Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects

Jian Hu[1*], **Jiayi Lin**[1*], **Weitong Cai**[1], **Shaogang Gong**[1]

[1]Queen Mary University of London

{jian.hu, jiayi.lin, weitong.cai, s.gong}@qmul.ac.uk

https://lwpyh.github.io/GenSAM/

Abstract

Camouflaged object detection (COD) approaches heavily rely on pixel-level annotated datasets. Weakly-supervised COD (WSCOD) approaches use sparse annotations like scribbles or points to reduce annotation efforts, but this can lead to decreased accuracy. The Segment Anything Model (SAM) shows remarkable segmentation ability with sparse prompts like points. However, manual prompt is not always feasible, as it may not be accessible in real-world application. Additionally, it only provides localization information instead of semantic one, which can intrinsically cause ambiguity in interpreting targets. In this work, we aim to eliminate the need for manual prompt. The key idea is to employ Cross-modal Chains of Thought Prompting (CCTP) to reason visual prompts using the semantic information given by a generic text prompt. To that end, we introduce a test-time instance-wise adaptation mechanism called Generalizable SAM (GenSAM) to automatically generate and optimize visual prompts from the generic task prompt for WSCOD. In particular, CCTP maps a single generic text prompt onto image-specific consensus foreground and background heatmaps using vision-language models, acquiring reliable visual prompts. Moreover, to test-time adapt the visual prompts, we further propose Progressive Mask Generation (PMG) to iteratively reweight the input image, guiding the model to focus on the targeted region in a coarse-to-fine manner. Crucially, all network parameters are fixed, avoiding the need for additional training. Experiments on three benchmarks demonstrate that GenSAM outperforms point supervision approaches and achieves comparable results to scribble supervision ones, solely relying on general task descriptions.

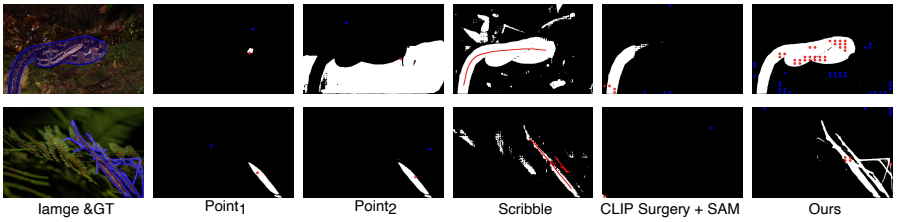
Introduction

Camouflaged Object Detection (COD) aims to accurately identify inconspicuous objects that have been carefully disguised, including those found in natural and artificial environments (Fan et al. 2017). The task's complexity is amplified by the indistinct boundaries between objects and backgrounds, necessitating a significant number of precisely annotated image-mask pairs. This places a rigorous demand on the annotation process (Hubel and Wiesel 1962; Pérez-de la Fuente et al. 2012; Pang et al. 2022). To alleviate this burden, weakly-supervised COD (WSCOD) is introduced to relax the annotation requirements. That only requires a sparse annotation in either the foreground or background. However, as annotations become sparser, they suffer from reduced accuracy. SAM introduces instance-level prompts to optimize segmentation, promising performance can be achieved with only a few points as prompts for each instance.

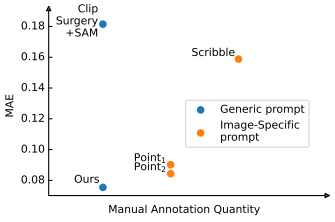
However, SAM has limited comprehension on its segmented object. Manual prompts can only provide location information of desired segmentation objects, but lack in semantic information, leading to potential ambiguity. In Fig. 1(a), despite both prompts targeting the same object, minor changes in the point prompt's position can make SAM misinterpret the desired object, greatly changing the results. Moreover, compared to human perception, it also exhibits a strong bias in interpreting prompts (Chen et al. 2023a). In Fig. 1(b), even with more information from scribble prompts than point prompts, SAM still misunderstands the target, leading to limited performance. To eliminate ambiguity and bias, recent work expands manual prompt input options to include reference regions, videos and even audios (Zou et al. 2023). However, regardless of the prompting method, SAM still requires instance-specific manual prompts, which may not always be practical in real-world scenarios, and the question of eliminating this need remains unexplored.

