Rombach et al., 2022) and Imagen (Saharia et al., 2022), which improves the generation quality by segmenting the foreground target objects from the background disturbance.

Large Models in Segmentation. As a fundamental task in computer vision, segmentation (Long et al., 2015; Jiang et al., 2022; Zhao et al., 2017; Xu et al., 2021; Jiang et al., 2023; Lin et al., 2022) requires a pixel-level comprehension of a image. Various segmentation-related tasks have been explored, such as semantic segmentation, classifying each pixel into a predefined set of classes (Badrinarayanan et al., 2017; Chen et al., 2017; Zheng et al., 2021; Cheng et al., 2022; Xie et al., 2021; Song et al., 2020b); instance segmentation, focusing on the identification of individual object instances (He et al., 2017; Wang et al., 2020; Tian et al., 2020a); panoptic segmentation, assigning both class labels and instance identification (Kirillov et al., 2019; Li et al., 2019); and interactive segmentation, involving human intervention for refinement (Hao et al., 2021; Chen et al., 2021). Recently, inspired by language foundation models (Zhang et al., 2023c; Brown et al., 2020), several concurrent works have proposed large-scale vision models for image segmentation. They are pre-trained by extensive mask data and exhibit strong generalization capabilities on numerous image distributions. Segment

Table 1: **Personalized Object Segmentation on the PerSeg Dataset**. We report the mIoU scores of 30 objects in addition to the 10 objects in Table 1 of the main paper. '*' denotes works concurrent to ours.

Method	Dog	Dog2	Dog3	Dog4	Dog5	Dog6	Tortoise Plushy	Round Bird	Colorful Sneaker	Colorful Teapot
Painter (Wang et al., 2022)	80.41	73.77	46.98	22.39	82.03	76.16	55.31	39.83	0.00	13.69
VP (Bar et al., 2022)	6.80	12.26	23.84	20.61	21.44	32.05	24.42	34.09	30.32	34.89
SEEM* (Zou et al., 2023)	71.04	35.35	67.02	81.87	75.02	72.99	78.75	38.74	20.08	44.44
SegGPT* (Wang et al., 2023)	73.82	65.07	61.11	81.66	82.94	76.44	77.85	82.89	72,24	80.44
PerSAM	96.79	95.66	88.85	95.22	97.10	94.66	93.06	96.79	94.48	96.27
PerSAM-F	96.81	95.79	88.67	95.18	97.22	94.85	97.09	96.85	95.13	84.41

Method	Dog7	Dog8	Candle	Fancy Boot	Sloth Plushie	Poop Emoji	Rc Car	Shiny Sneaker	Wolf Plushie	Wooden Pot
Painter (Wang et al., 2022)	40.97	57.15	24.36	49.06	45.78	23.42	23.69	0.00	38.97	57.61
VP (Bar et al., 2022)	17.67	12.24	12.71	39.13	29.31	37.55	29.98	30.88	28.86	34.30
SEEM* (Zou et al., 2023)	63.77	70.34	26.99	34.90	81.46	45.55	34.94	82.30	76.27	74.81
SegGPT* (Wang et al., 2023)	66.20	82.21	81.60	76.06	80.54	81.32	79.26	85.26	72.48	78.00
PerSAM	93.69	95.34	74.16	95.87	96.37	96.01	39.30	97.00	94.34	97.42
PerSAM-F	93.77	95.61	96.75	95.96	96.64	96.43	96.12	96.87	94.32	97.43

Method	Table	Teapot	Chair	Elephant	Duck Toy	Monster Toy	Dog Pack	Bear Plushie	Berry Bowl	Cat Statue
Painter (Wang et al., 2022)	16.92	7.00	50.09	40.80	29.24	34.80	40.73	81.30	45.98	19.96
VP (Bar et al., 2022)	16.00	10.00	27.20	22.01	52.14	30.92	22.80	23.95	11.32	27.54
SEEM* (Zou et al., 2023)	30.15	12.30	66.15	46.64	89.92	41.49	66.83	61.27	38.29	24.27
SegGPT* (Wang et al., 2023)	81.95	89.89	78.97	80.38	84.48	83.33	77.53	75.54	73.00	76.54
PerSAM	94.68	40.02	92.22	96.05	97.31	93.75	95.85	89.28	91.81	95.42
PerSAM-F	94.66	96.93	92.14	96.07	97.31	94.21	95.76	95.32	91.27	95.46

Anything Model (SAM) (Kirillov et al., 2023) utilizes a data engine with model-in-the-loop annotation to learn a promptable segmentation framework, which generalizes to downstream scenarios in a zero-shot manner. Painter (Wang et al., 2022) and SegGPT (Wang et al., 2023) introduce a robust in-context learning paradigm and can segment any images by a given image-mask prompt. SEEM (Zou et al., 2023) further presents a general segmentation model prompted by multi-modal references, e.g., language and audio, incorporating versatile semantic knowledge. In this study, we introduce a new task termed personalized object segmentation, and annotate a new dataset PerSeg for evaluation. Instead of developing large segmentation models, our goal is to personalize them to segment user-provided objects in any poses or scenes. We propose two approaches, PerSAM and PerSAM-F, which efficiently customize SAM for personalized segmentation.

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 (Sung et al., 2022; He et al., 2022; Rebuffi et al., 2017; Qin & Eisner, 2021) to freeze the weights of foundation models and append small-scale modules for fine-tuning. Prompt tuning (Lester et al., 2021; Zhou et al., 2022; Jia et al., 2022; Liu et al., 2021) 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) (Hu et al., 2021; Cuenca & Paul, 2023; Zhang et al., 2023b; Hedegaard et al., 2022) injects trainable rank decomposition matrices concurrently to each pre-trained weight, which significantly reduces the number of learnable parameters required for downstream tasks. Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020; Lin et al., 2020; Chen et al., 2022) are designed to be inserted between layers of the original transformer, introducing lightweight MLPs for feature transformation. Different from existing works, we adopt a more efficient adaption method delicately designed for SAM, i.e., the scale-aware fine-tuning of PerSAM-F with only 2 parameters and 10 seconds. This effectively avoids the over-fitting issue on one-shot data, and alleviates the ambiguity of segmentation scale with superior performance.

C Experimental Details and Visualization

C.1 Personalized Evaluation

Implementation Details. We adopt a pre-trained SAM (Kirillov et al., 2023) with a ViT-H (Dosovitskiy et al., 2020) backbone as the foundation model, and utilize SAM's encoder to calculate the location confidence map. For PerSAM, we apply the target-guided attention and target-semantic prompting to all three blocks in the decoder. The balance factor α in Equation 8 of the main paper is set as 1. For PerSAM-F, we conduct one-shot training for 1,000 epochs with a batch size 1, supervised by the dice loss (Milletari et al., 2016) and focal loss (Lin et al., 2017). We set the initial learning rate as 10^{-3}, and adopt an AdamW (Loshchilov & Hutter, 2017) optimizer with a cosine scheduler.

Complete Results on the PerSeg Dataset. In Table 1, we report the mIoU scores of the other 30 objects in the PerSeg dataset, except for the 10 objects in Table (1) of the main paper. As compared, our PerSAM without any training can achieve superior segmentation results to Painter (Wang et al., 2022), Visual Prompting (VP) (Bar et al., 2022), and SEEM (Zou et al., 2023) on most objects. Note that, we here compare the results of SEEM with the Focal-L (Yang et al., 2022) vision backbone, its best-performing variant. Aided by the 2-parameter fine-tuning, PerSAM-F further performs comparably with SegGPT (Wang et al., 2023), a powerful in-context segmentation framework. Therefore, our approach exhibits a high performance-efficiency trade-off by efficiently customizing the off-the-shelf SAM (Kirillov et al., 2023) for personalized object segmentation.

