CASE: Concept-Aware Segmentation from Positive and Negative Visual Exemplars

Yi-Jen Tsai[1], Yen-Yu Lin[1], and Chien-Yao Wang[2⋆]

[1] National Yang Ming Chiao Tung University, Taiwan
tsai.cs14@nycu.edu.tw, lin@cs.nycu.edu.tw
[2] Institute of Information Science, Academia Sinica, Taiwan
kinyiu@iis.sinica.edu.tw

Abstract. Few-shot semantic segmentation typically identifies target objects based on positive support examples of a single semantic class. In practice, however, positive cues alone are often underspecified: for instance, "red cars" may be confused with cars of all colors, "vehicles" may include cars that should be excluded, and "material-defined objects" may span multiple categories. Fulfilling precise requests requires positive exemplars to provide inclusion evidence and negative exemplars to act as intent-defining constraints for exclusion and disambiguation. We formulate this setting as **Specified Visual Prompt Segmentation (SVPS)**, where the target concept is specified by both positive and negative visual prompts. In this work, visual exemplars refer to support-image regions specified by bounding-box prompts. To study SVPS, we introduce **CASE** (**C**oncept-**A**ware **S**egmentation from Positive and Negative Visual **E**xemplars), a benchmark built on COCO and PACO. CASE encompasses six category- and attribute-level task families: exclusion, isolation, union, attribute exclusion, attribute isolation, and attribute union. We evaluate SAM 3 under this setting and find that, despite its powerful positive-only few-shot segmentation capability, frozen SAM 3 does not consistently treat negative visual prompts as exclusion evidence. To address this, we present **Concept Prompt Adaptation (CPA)**, a conservative LoRA-based adaptation baseline that improves positive-negative prompt interpretation on CASE while remaining competitive under positive-only few-shot segmentation. Code is available at https://github.com/WongKinYiu/FSS-CASE.

Keywords: Specified visual prompt segmentation · Few-shot segmentation · Visual prompting · Negative exemplars · SAM 3

1 Introduction

Semantic segmentation [3, 15, 38] is a fundamental task in computer vision that aims to assign a semantic label to every pixel in an image. Although deep models have achieved remarkable performance on standard benchmarks, they rely on large-scale datasets with dense, pixel-level annotations for training. Beyond

⋆ Corresponding author

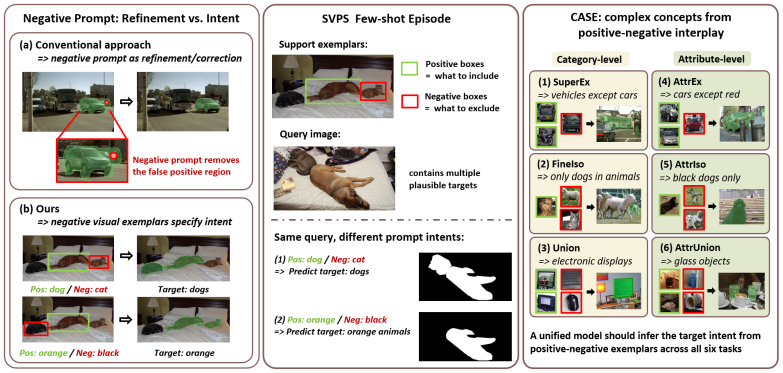


Fig. 1: Motivation and overview of CASE. Conventional negative prompts are commonly used as local refinement signals for an already specified target. In contrast, Specified Visual Prompt Segmentation (SVPS) treats positive and negative visual exemplars as intent-defining cues. The same query image can therefore induce different target under different prompts. CASE evaluates this setting through six task families, requiring a unified model to infer target intent from positive-negative exemplar.

expensive annotation costs, most models are trained on a fixed set of categories, resulting in limited generalization to unseen objects in open-world scenarios.

Few-shot semantic segmentation (FSS) [26, 28] addresses this limitation by segmenting novel objects in a query image using only a small number of annotated support examples. Existing FSS methods commonly adopt episodic training to learn transferable matching mechanisms between support and query images, such as prototype matching [26, 28], pixel-level correspondence [17, 20], or cross-image attention [31, 35]. Despite their success on standard FSS benchmarks, these methods largely formulate the problem as category-level matching: Namely, given one or more positive support examples, the model is expected to segment all objects of the same semantic class in the query image.

However, this formulation is overly restrictive for practical visual prompting. In real-world scenarios, user specifications frequently require both inclusion and exclusion cues, rather than purely positive category matching. For instance, positive examples may overgeneralize from "red cars" to all cars, from "buses and trains" to all vehicles (thereby incorrectly including cars), or from "wooden chairs" to all chairs rather than wooden objects. These complex targets cannot be precisely encapsulated by positive examples alone: While positive examples define what to include, they do not explicitly specify what should be excluded. Consequently, a model needs to jointly interpret positive and negative visual prompts to infer the target mask from their relationship.

We define this task as **Specified Visual Prompt Segmentation (SVPS)**, where the target concept is specified by both positive and negative visual exemplars. In this work, visual exemplars are support-image regions represented as bounding-box prompts. Positive exemplars provide visual evidence for inclusion, whereas negative exemplars act as intent-defining constraints that specify confusing categories, attributes, or instances to exclude. Unlike conventional FSS,

SVPS evaluates whether a model can infer a specified target concept from the interplay between positive and negative exemplars. As illustrated in Fig. 1, this changes the role of negative prompts from local mask refinement to concept-level intent specification: different positive-negative exemplar configurations can induce different target masks for the same query image.