In this work, we introduce a test-time adaptation mechanism called Generalizable SAM (GenSAM), a novel approach to alleviating the demand for accurate, instance-specific manual prompts in the SAM framework for the WSCOD task. Given a simple text description, detailed semantic information about the desired object is reasoned based on both text and image information. Subsequently, this generates unambiguous visual prompt to guide the segmentation without human intervention. In order to provide semantic information of the target objects for SAM, we introduce Cross-modal Chains of Thought Prompting (CCTP) which automatically reason pixel-level visual prompts from various chains. BLIP2 (Zhu et al. 2023) and our devised CLIP (Radford et al. 2021) are employed for this propose. BLIP2 generates various keywords related to the target and their background using multiple chains of prompting (Wei et al. 2022), ambiguously in the input text prompts is eliminated. The spatial CLIP component introduces a novel self-attention mechanism, mapping keywords from different

[*]These authors contributed equally.



(a) Segmentation results using different prompts in SAM with various approaches.



(b) Mean absolute error on S-COD dataset.

Figure 1: In SAM, manual point and scribble prompts suffer from ambiguity in interpreting targets and is sensitive to minor spatial variations. Using a generic task description as a generic prompt with CLIP Surgery+SAM enables the model to achieve some segmentation capability on obvious objects. However, it struggles to perform well in complex environments with camouflage-like patterns. In contrast, our proposed GenSAM can adaptively convert a generic task description into image-specific visual prompts, effectively enhancing the segmentation process by leveraging the unique characteristics of each image.

chains onto a consensus heatmap for generating consensus visual prompts, thus resolving the visual prompt ambiguity induced by a particular chain. To progressively adapt the visual prompts, we further propose Progressive Mask Generation (PMG), which uses an test-time prompt tuning approach to iteratively reweight the input image with the consensus heatmaps. This guides the model to focus on the targets in a coarse-to-fine manner. It encourages the model to concentrate on task-relevant regions, enhancing the performance.

Our contribution can be summarized as follows:

(1) To eliminate the need for specific annotations tailored to each image in WSCOD, our GenSAM approach automatically generates personalized prompts for multiple unlabeled images using only a general task description.

(2) To convert task descriptions into precise visual prompts, we introduce an Cross-modal Chains of Thought Prompting module. It uses a consensus mechanism and a novel self-attention to derive image-specific prompts for SAM. Additionally, Progressive Mask Generation module utilizes the consensus heatmap as a visual prompt, progressively enhancing the segmentation performance.

(3) Extensive experiments on three benchmarks have demonstrated the effectiveness of our proposed GenSAM.

Related Work

Concealed Object Segmentation

The goal of camouflage detection is to identify objects that blend into complex backgrounds. Initially, some studies utilize low-level features such as texture, brightness, and color to distinguish the foreground from the background (Pike 2018; Hou and Li 2011; Sengottuvelan, Wahi, and Shanmugam 2008). Recently, several end-to-end approaches (Fan et al. 2020b,a) are proposed that achieve superior performance. However, most of these methods require fully annotated samples for training, which imposes a significant annotation burden. Weakly-Supervised Concealed Object Detection (WSCOD) aims to train a segmentation model using sparsely annotation like points and scribbles. Although WSCOD alleviates the reliance on pixel-level annotations, its performance is still limited by the quality and diversity of the training data. The lack of representative samples in a single dataset, as well as the restricted coverage of various scenes and objects, hinder the generalization ability of the model. For example, (He et al. 2023b) only requires scribble supervision, which achieves decent segmentation performance with lower annotation requirements per image. But its performance is limited by the quality of the annotations and lacks in generalization ability. Furthermore, to achieve satisfied performance across different datasets, current WSCOD approaches still requires separate training on different datasets, which limits their generalization ability.