Visualization. In Figure 1, we visualize the location confidence maps, segmentation results of PerSAM with positive-negative location prior, and the bounding boxes from the cascaded post-refinement. As shown, the confidence map (hotter colors indicate higher scores) can clearly indicate the rough region of the target object in the image, which contributes to precise foreground (green pentagram) and background (red pentagram) point prompts selection. The bounding boxes in green also well enclose the targets and prompt SAM's decoder for accurate post-refinement.

C.2 Existing Segmentation Benchmarks

Implementation Details. For experiments in existing segmentation datasets, we utilize DINOv2 (Oquab et al., 2023) as the image encoder to calculate the location confidence map, which produces a more accurate location prior. Note that, the generality and extensibility of our approach enable us to apply any vision backbones for location confidence map calculation. For video object segmentation, different from PerSeg, where one image contains only one object, DAVIS 2017 dataset (Pont-Tuset et al., 2017) requires to personalize SAM to track and segment multiple different objects across the video frames. In PerSAM, we regard the top-2 highest-confidence points as the positive location prior, and additionally utilize the bounding boxes from the last frame to prompt the decoder. This provides more sufficient temporal cues for object tracking and segmentation. In PerSAM-F, we conduct one-shot fine-tuning on the first frame for 800 epochs with a learning rate 4^{-4}. As discussed in Section 2.5 of the main paper, for N objects, we only need to run SAM's large-scale encoder (2s) once to encode the visual feature of the new frame, while running the lightweight decoder for N times to segment different objects, which takes marginal $50N$ms. For one-shot semantic segmentation, we evaluate our method on FSS-1000 (Li et al., 2020) following HSNet (Min et al., 2021) and LVIS-92^i (Gupta et al., 2019) pre-processed by (Liu et al., 2023b). The benchmark contains objects in a wide range of semantic categories within various backgrounds. For one-shot part segmentation, we utilize the part-level benchmarks of PASCAL VOC (Morabia et al., 2020) and PACO (Ramanathan et al., 2023) built by (Liu et al., 2023b), requiring to segment partial objects with challenging scenarios.

Visualization. In Figure 2, we visualize more results of PerSAM-F for multi-object tracking and segmentation in consecutive frames of the DAVIS 2017 dataset. We utilize different colors to denote different objects, along with the additional prompts for SAM's decoder, including a bounding box from the last frame and its center point. Aided by our techniques and the last-frame temporal cues, PerSAM-F exhibits favorable video segmentation performance and tracking consistency, even for objects occluded by others or objects of the same category with similar appearances. In Figure 3, we also visualize the results of PerSAM-F for one-shot semantic and part segmentation on four

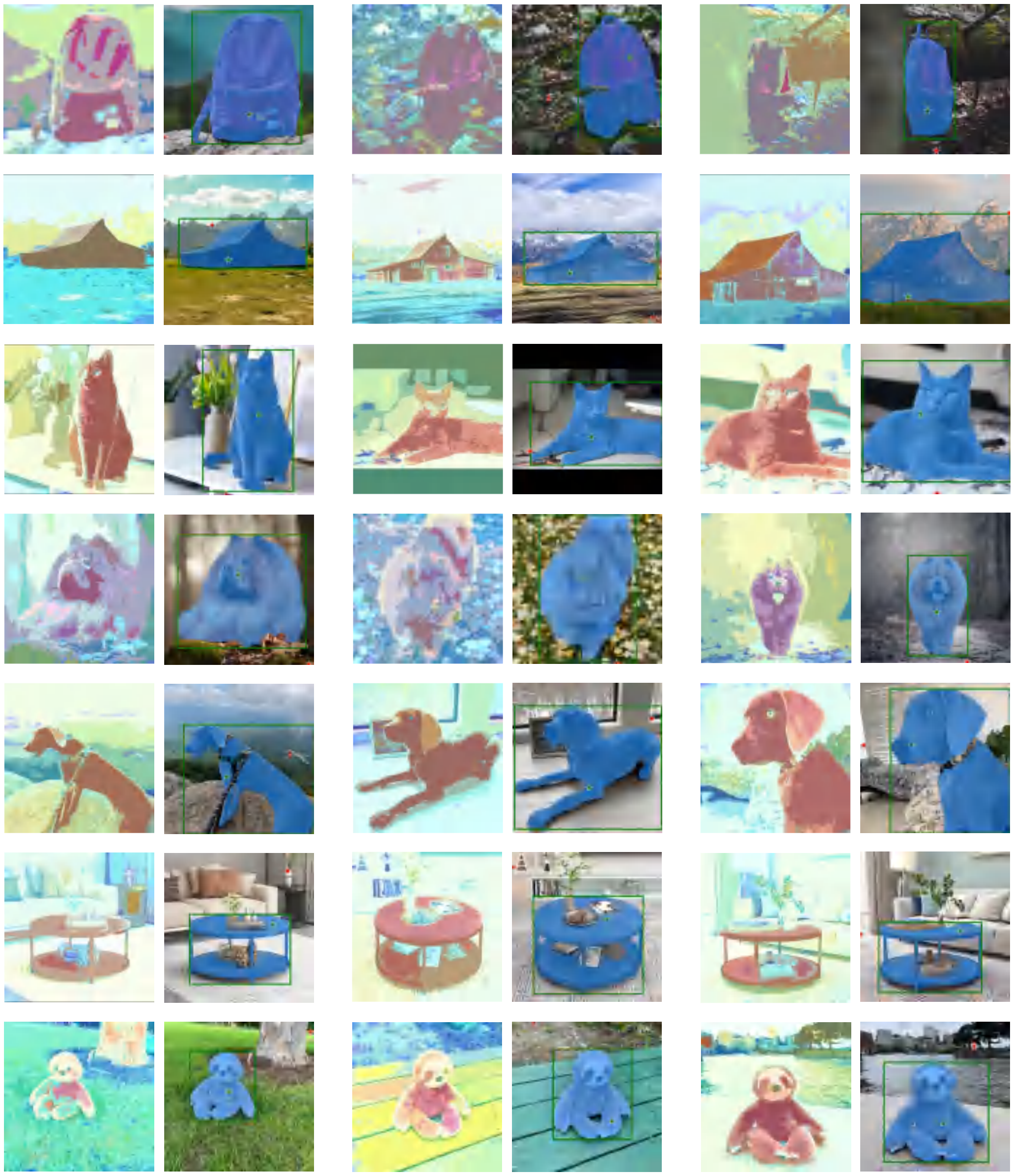

Figure 1: **Visualization of Location Confidence Maps and PerSAM's Segmentation Results.** We represent the positive (foreground) and negative (background) location prior by green and red pentagrams. The green bounding boxes denote the box prompts in the cascaded post-refinement.

datasets. The satisfactory performance illustrates that our approach is not limited to object-level personalization, but also part- and category-wise segmentation with good generalization capability.

C.3 PerSAM-assisted DreamBooth

Implementation Details. We follow most model hyperparameters and training configurations in DreamBooth (Ruiz et al., 2022), including a 10^{-6} learning rate and a batch size 1. We generate a 200-image regularization dataset by the pre-trained Stable Diffusion (Rombach et al., 2022) using the textual prompt: "photo of a [CLASS]". We fine-tune DreamBooth and our approach both for 1,000 iterations on a single A100 GPU, and adopt 't@y' as the word identifier [V] for the personal visual concepts. We utilize DDIM (Song et al., 2020a) sampling with 100 steps and a 10-scale classifier-free guidance for generation.