To evaluate this task, we introduce **CASE** (**C**oncept-**A**ware **S**egmentation from Positive and Negative Visual **E**xemplars), a benchmark constructed from COCO [13] and PACO [22] annotations. The CASE benchmark encompasses six task families spanning category-level and attribute-level concept specification: exclusion, isolation, and union. These tasks are designed to assess whether a segmentation model can leverage negative visual prompts as concept-level exclusion evidence, rather than simply appending them as additional prompts without modeling their relation to the positive visual evidence.

We further study Segment Anything Model 3 (SAM 3) [1] under the SVPS setting. SAM 3 is a strong vision foundation model for promptable concept segmentation, supporting segmentation from text descriptions, visual exemplars, or their combination. A recent shared-canvas formulation [27] has shown that support and query images can be placed in a single spatial canvas, enabling a fully frozen SAM 3 to perform few-shot segmentation without additional training or architectural modification. This makes SAM 3 a suitable testbed for studying whether promptable segmentation models can handle positive-negative visual prompts beyond positive-only matching. Unlike general promptable concept segmentation in SAM 3 [1], where different prompt types serve as alternative ways to refer to a concept, SVPS evaluates whether positive and negative exemplars can jointly specify a constrained target concept.

However, the same line of evidence also reveals a key limitation: directly adding negative visual prompts to the few-shot segmentation pipeline does not necessarily improve performance and may even suppress target objects [27]. This behavior contradicts the intended role of negative prompts as exclusion cues. It suggests that, although SAM 3 can effectively exploit positive visual prompts, it does not inherently parse negative prompts as precise exclusion constraints. In few-shot settings, negative prompts may instead compete with positive prompts and weaken or suppress predicted target masks.

To address this issue, we propose **Concept Prompt Adaptation (CPA)**, a lightweight, LoRA-based adaptation strategy tailored for SAM 3 under positive-negative visual prompting. We use the term *concept* to highlight that CPA adapts the model to target concepts specified by positive-negative visual evidence. CPA updates only low-rank adaptation parameters while keeping the pretrained model frozen. Furthermore, we introduce positive-only preservation regularization that keeps the adapted model close to frozen SAM 3 when negative prompts are absent. This provides a conservative yet effective adaptation baseline that balances SVPS improvement on CASE while mitigating standard FSS degradation.

The primary contributions of this work are summarized as follows:

- We introduce **SVPS**, a new segmentation paradigm where positive and negative visual exemplars jointly specify a target concept.

- We present **CASE**, a benchmark for SVPS built upon COCO and PACO, covering six category- and attribute-level concept specification tasks.
- We propose **CPA**, a lightweight LoRA-based adaptation baseline equipped with positive-only preservation regularization, and evaluate the adaptation-preservation trade-off on both SVPS and positive-only FSS benchmarks.

2 Related Work

2.1 Few-Shot Semantic Segmentation

Few-shot semantic segmentation (FSS) [26, 28] aims to segment novel categories in a query image given only a small number of annotated support examples. Early approaches commonly adopt episodic training to learn matching functions. A prominent line of work represents support evidence with prototypes or prototype-enhanced class summaries [7, 12, 26, 28]. Although prototype-based methods are efficient, compressing multiple support masks into a global vector discards fine-grained spatial details crucial for precise localization.

To preserve richer support-query correspondences, subsequent methods explore pixel- or region-level matching through hypercorrelation [17], dense matching [20], or cross-attention [31, 35]. Despite effectiveness on FSS, these methods work under the restrictive assumption that the support set provides solely positive evidence for a single semantic class. Hence, the task is reduced to category-level matching, where the model merely transfers a positive support concept to the query image. In contrast, our work studies a more expressive setting where a target concept is jointly specified by positive and negative visual prompts.

2.2 Foundation Models for Promptable Segmentation

Vision foundation models such as CLIP [21] and DINO [2, 19] provide general-purpose visual representations that have reshaped many recognition [37, 40, 41] and segmentation [16, 23, 30, 39] tasks. Building on this trend, the Segment Anything Model (SAM) [11] demonstrates remarkable class-agnostic segmentation capabilities through spatial prompts such as points and boxes. Several works adapt SAM to FSS by generating visual reference prompts from support images [14, 25, 34]. While effective, these methods often introduce additional components, such as visual reference encoders [25], graph-based prompt refinement modules [34], or prompt-learning/adaptation modules [4, 9], to bridge the gap between support examples and query segmentation.

SAM 2 [24] extends promptable segmentation to videos through a memory mechanism for frame-to-frame mask propagation. Recent FSS methods based on SAM 2, such as FSSAM [32] and SANSA [5], adapt this memory mechanism by treating support images as memory references and query images as target frames. However, these memory-based formulations are primarily designed for temporal consistency and object tracking, making them suboptimal for semantic correspondences across temporally disjoint images.

SAM 3 [1] introduces Promptable Concept Segmentation (PCS), where an object concept can be specified by text prompts, positive or negative image exemplars, or their combinations. Native prompting operates on the image being segmented, whereas CASE requires cross-image transfer. A recent *shared-canvas* pipeline [27] further enables frozen SAM 3 to perform training-free FSS by placing support and query images into a single canvas, but directly incorporating negative prompts can degrade segmentation quality and suppress target objects. Rather than treating negative exemplars as interactive refinements to an existing concept prompt, SVPS studies whether positive and negative visual exemplars can jointly define complex target concepts through inclusion and exclusion roles.