Segment Anything model

Segment Anything Model (SAM) (Kirillov et al. 2023) is trained on the extensive SA-1B dataset, which comprises a vast collection of 11 million images and over 1 billion masks. This extensive dataset enables SAM to establish a robust foundation model for image segmentation with strong zero-shot generalization ability. While SAM is good at segmenting images, it struggles with segmenting camouflaged objects (Tang, Xiao, and Li 2023; Ji et al. 2023a,b). Moreover, its impressive ability requires the use of carefully crafted prompts to guide segmentation, which can be subjective and unclear. To address the challenges SAM encounters in camouflaged object detection, SAM-adaptor (Chen et al. 2023b) leverages a fully supervised dataset of camouflaged objects to train the encoder, yielding favorable results. However, this approach is hampered by its substantial demand for pixel-level annotated data. On the other hand, PLFMG (He et al. 2023b) enhances SAM's performance in WSCOD task through the application of pseudo labeling and multi-scale feature grouping. Regrettably, this method remains contingent upon separate training for different datasets within the WSCOD task, indicative of a deficiency in robust generalization capabilities. In contrast, our proposed approach only mandates a generic task description, enabling us to perform effective segmentation of concealed objects in unsupervised images across diverse datasets within the WSCOD task through instance-level test-time adaptation.

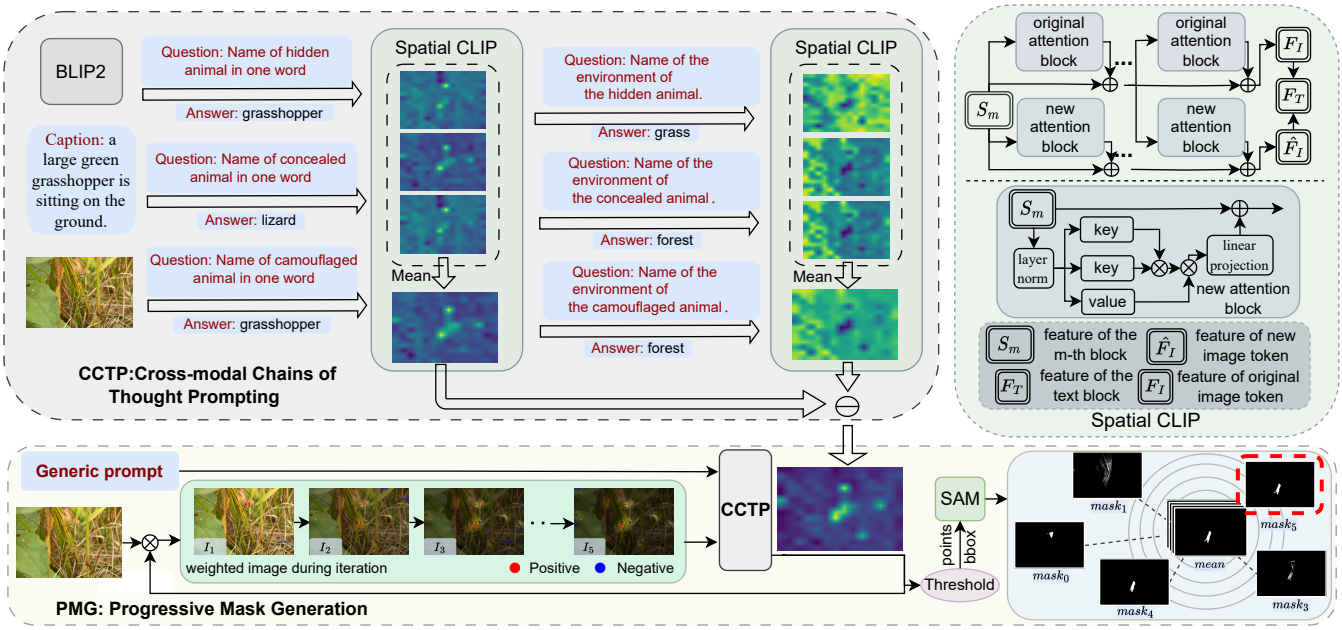


Figure 2: The framework of our proposed GenSAM. GenSAM consists of two components: Cross-modal Chains of Thought Prompting (CCTP) and Progressive Mask Generation (PMG). CCTP begins by taking a generic task prompt as input. BLIP2 generates an image caption for each image, using the input generic prompt as a foundation. Based on this prompt and generated caption, three parallel chains of thought are constructed to extract keywords about concealed animals and their corresponding background from unlabeled images. These keywords are then fed into our designed spatial CLIP module, which generates heatmaps for locating the camouflaged objects. High-confidence regions selected from these heatmaps serve as prompts to guide the segmentation process. The heatmaps generated by CCTP are weighted and utilized as visual prompts in PMG, gradually directing the model's attention towards task-relevant regions. In addition, during the adaptation process, the mask generated by a single iteration that is closest to the average mask obtained from multiple iterations is selected as the final output.