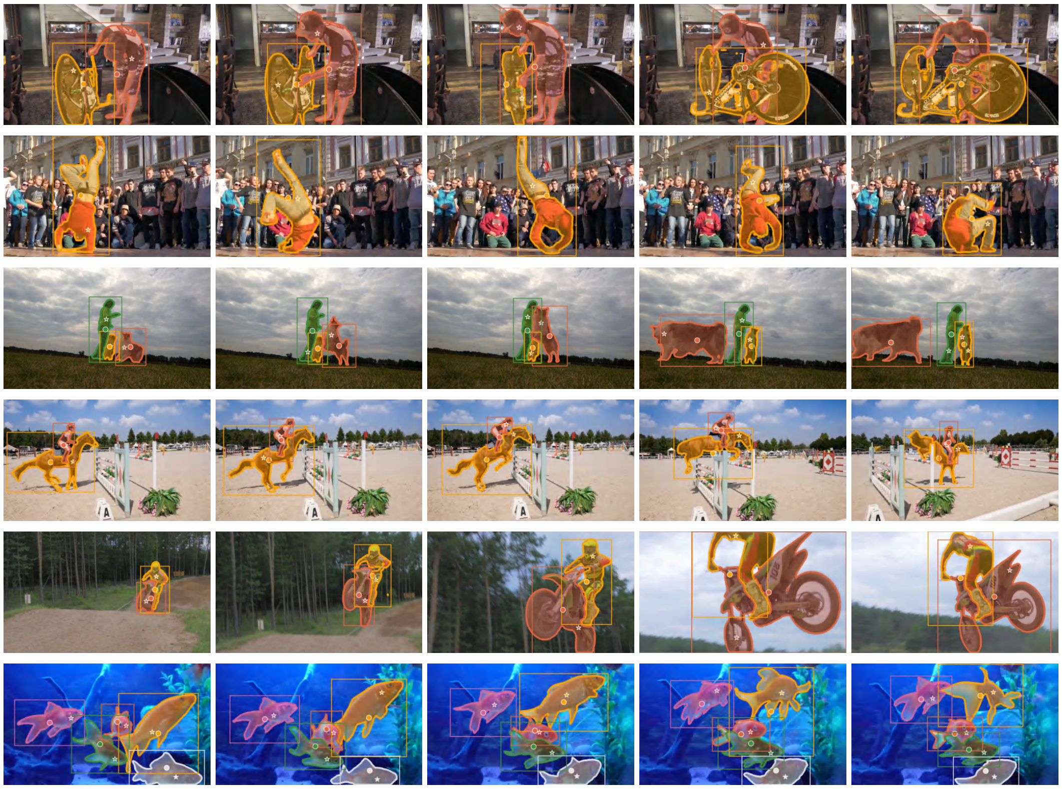

Figure 2: **Visualization of PerSAM-F for Video Object Segmentation** on the DAVIS 2017 (Pont-Tuset et al., 2017) dataset. We represent different objects in different colors, and visualize the input prompts for SAM's decoder: a positive location prior (pentagram), an enclosing bounding box from the last frame, and its center point (dot).

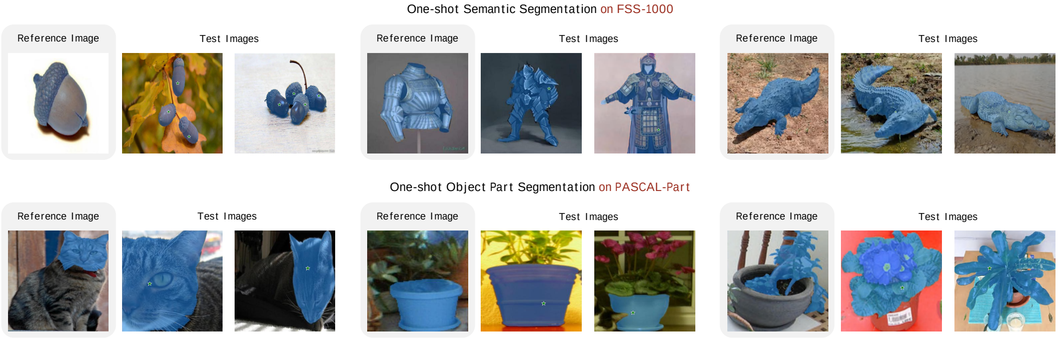

Figure 3: **Visualization of PerSAM-F for One-shot Semantic and Part Segmentation** on FSS-1000 (Li et al., 2020) and PASCAL-Part (Morabia et al., 2020) datasets. Our approach exhibits superior generalization capabilities for diverse segmentation scenarios.

Quantitative Evaluation. Besides visualization, we also evaluate the PerSAM-assisted DreamBooth by three quantitative metrics in Table 13. We leverage CLIP (Radford et al., 2021) to calculate the feature similarity of generated images with textual prompts ('Text-Align') and reference images ('Image-Align') (Kumari et al., 2022), along with KID (Bińkowski et al., 2018) (the smaller, the better). 'Text-Align' (Gal et al., 2022) and 'Image-Align' (Hessel et al., 2021) indicate the semantic correspondence of the synthesized images with the textual prompt and few-shot reference images, respectively. KID (Bińkowski et al., 2018) measures how much the fine-tuned models over-fit the specific visual concepts in few-shot images, for which we utilize Stable Diffusion to generate 500

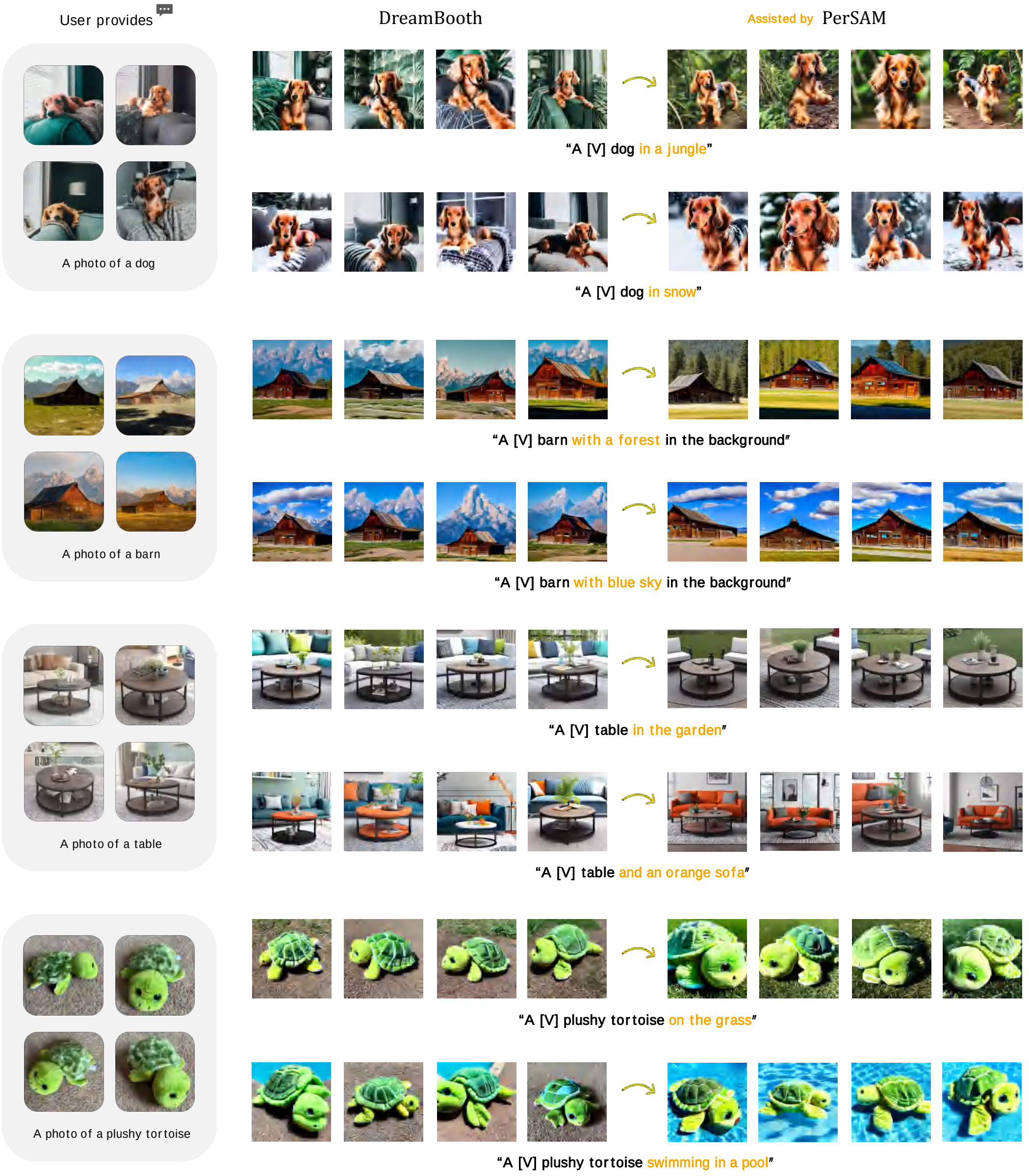
User provides
DreamBooth
Assisted by PerSAM
A photo of a dog
"A [V] dog in a jungle"
"A [V] dog in snow"
A photo of a barn
"A [V] barn with a forest in the background"
"A [V] barn with blue sky in the background"
A photo of a table
"A [V] table in the garden"
"A [V] table and an orange sofa"
A photo of a plushy tortoise
"A [V] plushy tortoise on the grass"
"A [V] plushy tortoise swimming in a pool"