2.3 Visual Prompting and Negative Prompts

Prompt design is crucial to foundation segmentation models. Existing works mainly improve positive guidance through visual reference prompts [25, 36], support-query feature matching [14, 34], learnable prompts or adaptation modules [4,9], and text- or image-guided prompting [6,16,29]. These mechanisms are used to localize a target object or transfer a positive support concept to a query image. Target concept ambiguity is commonly mitigated by strengthening positive evidence through refined localization [36], support-query matching [14, 34], or multimodal guidance [6, 16, 23, 29].

Negative prompts remain less explored in exemplar-based FSS. While negative points or boxes are commonly expected to suppress unwanted regions in user-interactive segmentation [11, 24], this assumption may not transfer directly to FSS, where negative exemplars define semantic exclusion constraints rather than merely correct local mask errors. Recent SAM 3-based FSS results [27] suggest that directly appending negative prompts reduces segmentation quality and suppresses target objects. This indicates that current promptable segmentation models fail to utilize negative visual prompts as exclusion evidence.

Beyond category-level segmentation, referring expression segmentation specifies targets through language descriptions involving attributes or context [6, 10, 29, 33], while datasets such as PACO [22] evaluate attribute-level visual understanding. Although language can describe constraints, subtle visual distinctions are easier to specify through visual examples. Motivated by this gap, we define Specified Visual Prompt Segmentation (**SVPS**), where positive and negative visual prompts jointly specify a target concept rather than serving as auxiliary inputs. The proposed **CASE** (**C**oncept-**A**ware **S**egmentation from Positive and Negative Visual **E**xemplars) benchmark systematically evaluates whether models can infer target concepts from positive-negative exemplar relationships, covering exclusion, isolation, union, and attribute-level disambiguation.

3 The CASE Benchmark

3.1 Benchmark Overview

We introduce **CASE**, a benchmark for **SVPS**. Traditional few-shot segmentation benchmarks evaluate whether a model can transfer a single positive support

Table 1: Task taxonomy of the CASE benchmark. Each task specifies a target concept through positive and negative visual exemplars.

Task	Target Concept	Positive Exemplars	Negative Exemplars	Challenge
SuperEx	Supercategory excluding one class, e.g., vehicles except cars.	Included classes from the supercategory, e.g., buses and trains.	The excluded class, e.g., cars.	Generalize to a broad group while enforcing explicit exclusion.
FineIso	One fine-grained class within a supercategory, e.g., dogs.	Target class instances, e.g., dogs.	Related distractor classes, e.g., cats or sheep.	Isolate a target class from visually close distractors.
Union	Classes sharing a common function, e.g., furniture for resting.	Functionally related classes, e.g., beds and sofas.	Related classes with different functions, e.g., chairs or tables.	Infer functional commonality across distinct objects.
AttrEx	A category excluding one attribute, e.g., non-red cars.	Instances without the excluded attribute, e.g., black or white cars.	Instances with the excluded attribute, e.g., red cars.	Apply visual negation by fine-grained attribute differences.
AttrIso	A category with one specified attribute, e.g., red cars.	Target attribute instances, e.g., red cars.	Same category with other attributes, e.g., blue or white cars.	Separate attribute-specific targets from general object identity.
AttrUnion	An attribute shared across categories, e.g., wooden objects.	Diverse classes with the attribute, e.g., wooden chairs or bowls.	Similar classes without the attribute, e.g., plastic chairs or bowls.	Group objects by attribute while avoiding category or shape bias.

concept to a query image. In contrast, CASE requires models to infer target concepts by joint reasoning over both positive and negative visual exemplars. Positive exemplars define regions to include, whereas negative exemplars specify related categories, attributes, or instances that must be excluded.

CASE is built on COCO [13] and PACO [22]. COCO provides category and supercategory annotations for category-level concept specification, while PACO offers annotations for fine-grained attribute-level concept specification. Each episode provides support-image bounding boxes as visual exemplars and requires the model to predict the target mask in the query image.

3.2 Task Taxonomy

As summarized in Tab. 1, CASE encompasses six task families that evaluate different facets of positive-negative concept specification. The category-level tasks comprise *supercategory exclusion* (SuperEx), *fine-grained isolation within a supercategory* (FineIso), and *functional union across related categories* (Union). The attribute-level tasks include *exclusion* (AttrEx), *isolation* (AttrIso), and *cross-category union* (AttrUnion). Together, these tasks assess whether a model can move beyond naive positive matching and leverage negative exemplars as exclusion or boundary constraints for precise target concept specification.

3.3 Episode Construction

Following the standard FSS protocol, CASE is organized in an episodic format. Each episode consists of a support set with positive and negative bounding-box prompts, a query image, and a binary ground-truth query mask. The support prompts specify the target concept, and the query mask serves as the expected segmentation output in the query image for evaluation.

To construct an episode, we first instantiate a positive-negative concept definition using COCO or PACO annotations, such as a supercategory paired with an excluded class, a function-driven category group, or an attribute-defined object group. We then sample support images containing the required positive and

negative exemplars. Query images are sampled independently to contain target objects and optionally include distractors. The target mask is generated from the underlying annotations according to the instantiated concept definition.

This design prevents the task from being solved by simply matching a positive support example. Because negative exemplars explicitly define what should be rejected or separated from the target, the model must infer the intended concept from the interplay between positive and negative prompts. Detailed per-task statistics and construction parameters are provided in the supplementary material, Sec. *CASE Benchmark Construction*.