Test-time Adaptation

Test-time domain adaptation aims to adapt the model to a test domain that exhibits a domain gap with the training data (Wang et al. 2020; Hu et al. 2020), in order to improve the performance on the test data (Niu et al. 2022). Currently, there are two main categories of test-time domain adaptation: backward-based adaptation and backward-free adaptation. The former often utilizes self-supervised learning methods to learn the data characteristics of the target domain through entropy minimization (Wang et al. 2020; Hu et al. 2019, 2022). The latter mostly achieves backward-free adaptation through batch normalization statistic adaptation. DUA (Mirza et al. 2022) employs a running average technique to update the statistics and achieve adaptation, while DIGA (Wang et al. 2023) utilizes distribution adaptation via batch normalization to effectively perform semantic segmentation under domain gaps. In our work, we employ instance-level test-time domain adaptation, which simply relies on a general task description for camouflage object segmentation. It allows accurate camouflage object segmentation across diverse datasets without sample-level supervision.

Methodology

We present GenSAM for segmenting camouflaged objects among different domains, based on a general task description. In specific, we (1) propose *Cross-modal Chains of Thought Prompting*, which reasons the description of targeted objects in each image and further derives a consensus attention heatmap to generate visual prompts for the SAM model, and (2) employ an iterative process *Progressive Mask Generation* to apply the consensus heatmap onto the original image as a visual prompt to further improve segmentation results.Note that GenSAM is entirely training-free, only relying on pretrained components without additional training data or extra parameters during test-time adaptation.

Given an image $X \in \mathbb{R}^{H\times W\times 3}$ from a test set and a task-generic prompt P_g ("the camouflaged animal"), GenSAM aim at inferring the visual prompts for SAM to get the final segmentation mask $M \in \mathbb{R}^{H\times W}$. We relax the requirement for each unlabeled image under the same task to have a unique supervision and instead, adopt a common task-generic prompt shared by different unlabeled images across datasets within the same task.

Cross-modal Chains of Thought Prompting

The task of converting task-generic text prompts into image-specific visual prompts poses two main challenges: generating robust and objective image-specific prompts, and accurately localizing camouflaged objects in the images for effective segmentation. Therefore, we use multiple cross-

modal chains of thought to evaluate unlabeled images from different perspectives, generating potential keywords for both the camouflaged objects and their backgrounds. These keywords are then fed into our spatial CLIP module, which generates specific heatmaps for each foreground and background keyword. The foreground and background heatmaps undergo individual consensus calculation, and the background consensus heatmap is subtracted from the foreground consensus heatmap. The resulting heatmap highlights regions of high confidence, which are selected as prompts and used for segmentation with the SAM model.

Keyword Generation with various chains of thought. We utilize multiple chains of thought to produce a variety of keywords for both interested objects and their background, and generate heatmaps for both to remove irrelevant highlights of the background in the heatmaps. Specifically, we utilize BLIP2, an image-to-caption model that generates task-relevant object keywords based on generic prompts. These keywords are used to aid in localizing the objects of interest. However, directly using the task-generic prompt P_g to query BLIP2 for camouflaged objects in the image leads to inaccurate answers (Tab. 2). Inspired by generated knowledge prompting (Liu et al. 2021), we propose a method that involves having BLIP2 initially generate a caption C for the image X. This generated caption is then incorporated to enhance the model's ability to make more precise predictions when querying about the image-specific targets.