Figure 4: **Visualization of PerSAM-assisted DreamBooth.** Our approach can alleviate the background disturbance, and boost DreamBooth (Ruiz et al., 2022) for better personalized synthesis.

images as the validation set. These quantitative results demonstrate our effectiveness in generating better visual correspondence with the target objects and input prompts.

Visualization. In Figure 4, we visualize more examples that demonstrate our effectiveness to enhance DreamBooth for higher-fidelity personalized synthesis. We utilize PerSAM-F to decouple the table and plushy tortoise from their backgrounds in the few-shot images, i.e., the couch and carpet. In this way, the PerSAM-assisted DreamBooth generates new backgrounds corresponding to the textual prompts of "in the garden", "and an orange sofa", "on the grass", and "swimming in a pool". In addition, our approach can boost the appearance generation of target objects with high text-image correspondence, while the vanilla DreamBooth might be interfered by textual prompts, e.g., the orange on the table and the blue on the turtle shell. The experiments fully verify our efficacy for better personalizing text-to-image models.

Table 2: **One-shot segmentation on COCO-20^i (Nguyen & Todorovic, 2019).**

Method	In-domain Train	mIoU
FPTrans	✓	47.0
SCCAN	✓	48.2
HDMNet	✓	50.0
PerSAM	-	47.9
PerSAM-F	-	50.6

Table 3: **One-shot segmentation on Tokyo Multi-Spectral-4^i (Bao et al., 2021).**

Method	In-domain Train	mIoU
PFENet	✓	14.0
PGNet	✓	17.5
V-TFSS	✓	26.1
PerSAM	-	18.4
PerSAM-F	-	25.6

Table 4: **Comparison with two text-guided models: OVSeg (Liang et al., 2023) and Grounded-SAM (gro, 2023).**

Method	Prompt	PerSeg	COCO-20^i
OVSeg	Category Name	76.5	37.8
Grounded-SAM	Category Name	93.2	51.3
PerSAM	One-shot Data	89.3	47.9
PerSAM-F	One-shot Data	95.3	50.6

Table 5: **Running Efficiency compared to SAM (Kirillov et al., 2023).**

Method	FPS↑	Memory (MB)↓
SAM	2.16	5731
PerSAM	2.08	5788
PerSAM-F	1.98	5832

Table 6: **Comparison with SAM-PT (Rajič et al., 2023) on DAVIS 2017 (Pont-Tuset et al., 2017).**

Method	Propagation	J&F
SAM-PT	Point Tracking	76.6
PerSAM	Feature Matching	66.9
PerSAM	+Point Tracking	68.2
PerSAM-F	Feature Matching	76.1
PerSAM-F	+Point Tracking	77.2

D Additional Experiments and Analysis

D.1 Evaluation on Additional Benchmarks

COCO-20^i (Nguyen & Todorovic, 2019). Constructed from MSCOCO (Lin et al., 2014), COCO-20^i divides the diverse 80 classes evenly into 4 folds for one-shot semantic segmentation. We directly test our method on the validation set without specific in-domain training. As shown in Table 2, our PerSAM(-F) achieves favorable segmentation performance over a wide range of object categories, comparable to previous in-domain methods, i.e., FPTrans (Zhang et al., 2022), SCCAN (Xu et al., 2023), and HDMNet (Peng et al., 2023).

Tokyo Multi-Spectral-4^i (Bao et al., 2021). Sampled from Tokyo Multi-Spectral (Ha et al., 2017), Tokyo Multi-Spectral-4^i contains 16 classes within outdoor city scenes, similar to CityScapes (Cordts et al., 2016). Different from existing methods, we only take as input the RGB images without the paired thermal data, and do not conduct in-domain training. As shown in Table 3, our approach still exhibits good generalization capacity in street scenarios, compared to the specialist models: PFENet (Tian et al., 2020b), PGNet (Zhang et al., 2019), and V-TFSS (Bao et al., 2021).

Table 7: **Few-shot segmentation on the PerSeg dataset.**

Method	Shot	mIoU	bIoU
SegGPT	1-shot	94.3	76.5
SegGPT	3-shot	96.7	78.4
PerSAM	1-shot	89.3	71.7
PerSAM	3-shot	90.2	73.6
PerSAM-F	1-shot	95.3	77.9
PerSAM-F	3-shot	97.4	79.1

Table 8: **Few-shot segmentation on FSS-1000 (Li et al., 2020) benchmark.**

Method	Shot	mIoU
SegGPT	1-shot	85.6
SegGPT	5-shot	89.3
PerSAM	1-shot	81.6
PerSAM	5-shot	82.3
PerSAM-F	1-shot	86.3
PerSAM-F	5-shot	89.8

D.2 Comparison to Additional Methods

Text-guided Segmenters. Recently, open-world segmentation models guided by text prompts have driven increasing attention. To compare our approach with them, we select two popular methods: OVSeg (Liang et al., 2023) and Grounded-SAM (gro, 2023). OVSeg leverages MaskFormer (Cheng et al., 2021) to first generate class-agnostic mask proposals, and then adopts a fine-tuned CLIP for zero-shot classification. Grounded-SAM utilizes a powerful text-guided detector, Grounding DINO (Liu et al., 2023a), to generate object bounding boxes, and then utilize them to prompt SAM for segmentation. Instead of giving a one-shot reference, we directly prompt them by the category name of the target object for text-guided segmentation, e.g., “cat”, “dog”, or “chair”. As shown in Table 4, our PerSAM-F consistently achieves competitive results on two different datasets: PerSeg and COCO-20^i. This indicates that, utilizing PerSAM with a class-agnostic one-shot reference is on par with recognizing the category and then segmenting it with text-guided methods.