4 Method

4.1 Task Formulation

Specified Visual Prompt Segmentation (SVPS) extends conventional FSS by defining the target concept through both positive and negative visual exemplars. Specifically, each *episode* contains a K-shot support set $\mathcal{S} = \{(I_i^s, \mathcal{B}_i^+, \mathcal{B}_i^-)\}_{i=1}^K$ and a query image I^q with the ground-truth mask M^q. The positive bounding boxes $\mathcal{B}_i^+$ enclose visual evidence that must be included, whereas the negative boxes $\mathcal{B}_i^-$ indicate irrelevant distractors to be excluded. Given these inputs, the model f_θ, parameterized by θ, predicts a binary mask $\hat{M}^q$ for the query image:

$$\hat{M}^q = f_\theta(\mathcal{S}, I^q). \tag{1}$$

Conventional FSS can be viewed as a special case of SVPS where $\mathcal{B}_i^- = \emptyset$ for all support examples. Therefore, an SVPS model should leverage negative exemplars to suppress distractors when they are provided, while retaining competitive segmentation performance under the positive-only setting.

4.2 Model Overview

Fig. 2 illustrates our **Concept Prompt Adaptation (CPA)** pipeline. CPA adapts SAM 3 [1] to SVPS by augmenting the original SAM 3 detector with two trainable LoRA [8] modules while keeping the original SAM 3 weights frozen. We separate the pipeline into two parts: (i) the shared-canvas input construction from the support set $\mathcal{S}$ and query image I^q, and (ii) the CPA-adapted SAM 3 forward path, where one LoRA module is attached to the transformer encoder and the other is attached to the transformer decoder.

Given $\mathcal{S}$ and I^q, CPA first composes them into a shared canvas I^c and transforms the positive and negative boxes into the canvas coordinate system [27]. The canvas image is processed by the SAM 3 image encoder, while the transformed prompt boxes are processed by the geometry encoder. The resulting features are passed through the SAM 3 transformer encoders and decoders, then through the prediction heads. The detector outputs mask logits Z^c, boxes, and confidence scores s^c on the composite canvas. We crop the query region from Z^c to obtain

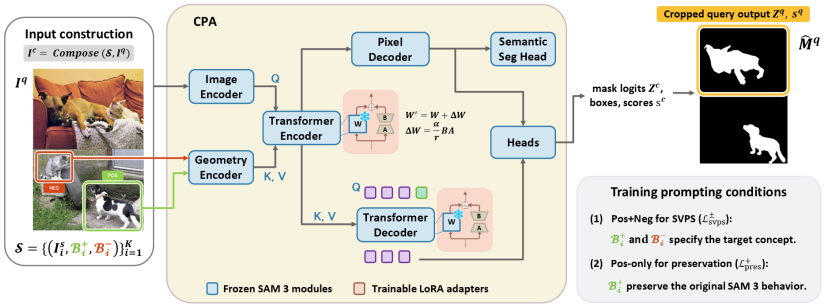


Fig. 2: Overview of Concept Prompt Adaptation (CPA). A K-shot support set $\mathcal{S}$ and query image I^q are composed into a shared canvas I^c, with positive and negative support boxes used as visual prompts. CPA augments the original SAM 3 detector with two trainable LoRA modules while keeping the original SAM 3 weights frozen. The final query prediction is obtained from the cropped query outputs Z^q and s^q. The inset summarizes the two prompting conditions used during training; their formal objectives are defined in Sec. 4.5.

query logits Z^q, with corresponding confidence scores denoted as s^q. The final query prediction $\hat{M}^q$ is derived from Z^q.

During training, CPA is optimized under two prompting conditions. In the positive-negative condition, the adapted model receives both $\mathcal{B}^+$ and $\mathcal{B}^-$ and is supervised by the CASE target mask M^q. In the positive-only preservation condition, only $\mathcal{B}^+$ is provided, and the adapted model is regularized to remain close to the original frozen SAM 3 prediction under the same positive-only input. The following subsections describe the shared-canvas formulation, LoRA placement in CPA, and the training objective, respectively.

4.3 Shared-Canvas Prompting

Following FSS-SAM3 [27], we adopt a shared-canvas formulation to process the support and query images within a unified spatial and attention space. The support set and query image are composed into a single image-like canvas:

$$I^c = \text{Compose}(\mathcal{S}, I^q), \tag{2}$$

and all support boxes $\{\mathcal{B}_i^+, \mathcal{B}_i^-\}_{i=1}^K$ are transformed into the coordinate system of I^c. The composite image and box prompts are then fed into SAM 3.

This formulation allows support and query regions to interact within the same attention space while reusing SAM 3's native box-prompt interface, eliminating the need for an explicit module. The SAM 3 detector heads output composite mask logits Z^c, boxes, and confidence scores s^c. We crop the query region from the composite logits to obtain query logits $Z^q = \text{Crop}_q(Z^c)$, and use the corresponding confidence scores s^q to select or rank predictions. The binary query mask is obtained from the selected query logits as $\hat{M}^q = \mathbf{1}[\sigma(Z^q) > \tau]$, where $\sigma(\cdot)$ denotes the sigmoid function and τ is the mask threshold. For CASE

episodes, both positive and negative boxes are used. For positive-only evaluation and preservation regularization, only positive boxes are provided.