$$C = BLIP2(X), \tag{1}$$

As in large-scale generative models (LLM) (Bai et al. 2021; Zhu et al. 2023) like BLIP2, prompts require careful design and even slight variations in the prompts can lead to significant differences in the generated keywords. (Wang et al. 2022) proposes to design various chains of prompts for generative language models, and then derive a consensus from the output results to serve as the final output. It assumes that the output results of BLIP2 can be determined through majority voting among a limited number of possible outcomes. Therefore, a single prompt often provides only partial and biased descriptions of objects. Therefore, for the same image X, we inquire from different perspectives using various prompts to obtain different descriptive keywords A_j^{fore} for interested foreground objects (camouflaged objects), where $fore$ represents the foreword keyword and j denotes the $j-$th chain of thought. As shown in Fig. 2, with the corresponding task description P_g as "camouflaged animal" we first have BLIP2 generate a caption C for the image X. Then, using C as a basis, we replace P_g "camouflaged animal" with two synonymous phrases, "hidden animal" and "concealed animal" creating J different chains of thought $\{(Q_j^{fore}, Q_j^{back})\}_{j=1}^{J}$ with similar meanings from different perspective, simultaneously propose different questions in parallel to inquire about X. For example, Q_1^{fore}:"Name of the hidden animal in one word", Q_2^{fore}:"Name of the concealed animal in one word" and Q_3^{fore}:"Name of the camouflaged animal in one word.". Then, the obtained foreground keyword A_j^{fore} can be denoted as follows,

$$A_j^{fore} = BLIP2(X, C, Q_j^{fore}), \tag{2}$$

Moreover, camouflaged objects often hide themselves using textures or backgrounds. Therefore, identifying the background of camouflaged objects can significantly mitigate interference from unrelated objects. To achieve this, we further use the background query Q_j^{back} by including inquiries about the background of the A_j^{fore} (e.g, "grasshopper" in Fig. 2). This enables us to obtain the background keywords A_j^{back} as follows,

$$A_j^{back} = BLIP2(X, C, Q_j^{fore}, A_j^{fore}, Q_j^{back}). \tag{3}$$

Spatial CLIP. Due to CLIP's powerful open-vocabulary capability, it can handle various text descriptions. Hence, we input the generated keywords A_j^{fore} and A_j^{back} into CLIP, aiming to leverage its cross-modal alignment capability to highlight the corresponding regions in the image related to the interested object. However, CLIP's open-vocabulary capability also results in the generated heatmap containing numerous irrelevant information unrelated to the task.

Clip Surgery (Li et al. 2023) employs v-v-v self-attention to address it, where the queries (Q), keys (K), and values (V) in self-attention mechanism are all replaced by the values (V). The similarity between queries and keys is computed using the same vectors. This design enhances computational efficiency as there is no need to differentiate between queries and keys. But using the same representation for queries, keys, and values might limit the model's ability to capture internal correlations and features within the input image token, as they are mixed in the representation space.

To further enhance the location accuracy of the heatmap and effectively explore the internal structures and semantic correlations within the image, we propose the k-k-v self-attention paralleled to the original k-q-v path. The element-wise multiplication of the keys vectors reduces interference from redundant features, enabling the self-attention mechanism to focus more on internal correlations within the input image. This results in a better representation of the image's internal structure and patterns. Additionally, the "kkv" approach, using different vectors for values, preserves more information from the original inputs, enriching the context and enhancing the model's expressive capabilities. For the m-th block of the visual encoder in CLIP, its input features are denoted as S_m, where $S_m \in \mathbb{R}^{L\times d}$, with $K = S_m \cdot W_k$, $V = S_m \cdot W_v$, and $Q = S_m \cdot W_q$. Here, $W_k \in \mathbb{R}^{d\times d_k}$, $W_v \in \mathbb{R}^{d\times d_v}$, and $W_q \in \mathbb{R}^{d\times d_q}$ are learnable parameter matrices. Next, we split K, V, and Q into h_0 individual heads, each with dimensions d_k, d_v, and d_q, respectively. Assuming we have h_0 heads, the dimensions of K, V, and Q can be represented as follows: $K = \{k_1, k_2, ..., k_{h_0}\}$, $Q = \{q_1, q_2, ..., q_{h_0}\}$, and $V = \{v_1, v_2, ..., v_{h_0}\}$. Here, $k_n \in \mathbb{R}^{L\times d_k}$, $v_n \in \mathbb{R}^{L\times d_v}$, and $q_n \in \mathbb{R}^{L\times d_q}$ represent the transformation results of the n-th head. In our proposed k-k-v self-attention strategy, Q is replaced by K, and the expression for k-k-v self-attention is as follows:

$$\text{attention}_{kkv} = \text{softmax}(K * K^T * scale) * V, \tag{4}$$

where $scale$ is the scaling factor. The output of the K-Q-V block S_{m+1} and modified K-K-V self attention block $\hat{S}_{m+1}$