SAM-PT (Rajič et al., 2023). Although both our PerSAM(-F) and the concurrent SAM-PT are developed based on SAM, our approach can be generalized to most one-shot segmentation tasks (personalized/video/semantic/part segmentation), while SAM-PT specifically aims at video object segmentation. One key difference between our approach and SAM-PT is how to locate and associate objects from the previous to the current frame, i.e., propagating the location prompt for SAM across frames. In detail, our PerSAM(-F) simply calculates a location confidence map by feature matching, while SAM-PT relies on an external point tracking network, PIPS (Harley et al., 2022). As shown in Table 6, on DAVIS 2017 dataset (Pont-Tuset et al., 2017), SAM-PT performs slightly better than the original PerSAM-F. However, inspired by SAM-PT, we can also incorporate its point tracking strategy (the PIPS tracker) with PerSAM(-F) to propagate the positive-negative point prompt, which effectively enhances the segmentation performance. This demonstrates the flexible extensibility of our approach for applying more advanced trackers in a plug-and-play way.

D.3 Few-shot Segmentation by PerSAM

Our approach is not limited to one-shot segmentation, and can accept few-shot references for improved results. As an example, given 3-shot references, we independently calculate 3 location confidence maps for the test image, and adopt a pixel-wise max pooling to obtain the overall location estimation. For PerSAM-F, we regard all 3-shot data as the training set to conduct the scale-aware fine-tuning.

We respectively conduct experiments for 3-shot segmentation on PerSeg dataset and 5-shot segmentation on FSS-1000 dataset (Li et al., 2020). The results are respectively shown in Tables 7 and 8. By providing more visual semantics in few-shot data, both our training-free PerSAM and the fine-tuned PerSAM-F can be further enhanced.

Table 9: **Different pre-trained encoders for obtaining the positive-negative location prior.**

Method	Encoder	DAVIS 2017	FSS-1000	LVIS-92^i	PASCAL-Part	PACO-Part
Painter	-	34.6	61.7	10.5	30.4	14.1
SegGPT	-	75.6	85.6	18.6	-	-
PerSAM	SAM	62.8	74.9	12.9	31.3	21.2
PerSAM	DINOv2	66.9	81.6	15.6	32.5	22.5
PerSAM-F	SAM	73.4	79.4	16.2	32.0	21.3
PerSAM-F	DINOv2	76.1	86.3	18.4	32.9	22.7

Table 10: **Different Image Encoders** of SAM for PerSAM and PerSAM-F.

Method	Encoder	mIoU	bIoU
PerSAM	ViT-B	63.98	49.30
	ViT-L	86.61	69.86
	ViT-H	**89.32**	**71.67**
PerSAM-F	ViT-B	87.24	69.36
	ViT-L	92.24	75.36
	ViT-H	**95.33**	**77.92**

Table 11: **Robustness to Mask Reference**. We resize the reference mask by 'erode' and 'dilate' functions in OpenCV (Bradski, 2000).

Method	Shrink↑↑	Shrink↑	Enlarge↑	Enlarge↑↑
SegGPT	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

D.4 Ablation Study

Different Pre-trained Encoders. For video object segmentation in Table 2 and other one-shot segmentation in Table 3 of the main paper, we adopt the DINOv2 (Oquab et al., 2023) encoder to obtain the positive-negative location prior by default. In Table 9, we show the results by using SAM's original image encoder. As DINOv2 is particularly pre-trained by large-scale contrastive data, it produces more discriminative image features than SAM's encoder. This contributes to a more precise positive-negative location prior for better segmentation results, especially on the challenging FSS-1000 dataset (Li et al., 2020). Despite this, with SAM's original encoder, our PerSAM-F and the training-free PerSAM still obtain better segmentation accuracy than Painter (Wang et al., 2022) or SEEM (Zou et al., 2023) on different datasets, demonstrating the effectiveness of our approach.

Image Encoders of SAM. By default, we adopt a pre-trained ViT-H (Dosovitskiy et al., 2020) in SAM as the image encoder for PerSAM and PerSAM-F. In Table 10, we investigate the performance of other vision backbones for our models, i.e., ViT-B and ViT-L. As shown, stronger image encoders lead to higher segmentation mIoU and bIoU scores. When using ViT-B as the encoder, the accuracy of training-free PerSAM is largely harmed, due to weaker feature encoding ability, while the one-shot training of PerSAM-F can effectively mitigate the gap by +23.26% mIoU and +20.06% bIoU scores, which demonstrates the significance of our fine-tuning on top of a weak training-free baseline.

Robustness to the Quality of Mask Reference. For more robust interactivity with humans, we explore how our approach performs if the given mask is of low quality. In Table 11, we respectively shrink and enlarge the area of the reference mask, and compare the results on PerSeg dataset. When the mask is smaller than the target object (shrink), PerSAM-F, aided by one-shot fine-tuning, exhibits the best robustness. In this case, the target embedding cannot incorporate complete visual appearances from the reference image, which largely harms the training-free techniques in PerSAM. When the mask becomes larger (enlarge), the oversize mask would mislead the scale-aware training of PerSAM-F. In contrast, despite some background noises, the target embedding can include all the visual semantics of objects, which, thereby, brings little influence to PerSAM.