4.4 LoRA-Based Concept Prompt Adaptation

Frozen SAM 3 already provides strong positive-only few-shot segmentation behavior, but it does not interpret negative visual prompts as concept-level exclusion cues. To address this limitation, CPA adapts SAM 3 to positive-negative prompting while mitigating degradation of its original positive-only behavior. Let f denote the original frozen SAM 3. CPA constructs an adapted model f_θ by augmenting f with two trainable LoRA modules: one attached to the transformer encoder and the other attached to the transformer decoder. All original SAM 3 parameters remain frozen, and θ contains only the parameters of these LoRA modules. For a pretrained attention projection matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, each LoRA module replaces the effective projection by

$$W' = W + \Delta W, \quad \Delta W = \frac{\alpha}{r} BA, \tag{3}$$

where $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$ are trainable matrices of rank $r \ll \min(d_{\text{in}}, d_{\text{out}})$, and α is a scaling factor. The LoRA modules shown in Fig. 2 correspond to low-rank updates injected into the transformer encoder and decoder.

4.5 Training Objective

CPA is optimized with two objectives: an SVPS segmentation loss for positive-negative prompting and a preservation loss for maintaining positive-only behavior. Let f_θ denote the adapted model and f denote the original frozen SAM 3 model. Given positive and negative prompts, the adapted model predicts query logits $Z^q_{\theta,\pm}$. We supervise this prediction using the ground-truth query mask M^q:

$$\mathcal{L}^{\pm}_{\text{svps}} = \mathcal{L}_{\text{mask}}(Z^q_{\theta,\pm}, M^q), \text{ where } Z^q_{\theta,\pm} = \text{Crop}_q(f_\theta(I^c, \mathcal{B}^+, \mathcal{B}^-)) \tag{4}$$

and $\mathcal{L}_{\text{mask}}$ combines sigmoid focal loss and Dice loss. The superscript $\pm$ indicates that the prediction is conditioned on both positive and negative prompts.

The preservation loss is computed under the positive-only prompting condition. For the same episode, the original frozen model f produces reference query logits $Z^q_{f,+}$ using only positive boxes. Under the same prompting condition, the adapted model f_θ outputs query logits $Z^q_{\theta,+}$. We penalize the discrepancy between the two probability maps using the squared L_2 norm:

$$\mathcal{L}^{+}_{\text{pres}} = \left\| \sigma(Z^q_{\theta,+}) - \sigma(Z^q_{f,+}) \right\|_2^2, \text{ where} \tag{5}$$

$$Z^q_{\theta,+} = \text{Crop}_q(f_\theta(I^c, \mathcal{B}^+)) \text{ and } Z^q_{f,+} = \text{Crop}_q(f(I^c, \mathcal{B}^+)), \tag{6}$$

and $\sigma(\cdot)$ denotes the sigmoid function. The superscript $+$ indicates that this term is computed under the positive-only prompting condition.

The final training objective is:

$$\mathcal{L} = \mathcal{L}^{\pm}_{\text{svps}} + \lambda_{\text{pres}} \mathcal{L}^{+}_{\text{pres}}, \tag{7}$$

where λ_{pres} controls the strength of positive-only preservation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate CPA under two settings.

CASE. CASE evaluates specified visual prompt segmentation with positive and negative visual exemplars. It contains six task families constructed from COCO [13] and PACO [22]: SuperEx, FineIso, and Union for category-level concepts, and AttrEx, AttrIso, and AttrUnion for attribute-level concepts. Each episode provides positive and negative support boxes, a query image, and a binary target mask defined by a structured concept specification, such as category exclusion, fine-grained isolation, functional union, attribute isolation, attribute union, or attribute exclusion. The evaluation split is fixed at the episode level to ensure deterministic comparison across models.

Standard FSS benchmark. We evaluate the CASE-trained CPA checkpoints directly on all four COCO-20^i folds [13, 18], without retraining or fold-specific tuning. COCO-20^i is a standard few-shot segmentation benchmark derived from COCO, where the 80 object categories are divided into four folds with 20 categories per fold. We report results under both 1-shot and 5-shot settings. Since standard FSS uses only positive support annotations, each support mask is converted into its tight bounding box and used as a positive visual prompt.

Evaluation Metrics. We use mean Intersection-over-Union (mIoU) as the primary metric. For CASE, we report task-wise mIoU and the average mIoU across all tasks. We further aggregate results into category-level and attribute-level groups. For COCO-20^i, we report fold-wise and mean mIoU under the standard positive-only 1-shot and 5-shot protocols.

Implementation Details. CPA is built on SAM 3 using the shared-canvas formulation of FSS-SAM3 [27], where support and query images are composed into a single 1008×1008 canvas. Positive and negative visual prompts are represented as support bounding boxes. We keep the pretrained SAM 3 weights frozen and insert LoRA modules into transformer encoder and decoder layers, resulting in 1.13M trainable parameters. CPA is trained with AdamW at a learning rate of 1×10^{-4}, with LoRA rank $r = 16$ and scaling factor $\alpha = 32$. Standard FSS evaluation follows the positive-only prompting protocol, which is also used for the preservation regularization. Query masks are binarized at $\tau = 0.5$. Additional details are provided in the supplementary material, Sec. *Implementation Details*.

Table 2: CASE benchmark results: task-wise and average mIoU across six tasks.