are defined as follows:

$$S_{m+1} = f_{\text{FFN}}(\text{attention}_{kqv}(S_m) + S_m),$$
$$\hat{S}_{m+1} = \begin{cases} \text{None, if } m < \delta \\ f_{\text{FFN}}(\text{attention}_{kkv}(S_m) + S_m) & \text{if } m = \delta \\ f_{\text{FFN}}(\text{attention}_{kkv}(S_m) + \hat{S}_m) & \text{if } m > \delta \end{cases}, \tag{5}$$

where $\hat{S}_m$ and S_m represent the output of k-k-v self-attention and k-q-v self-attention, respectively, for the m-th layer block. Shallow layers in the spatial CLIP focus more on low-level features and less on higher level concepts (e.g. meanings of input keywords). The original k-q-v attention mechanism is used in these layers. Deep layers start to represent higher level semantics like people and animals. Thus, for layers at depth δ, S_{m-1} is utilized as the input for the new module. Beyond depth δ, $\hat{S}_{m-1}$ and $S_{m-1}$from the previous block are both used as input. Following Clip Surgery (Li et al. 2023), the value of δ is set to 7. These features are then accumulated and mixed using the fully connected layer f_{FFN}. Image X, processed through the original image path and the k-k-v image path, yields image features F_I and $\hat{F}_I$, respectively. The corresponding text feature of the keyword A_j^{fore} and A_j^{back} are F_j^{fore} and F_j^{back}.

Visual prompt with consensus. After obtaining the corresponding features of different keywords, our objective is to identify a consensus among these features to locate specific regions of interest related to the task. Specifically, the consensus derived from foreground keyword generation is utilized to pinpoint the precise location of the camouflaged objects, while the consensus derived from background keyword generation helps eliminate interference from the background in object localization. For foreground keywords, the feature dimensions of the foreground keyword F_j^{fore} and the image feature $\hat{F}_I$ are $N_i \times 1 \times C$, where N_i represents the size of the image tokens, 1 represents the size of the text token, and C represents the number of channels. Consequently, the foreground similarity vector SI_{fore}^j obtained for the j-th keyword, is defined as:

$$SI_{fore}^j = \frac{F_j^{fore}}{||F_j^{fore}||_2} \odot \frac{\hat{F}_I}{||\hat{F}_I||_2}, \tag{6}$$

where $\odot$ is element-wise multiplication. L2 normalization is applied across the channel dimension. To derive a consensus among various image features corresponding to different foreground keywords, SI_{fore} can be obtained as follows:

$$SI_{fore} = \frac{\sum_{j=1}^{J}(SI_{fore}^j)}{J}, \tag{7}$$

where j is the number of the chains, we set J as 3. We also obtain the corresponding background consensus SI_{back} in a similar way. Then the resulting similarity heatmap SI is:

$$SI = SI_{fore} - SI_{back}, \tag{8}$$

where $SI \in \mathbb{R}^{N_i \times 1}$. SI is then upsampled using bilinear interpolation. After upsampling SI to match the original size of image X, the resulting output can be regarded as the consensus heatmap H corresponding to X. We further sample highly-activated pixels on H as positive point prompts and the same number of the most unactivated pixels as negative point prompts to guide the segmentation process in SAM.

Progressive Mask Generation

However, a single inference may not provide satisfactory segmentation result. For image with complicated background, some backgound objects can also be highly activated in the heatmap, causing some noises for inference the point prompts. In order to get more roubust prompt, we use the heatmap as a visual prompt to reweight the original image and guide the model during test time adaptation. The weighted image X' is as follows:

$$X' = X * H * w_{pic} + X * (1 - w_{pic}), \tag{9}$$

where $w_{pic} = 0.3$ is a hyper-parameter. The weighted image X' is then used as input image for next iteration. In this way, we develop a circularly test-time adaptation framework which involves multiple iterations of inference in a coarse-to-fine manner.