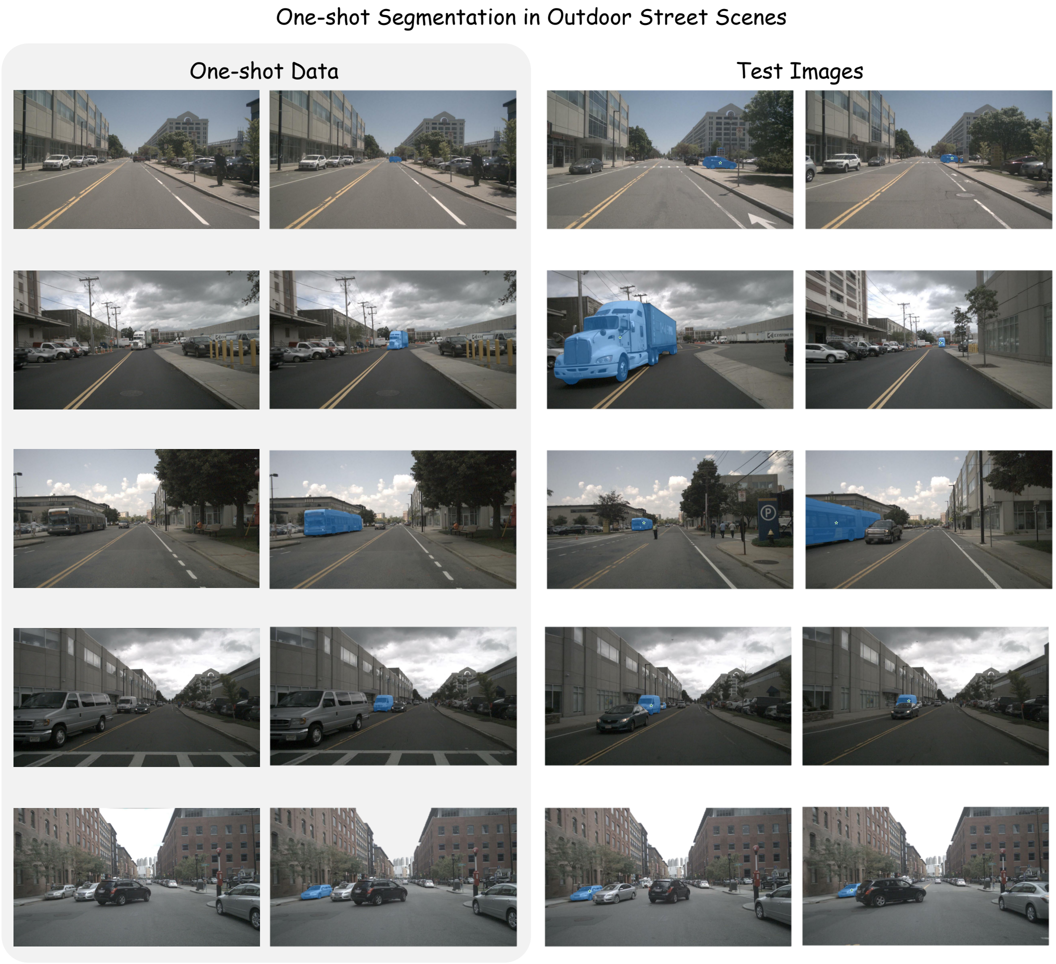

Figure 5: **One-shot segmentation of PerSAM-F in outdoor street scenes.**

E Discussion

E.1 What's the Additional Running Speed/Memory Compared to SAM?

We test the additional running consumption of PerSAM and PerSAM-F on a single NVIDIA A100 GPU with batch size 1. As shown in Table 5, our PerSAM and PerSAM-F bring marginal latency and GPU memory consumption over SAM, indicating superior running efficiency.

E.2 How to Differentiate Similar Objects in Video Object Segmentation?

For video object segmentation, our approach tries to accurately locate the target object among similar ones by the following three aspects.

Discriminative Features from the Encoder. Due to large-scale pre-training, the SAM's image encoder, or the more powerful DINOv2, can already produce discriminative visual features for different similar objects, which is fundamental to the calculation of location confidence map.

Comprehensive Location Confidence Map. We calculate a set of confidence maps for all foreground pixels within the target object, such as the head, the body, or the paws of a dog, and then

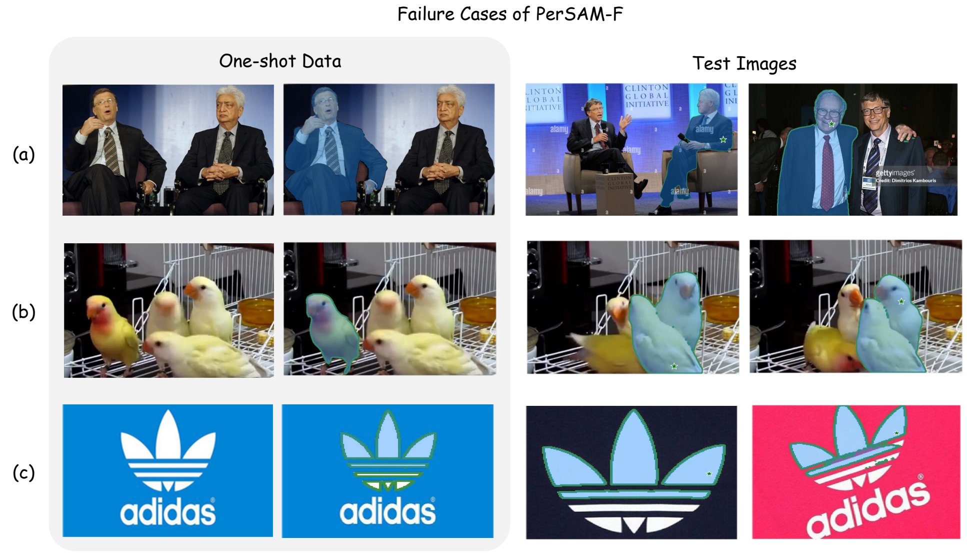

Figure 6: **Three types of failure cases of PerSAM-F.**

aggregate them to obtain an overall location estimation. This strategy can comprehensively consider the slight differences in any local parts between similar objects.

Temporal Cues between Adjacent Frames. To better leverage the temporal consistency along the video, we prompt SAM's decoder additionally with the object bounding box from the last frame. As different objects have different trajectories, such temporal constraints can better differentiate similar objects by spatial locations.

As visualized in Figures 2, our method can precisely segment the dancing man in front of a crowd (the 2^{nd} row) and differentiate different fishes within a group (the last row).

E.3 Can PerSAM also Work on Self-driving Scenarios?

Yes. In most cases, our model can segment the designated cars with distinctive appearances in dense traffic. As visualized in Figure 5, for the user-provided target (e.g., a red car, a truck, and a bus), our PerSAM-F can well locate and segment them under severe occlusion or surrounded by similar cars.

E.4 Failure Cases of PerSAM-F

After solving the scale ambiguity issue, the three types of failure cases of PerSAM-F are shown in Figure 6: (a) different people with the same clothes, indicating our approach is not very sensitive to fine-grained human faces; (b) the key appearance of the target object is occluded by in test images (the red chest of the bird), indicating that we still need to improve our robustness when there is too large appearance change in test images; (c) discontinuous objects that SAM cannot tackle, for which we can replace SAM with stronger segmentation foundation model for assistance.

E.5 Can PerSAM Segment Multiple Identical Objects in an Image?

Yes. As shown in Figure 7, given the one-shot image of a reference cat, if the test image contains two similar cats that are expected to be both segmented, we propose two simple strategies for PerSAM:

Iterative Masking. For two similar cats, we first calculate the location confidence map S_1, and utilize PerSAM to segment one of the cats, denoting the obtained mask prediction as M_1. Then, we

Table 12: **Statistic of Location Confidence Scores for Different Objects** in the PerSeg dataset.

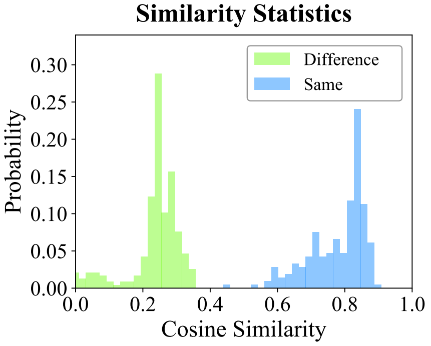

Table 13: **DreamBooth Assisted by PerSAM** with quantitative results. We adopt CLIP (Radford et al., 2021) to calculate the image-text and -image similarity.

Method	Text-Align	Image-Align	KID ($\times 10^3$)
DreamBooth	0.812	0.793	29.7
+ PerSAM	0.830	0.814	29.2
+ PerSAM-F	**0.834**	**0.818**	**28.9**

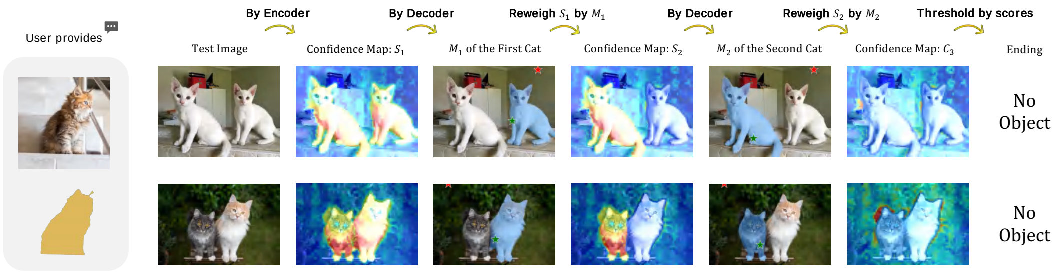

Figure 7: **Segmenting Multiple Similar Objects in an Image.** We adopt two strategies for PerSAM to simultaneously segment multiple similar objects: iterative masking and confidence thresholding. We denote the positive and negative location prior by green and red pentagrams, respectively.

reweigh the confidence map S_1 by assigning zeros to the area within M_1. We denote the masked confidence map as S_2. After this, we enable PerSAM's decoder to subsequently segment the second cat and acquire M_2. In this way, our approach can iteratively mask the already segmented objects and segment all the expected similar objects, until there is no target in the image.