Method	Prompt	SuperEx	FineIso	Union	AttrEx	AttrIso	AttrUnion	**Mean**
				1-shot				
GF-SAM [34]	Pos Only	24.87	35.61	33.25	16.74	17.44	10.26	23.03
FSSAM [32]	Pos Only	28.55	38.97	35.89	17.28	15.87	14.83	25.23
SANSA [5]	Pos Only	26.96	53.65	29.83	15.65	22.48	17.07	27.61
FSS-SAM3 [27]	Pos Only	28.80	56.23	43.43	20.05	19.83	15.32	30.61
FSS-SAM3 [27]	Pos+Neg	26.22	20.43	30.52	17.06	6.33	11.17	18.62
CPA (Ours)	Pos Only	27.04	46.64	38.11	20.29	24.46	14.36	28.48
CPA (Ours)	**Pos+Neg**	**37.11**	**56.70**	**52.07**	**21.98**	**28.74**	**18.39**	**35.83**
				5-shot				
GF-SAM [34]	Pos Only	35.60	49.44	45.36	20.87	22.76	19.91	32.32
FSSAM [32]	Pos Only	37.11	55.13	44.51	17.76	17.87	20.61	32.17
SANSA [5]	Pos Only	34.37	61.55	36.89	18.45	25.24	23.87	33.40
FSS-SAM3 [27]	Pos Only	39.07	62.38	53.72	23.11	25.21	25.13	38.10
FSS-SAM3 [27]	Pos+Neg	39.76	50.83	47.50	20.74	21.48	24.55	34.14
CPA (Ours)	Pos Only	45.85	62.23	49.23	25.82	25.83	26.84	39.30
CPA (Ours)	**Pos+Neg**	**49.39**	**66.00**	**56.05**	**26.08**	**27.64**	**27.71**	**42.15**

Table 3: Aggregated CASE results by category-level and attribute-level concepts.

Method	Prompt	1-shot			5-shot		
		Category	Attribute	**Mean**	Category	Attribute	**Mean**
GF-SAM [34]	Pos Only	31.24	14.81	23.03	43.47	21.18	32.32
FSSAM [32]	Pos Only	34.47	15.99	25.23	45.58	18.75	32.17
SANSA [5]	Pos Only	36.81	18.40	27.61	44.27	22.52	33.40
FSS-SAM3 [27]	Pos Only	42.82	18.40	30.61	51.72	24.48	38.10
FSS-SAM3 [27]	Pos+Neg	25.72	11.52	18.62	46.03	22.26	34.14
CPA (Ours)	Pos Only	37.26	19.70	28.48	52.44	26.16	39.30
CPA (Ours)	**Pos+Neg**	**48.63**	**23.04**	**35.83**	**57.15**	**27.14**	**42.15**

5.2 Main Results on CASE

Tab. 2 shows that CASE exposes limitations that are not visible in standard positive-only FSS. Positive-only methods remain competitive on tasks close to category matching, especially FineIso, but they cannot explicitly encode what should be rejected. This becomes clear on exclusion, union, and attribute-based tasks, where a target concept cannot be specified by positive similarity alone. Directly adding negative prompts to frozen SAM 3 is also ineffective: FSS-SAM3 drops from 30.61 to 18.62 in 1-shot and from 38.10 to 34.14 in 5-shot. This suggests that the frozen model is not calibrated to treat negative boxes as concept-level exclusion cues; instead, they can act as unstable suppression signals.

By adapting SAM 3 to use negative exemplars relative to positive anchors, CPA improves over its positive-only variant from 28.48 to 35.83 in 1-shot and from 39.30 to 42.15 in 5-shot. The improvement is consistent across both category-level and attribute-level tasks, showing that lightweight adaptation enables the model to interpret positive and negative supports as a structured concept specification rather than as independent box prompts.

As shown in Tab. 3, foundation models favor category-level over attribute-level concepts due to object-centric priors. This highlights the value of CASE as

Table 4: Comparison with state-of-the-art FSS methods on COCO-20^i. CPA is evaluated under the standard positive-only protocol.

Method	Type	1-shot					5-shot				
		20^0	20^1	20^2	20^3	Mean	20^0	20^1	20^2	20^3	Mean
PFENet [26]	Classical FSS	36.5	38.6	34.5	33.8	35.8	36.5	43.3	37.8	38.4	39.0
HSNet [17]	Classical FSS	36.3	43.1	38.7	38.7	39.2	43.3	51.3	48.2	45.0	46.9
BAM [12]	Classical FSS	43.4	50.6	47.5	43.4	46.2	49.3	54.2	51.6	49.6	51.2
HDMNet [20]	Classical FSS	43.8	55.3	51.6	49.4	50.0	50.6	61.6	55.7	56.0	56.0
Matcher [14]	Foundation-based	52.7	53.5	52.6	52.1	52.7	60.1	62.7	60.9	59.2	60.7
GF-SAM [34]	Foundation-based	56.6	61.4	59.6	57.1	58.7	67.1	69.4	66.0	64.8	66.8
SANSA [5]	Foundation-based	58.9	62.6	61.5	58.0	60.2	–	–	–	–	67.3
FSSAM [32]	Foundation-based	59.9	65.6	62.1	61.6	62.3	68.6	74.0	64.5	69.9	69.3
FSS-SAM3 [27]	Foundation-based	**64.2**	**66.8**	**67.7**	**67.5**	**66.6**	68.6	72.5	72.5	73.6	71.8
CPA (Ours)	Ours	61.1	64.3	59.2	61.6	61.6	**74.4**	**74.1**	**74.6**	**76.2**	**74.8**

Table 5: Effect of positive-only preservation. CASE is evaluated with positive-negative prompts, while COCO-20^i is evaluated under the standard FSS protocol.