Confidence Thresholding. How to stop the iteration when there is no other expected object in the image? We introduce a thresholding strategy for adaptive control. As shown by the statistics in Table 12, we count the confidence scores of the positive location prior (the maximum score on the confidence map) for two groups of objects in the PerSeg dataset: 'Same' and 'Different', where we utilize DINOv2 (Oquab et al., 2023) as the image encoder. 'Same' utilizes the same object for reference and test, just like the normal evaluation. 'Different' utilizes one object for reference, but tests on all other 39 objects. We observe the scores in 'Same' are almost all larger than 0.5, while those in 'Different' are lower than 0.4. Therefore, we adopt a simple thresholding strategy to stop the iterative segmentation based on the confidence map with a threshold of 0.45, which can well discriminate different objects or categories for most cases, e.g., segmenting all the cats or dogs in the image shown in Figure 8. In this way, for a test image, if the maximum score in the confidence map is lower than 0.45, there is no more target object in the image and we would stop the iteration.

E.6 Is PerSAM-F Generalized Only to a Specific Object?

Our PerSAM-F can not only be personalized by a specific object, but also generalize to a certain category with the same amount of parameters. As visualized in Figure 3, given a reference cone/armour/crocodile in FSS-1000 dataset (Li et al., 2020), our PerSAM-F can well segment other similar cones/armours/crocodiles in test images. This is because objects of the same category can contain similar hierarchical structures, so the learned scale weights of PerSAM-F by one sample can

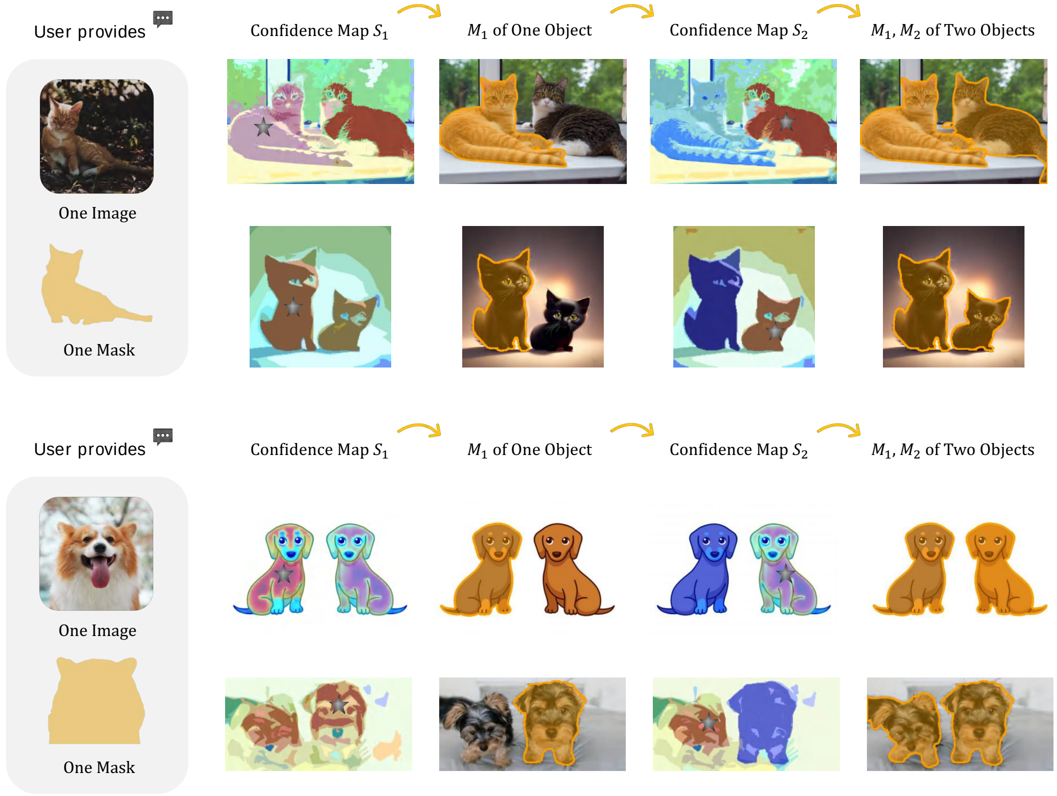

Figure 8: **Segmenting Objects of the Same Category.** Besides specific visual concepts, our approach can also be personalized by a category, cat or dog, with a confidence thresholding strategy.

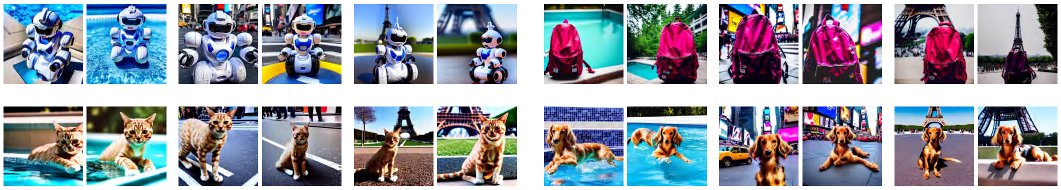

Figure 9: **Visualization of the Enlarged PerSeg Dataset** generated by a fine-tuned DreamBooth (Ruiz et al., 2022). We show the examples of four objects with three different text prompts: 'A photo of an [OBJECT] in a swimming pool/in Times Square/in front of Eiffel Tower.'

also be applicable to different objects within the same category. In contrast, for different categories, one needs to fine-tune two sets of scale weights to respectively fit their scale information.

E.7 Will PerSAM be Constrained by SAM's Limited Semantics by Class-agnostic Training?

Yes, due to SAM's inherent class-agnostic training, the visual features extracted by SAM's encoder contain limited category-level semantics. This might constrain the category-level discriminative capability for complex multi-object scenes. Observing this limitation, we locate the target object among other objects in test images entirely by feature matching, i.e., the location confidence map. Such a matching strategy only considers the appearance-based class-agnostic similarity, without category semantics. To this end, we can leverage other semantically rich image encoders, e.g., CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023), for PerSAM(-F) to improve the multi-object performance. We conduct an ablation study of different image encoders on DAVIS 2017