Method	CASE			COCO-20^i	
	Category	Attribute	Mean	1-shot	5-shot
FSS-SAM3 [27]	25.7	11.5	18.6	**66.6**	71.8
CPA w/o Preservation	**50.6**	**25.2**	**37.9**	50.8	66.8
CPA	48.6	23.0	35.8	61.6	**74.8**

a broader benchmark evaluating visual exclusions, compositional groupings, and attribute-defined targets. Comparing CPA with and without negative prompts shows that negative exemplars provide complementary support, improving the CASE mean from 28.48 to 35.83 in 1-shot and from 39.30 to 42.15 in 5-shot. The gains are more pronounced on category-level concepts, suggesting that negative prompts are especially useful when the target set can be refined by excluding related object categories or disambiguating broad concept groups. Attribute-level gains are smaller but consistent, indicating that material and color reasoning remains difficult. Detailed breakdowns are provided in the supplementary material, Sec. *Additional Experimental and Qualitative Results.*

5.3 Positive-Only FSS Evaluation

Although CPA is trained for SVPS, we evaluate its positive-only FSS behavior on COCO-20^i under the standard protocol, where support masks are converted into tight bounding boxes. As shown in Tab. 4, CPA obtains 61.6 mIoU in 1-shot, which is lower than FSS-SAM3. However, CPA achieves the best 5-shot performance, 74.8 mIoU, among all methods. This suggests that the adapted model remains responsive to positive visual prompts and can effectively exploit additional support evidence when multiple exemplars are available.

5.4 Ablation and Sensitivity Analysis

Effect of preservation regularization. Tab. 5 shows the adaptation–preservation trade-off. Removing preservation gives the best CASE performance, but substan-

Table 6: Negative prompt sensitivity on CASE-NegStress. We report mIoU (%). The full per-task results are provided in the supplementary material.

Group	Method	0 Neg	1 Neg	2 Neg	All Neg
Category	FSS-SAM3 [27]	37.28	29.82	21.86	12.53
	CPA	**49.29**	**59.66**	**60.27**	**59.19**
Attribute	FSS-SAM3 [27]	10.04	8.68	5.40	2.74
	CPA	**18.22**	**24.12**	**25.49**	**22.75**
Overall	FSS-SAM3 [27]	23.66	19.25	13.63	7.63
	CPA	**33.76**	**41.89**	**42.88**	**40.97**

tially degrades COCO-20^i positive-only FSS. Adding preservation slightly lowers CASE mIoU, yet recovers much of the positive-only FSS ability and yields strong 5-shot COCO performance. We therefore use the preservation-regularized CPA as the default, since it improves SVPS while maintaining the usability of SAM 3 under conventional positive prompts.

Negative prompt sensitivity. CASE-NegStress is a dedicated 1-shot subset that fixes the query image, target mask, positive prompts, and candidate negative pool while varying only the negative-prompt count; 0-Neg denotes Pos-only on this subset, not full CASE. As shown in Tab. 6, FSS-SAM3 degrades sharply as more negatives are added, dropping from 23.66 to 7.63 overall mIoU. In contrast, CPA improves from 33.76 to 40.97. This indicates that CPA learns to interpret negative prompts relative to positive anchors, while the original SAM 3 treats them as unstable suppression signals.

5.5 Qualitative Results and Limitations

Qualitative Results. Fig. 3 presents representative qualitative comparisons on CASE. SAM 3 with positive-only prompting often misses valid target regions or fails to respect the intended exclusion or attribute constraint. Naively adding negative prompts to SAM 3 remains unstable and may suppress valid target regions together with distractors. CPA with positive-only prompting remains responsive to positive prompts, but may retain distractor regions because the target concept remains underspecified without negative exemplars, especially for exclusion and attribute-based tasks. CPA with positive-negative prompting provides the most consistent predictions, producing masks that better match the concept jointly specified by positive and negative visual exemplars.

Limitations. Attribute-level tasks remain challenging when color or material cues are subtle or ambiguously annotated, and cross-category attribute union is particularly difficult because the model must group visually diverse objects by a shared property rather than by category or shape. CASE is also limited by the categories, masks, and attributes available in COCO and PACO. Additional qualitative results and failure cases are provided in the supplementary material, Secs. *Additional Qualitative Results* and *Failure Cases and Limitations*.

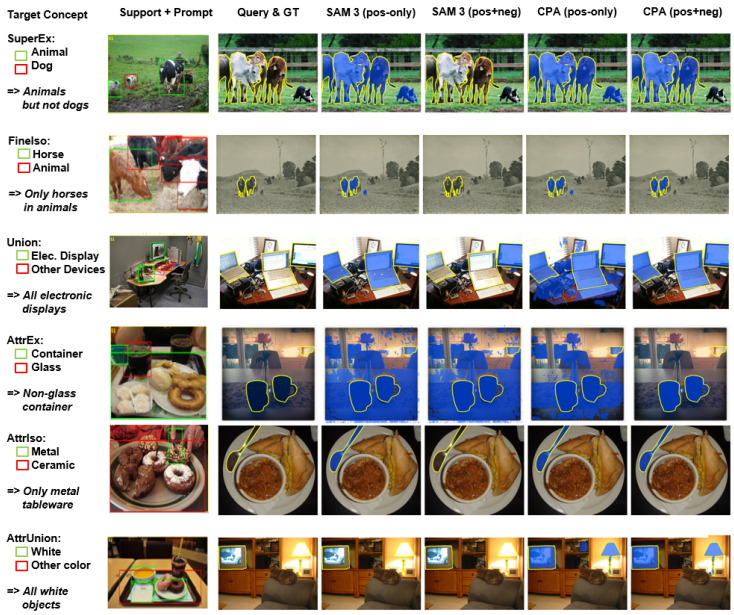


Fig. 3: Qualitative comparison on CASE. Green/red boxes in the support image indicate positive/negative visual prompts. Yellow contours denote ground-truth query masks, and blue overlays denote model predictions. We compare SAM 3 (FSS-SAM3) and CPA under positive-only and positive-negative prompting.

6 Conclusion

We introduced **Specified Visual Prompt Segmentation (SVPS)**, a new segmentation setting, where positive and negative visual exemplars jointly specify target concepts. To evaluate this setting, we constructed **CASE** (**C**oncept-**A**ware **S**egmentation from Positive and Negative Visual **E**xemplars), a benchmark built upon COCO and PACO annotations with six category- and attribute-level task families. Our experiments show that while frozen SAM 3 exhibits strong few-shot segmentation capabilities with solely positive prompts, it struggles to reliably interpret negative visual exemplars as exclusion evidence. To address this limitation, we presented **Concept Prompt Adaptation (CPA)**, a lightweight LoRA-based adaptation baseline to enhance positive-negative prompt interpretation. CPA significantly improves performance on CASE while keeping competitive performance under positive-only FSS setups. We believe that CASE can serve as a useful testbed for future research on positive-negative visual prompting, negative exemplars, and concept-level visual prompt segmentation.

Acknowledgements. This work was supported in part by the National Science and Technology Council (NSTC) under grants 114-2221-E-A49-038-MY3, 115-2221-E-A49-119-MY3, 115-2628-E-001-002 and 115-2634-F-A49-011, by IIS grant 3006-C4741, and by MediaTek.

References

1. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. In: ICLR (2026)
2. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 801–818 (2018)
4. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: IEEE/CVF International Conference on Computer Vision Workshops. pp. 3367–3375 (2023)
5. Cuttano, C., Trivigno, G., Averta, G., Masone, C.: Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation. In: NeurIPS (2025)
6. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: CVPR. pp. 16321–16330 (2022)
7. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: ECCV. pp. 701–719. Springer (2022)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
9. Huang, J., Jiang, K., Zhang, J., Qiu, H., Lu, L., Lu, S., Xing, E.: Learning to prompt segment anything models. arXiv preprint arXiv:2401.04651 (2024)
10. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 787–798 (2014)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
12. Lang, C., Cheng, G., Tu, B., Han, J.: Learning what not to segment: A new perspective on few-shot segmentation. In: CVPR. pp. 8057–8067 (2022)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll'ar, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
14. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: ICLR (2024)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
16. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: CVPR. pp. 7086–7096 (2022)
17. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV. pp. 6941–6952 (2021)
18. Nguyen, K., Todorovic, S.: Feature weighting and boosting for few-shot segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6231–6240 (October 2019)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

20. Peng, B., Tian, Z., Wu, X., Wang, C., Liu, S., Su, J., Jia, J.: Hierarchical dense correlation distillation for few-shot segmentation. In: CVPR. pp. 23641–23651 (2023)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
22. Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., Mousavi, A., Song, Y., Dubey, A., Mahajan, D.: Paco: Parts and attributes of common objects. In: CVPR. pp. 7141–7151 (2023)
23. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022)
24. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., R"adle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: ICLR (2025)
25. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Ding, E., Wang, J., Li, Z.: Vrp-sam: Sam with visual reference prompt. In: CVPR. pp. 23565–23574 (2024)
26. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. IEEE TPAMI **44**(2), 1050–1065 (2020)
27. Tsai, Y.J., Lin, Y.Y., Wang, C.Y.: Few-shot semantic segmentation meets sam3. arXiv preprint arXiv:2604.05433 (2026)
28. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Few-shot image semantic segmentation with prototype alignment. In: ICCV. pp. 9197–9206 (2019)
29. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR. pp. 11686–11695 (2022)
30. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: CVPR. pp. 18134–18144 (2022)
31. Xu, Q., Zhao, W., Lin, G., Long, C.: Self-calibrated cross attention network for few-shot segmentation. In: ICCV. pp. 655–665 (2023)
32. Xu, Q., Zhu, L., Liu, X., Lin, G., Long, C., Li, Z., Zhao, R.: Unlocking the power of sam 2 for few-shot segmentation. In: ICML (2025)
33. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85. Springer (2016)
34. Zhang, A., Gao, G., Jiao, J., Liu, C., Wei, Y.: Bridge the points: Graph-based few-shot segment anything semantically. In: NeurIPS. vol. 37, pp. 33232–33261 (2024)
35. Zhang, J.W., Sun, Y., Yang, Y., Chen, W.: Feature-proxy transformer for few-shot segmentation. In: NeurIPS. vol. 35, pp. 6575–6588 (2022)
36. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Ma, X., Dong, H., Gao, P., Li, H.: Personalize segment anything model with one shot. In: ICLR (2024)
37. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: ECCV. pp. 493–510. Springer (2022)
38. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 2881–2890 (2017)
39. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: ECCV. pp. 696–712. Springer (2022)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)

41. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)