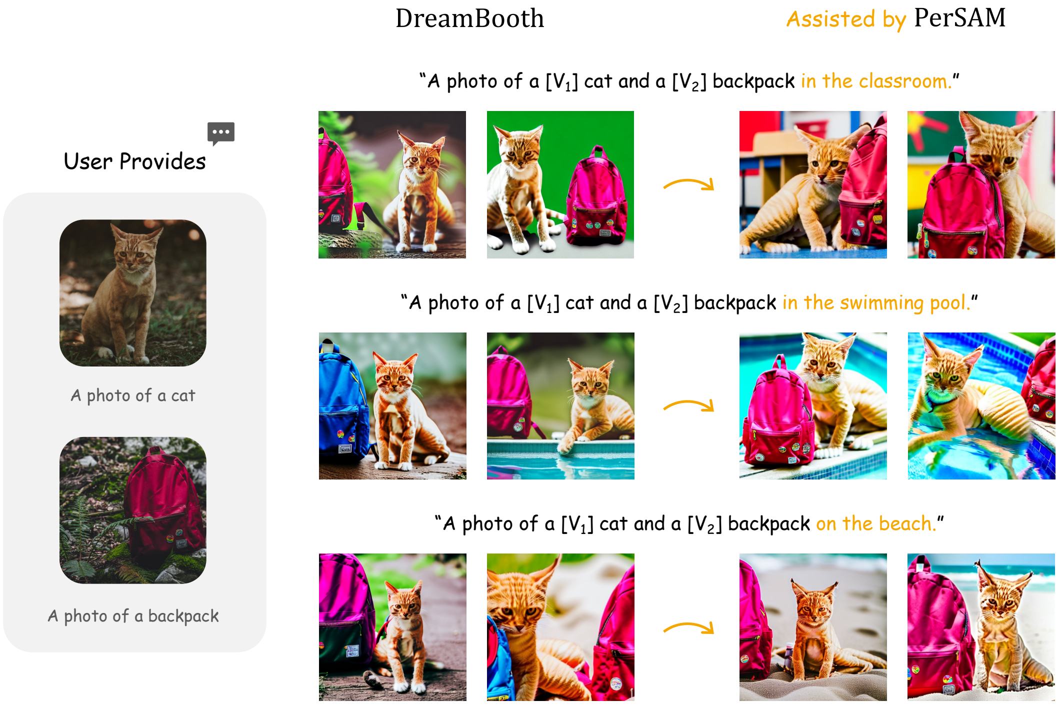
DreamBooth
Assisted by PerSAM
"A photo of a [V$_1$] cat and a [V$_2$] backpack in the classroom."
User Provides
A photo of a cat
"A photo of a [V$_1$] cat and a [V$_2$] backpack in the swimming pool."
A photo of a backpack
"A photo of a [V$_1$] cat and a [V$_2$] backpack on the beach."

Figure 10: **Multi-object text-to-image generation of PerSAM-assisted DreamBooth (Ruiz et al., 2022).**

Table 14: **Personalized Object Segmentation on the Enlarged PerSeg Dataset** with 5x largaer in size. We compare the overall mIoU and bIoU for different methods (Bar et al., 2022; Wang et al., 2022; 2023; Zou et al., 2023).

Method	Painter	SEEM	SegGPT	PerSAM	PerSAM-F
mIoU	43.6	82.8	87.8	85.9	89.6
bIoU	37.5	51.3	69.7	66.2	72.4

dataset (Pont-Tuset et al., 2017) for video object segmentation, which contains multiple similar objects within a video. As shown in Table 9, applying CLIP and DINOv2 with more sufficient semantic knowledge can improve the results of PerSAM-F for more challenging multi-object segmentation.

E.8 Can PerSAM help DreamBooth Achieve Better Multi-object Customization?

Yes. Similar to single-object personalization, we only calculate the loss within foreground regions for DreamBooth (Ruiz et al., 2022) with multi-object training samples. As visualized in Figure 10, we show the improvement for two-object customization assisted by our PerSAM. The backgrounds within images generated by DreamBooth are severely disturbed by those within few-shot training images, while the PerSAM-assisted DreamBooth can accurately synthesize new backgrounds according to the input language prompts.

E.9 Scaling PerSeg Dataset

Although our newly constructed PerSeg dataset contains different objects in various contexts, it is relatively small in scale compared to existing segmentation benchmarks. For a more robust evaluation, we enlarge the PerSeg dataset (40 objects with 5$\sim$7 images per object) to 30 images per object, **5x larger** in scale. We leverage the existing few-shot images to fine-tune DreamBooth (Ruiz et al.,

2022) respectively for each object, and then generate new images with diverse backgrounds or poses (swimming pool, Times Square, Eiffel Tower, etc. . . .), including richer data examples as shown in Figure 9. We report the segmentation results in Table 14, the scale-aware fine-tuned PerSAM-F still achieves the best performance, and the training-free PerSAM can also surpass Painter and SEEM, demonstrating the superior robustness of our approach.

E.10 Any Other Applications for PerSAM?

Besides improving the generation of DreamBooth (Ruiz et al., 2022), our PerSAM and PerSAM-F can also be utilized to assist other models and applications, such as CLIP (Radford et al., 2021) and NeRF (Mildenhall et al., 2021). For CLIP-based few-shot image classification, a series of works (Zhang et al., 2021; 2023d; Udandarao et al., 2022) extract the visual features of few-shot images by CLIP, and cache them as category prototypes for downstream adaption of CLIP. However, such prototypes contain the visual noises of the backgrounds that disturb the category semantics. Therefore, via the category-wise personalization approach, our PerSAM is helpful in segmenting the objects of the same category in few-shot images, and enables the CLIP-based methods to cache only foreground informative features. For 3D reconstruction by NeRF, existing approaches can only lift the objects, which are annotated with multi-view masks, into 3D space. Considering that the multi-view annotation is labor-intensive, our approach provides a solution for NeRF to lift any object in a scene, simply by prompting SAM to segment the object in one view. On top of that, PerSAM can be personalized to generate the masks in all multi-view images, allowing for efficient and flexible 3D reconstruction. We leave these applications as future works.

References

Grounded-segment-anything. https://github.com/IDEA-Research/Grounded-Segment-Anything, 2023.

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.

Yanqi Bao, Kechen Song, Jie Wang, Liming Huang, Hongwen Dong, and Yunhui Yan. Visible and thermal images fusion architecture for few-shot semantic segmentation. *Journal of Visual Communication and Image Representation*, 80:103306, 2021.

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.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

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

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.

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Xiaopeng Zhang, and Qi Tian. Segment anything in 3d with nerfs. *arXiv preprint arXiv:2304.12308*, 2023.

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.

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*, pp. 7345–7354, 2021.

Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

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*, pp. 1290–1299, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

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

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.

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.

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.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. IEEE, 2017.

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*, pp. 1551–1560, 2021.

Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pp. 59–75. Springer, 2022.

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. URL https://openreview.net/forum?id=0RDcd5Axok.

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

Lukas Hedegaard, Aman Alok, Juby Jose, and Alexandros Iosifidis. Structured pruning adapters. *arXiv preprint arXiv:2211.10155*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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*, pp. 2790–2799. PMLR, 2019.

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.

Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023.

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*, pp. 4904–4916. PMLR, 2021.

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*, pp. 709–727. Springer, 2022.

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*, pp. 36–54. Springer, 2022.

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*, pp. 539–556. Springer, 2023.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.

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.

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.

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

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2869–2878, 2020.

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*, pp. 7026–7035, 2019.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*, 2020.

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*, pp. 388–404. Springer, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023a.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023b.

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*, pp. 3431–3440, 2015.

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

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)*, pp. 13–23, 2019.

Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6941–6952, 2021.

Keval Morabia, Jatin Arora, and Tara Vijaykumar. Attention-based joint detection of object and semantic part. *arXiv preprint arXiv:2007.02419*, 2020.

Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 622–631, 2019.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23641–23651, 2023.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

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.

Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021.

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

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.

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*, pp. 8748–8763. PMLR, 2021.

Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7141–7151, 2023.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in Neural information processing systems*, 30, 2017.

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.

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.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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, 2020b.

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*, pp. 5227–5237, 2022.

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

Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 44(2):1050–1065, 2020b.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*, 2022.

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.

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.

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

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.

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, pp. 3056–3064, 2021.

Qianxiong Xu, Wenting Zhao, Guosheng Lin, and Cheng Long. Self-calibrated cross attention network for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 655–665, 2023.

Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9587–9595, 2019.

Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, 35:6575–6588, 2022.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023b.

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.

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*, 2023c.

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*, 2023d.

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*, pp. 2881–2890, 2017.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

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*, pp. 6881–6890, 2021.

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.

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.