Moreover, in subsequent iterations, we use the previous iteration's mask to guide the segmentation by drawing bounding boxes as a post-process. We select the box with the highest Intersection over Union (IoU) value with the mask as our choice. It optimizes the current iteration and improves the consistency of segmentation results. The mask obtained at the i-th iteration is defined as M_i, where $i \in \{1, .., \mathbf{Iter}\}$, $\mathbf{Iter}$ is set as 6. To eliminate the impact of ambiguity caused by inconsistent prompts in each iteration, the mask obtained in each iteration is averaged. Finally, the selected iteration i^* is determined by selecting the iteration's result that closely resembles the average mask across all iterations as follows:

$$\mathrm{i}^* = \arg\min_i \left(\left| M_i - \frac{\sum_i (M_1, \ldots, M_{\mathbf{Iter}})}{\mathbf{Iter}} \right| \right), \tag{10}$$

then M_{i^*} is the corresponding final mask for X.

Experiments

To evaluate GenSAM in different scenarios, we choose challenging camouflaged object detection (COD) datasets to evaluate our GenSAM under various settings.

Setup

Datasets. Camouflaged object detection tasks aim to identify organisms attempting to camouflage themselves from complex backgrounds. In this study, we select three representative datasets containing samples of camouflage objects: CHAMELEON (Skurowski et al. 2018), CAMO (Le et al. 2019) and COD10K (Fan et al. 2021a). CHAMELEON dataset comprises 76 images sourced from the Internet for testing purposes. CAMO dataset consists of 1,250 images, with 1,000 images allocated for training and 250 images for testing. COD10K dataset encompasses a total of 3,040 training samples and 2,026 testing samples.

Baseline. We compare current SOTA weakly supervised segmentation methods, namely SAM (Kirillov et al. 2023), WSSA (Zhang et al. 2020), SCWS (Yu et al. 2021), TEL

Table 1: Results on COD with point supervision and scribble supervision. Best are in **bold** .

Methods	Venue	CHAMELEON				CAMO				COD10K			
		$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
Scribble Supervision Setting													
WSSA(Zhang et al. 2020)	CVPR20	0.067	0.692	0.860	0.782	0.118	0.615	0.786	0.696	0.071	0.536	0.770	0.684
SCWS(Yu et al. 2021)	AAAI21	0.053	0.758	0.881	0.792	0.102	0.658	0.795	0.713	0.055	0.602	0.805	0.710
TEL(Zhang et al. 2020)	CVPR22	0.073	0.708	0.827	0.785	0.104	0.681	0.797	0.717	0.057	0.633	0.826	0.724
SCOD(He et al. 2023b)	AAAI23	**0.046**	**0.791**	**0.897**	**0.818**	**0.092**	**0.709**	**0.815**	**0.735**	0.049	0.637	**0.832**	0.733
SAM(Kirillov et al. 2023)	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730
SAM-S(Kirillov et al. 2023)	ICCV23	0.076	0.729	0.820	0.650	0.105	0.682	0.774	0.731	**0.046**	**0.695**	0.828	**0.772**
Point Supervision Setting													
WSSA(Zhang et al. 2020)	CVPR20	0.105	0.660	0.712	0.711	0.148	0.607	0.652	0.649	0.087	0.509	0.733	0.642
SCWS(Yu et al. 2021)	AAAI21	0.097	0.684	0.739	0.714	0.142	0.624	0.672	**0.687**	0.082	0.593	0.777	0.738
TEL(Zhang et al. 2020)	CVPR22	0.094	**0.712**	**0.751**	**0.746**	0.133	**0.662**	0.674	0.645	0.063	0.623	**0.803**	0.727
SCOD(He et al. 2023b)	AAAI23	**0.092**	0.688	0.746	0.725	0.137	0.629	0.688	0.663	**0.060**	0.607	0.802	0.711
SAM(Kirillov et al. 2023)	ICCV23	0.207	0.595	0.647	0.635	0.160	0.597	0.639	0.643	0.093	0.673	0.737	0.730
SAM-P(Kirillov et al. 2023)	ICCV23	0.101	0.696	0.745	0.697	**0.123**	0.649	**0.693**	0.677	0.069	**0.694**	0.796	**0.765**
Task-Generic Prompt Setting													
CLIP Surgery+SAM(Li et al. 2023)	Arxiv2023	0.180	0.557	0.710	0.637	0.206	0.466	0.666	0.573	0.187	0.448	0.672	0.601
GenSAM	Ours	**0.090**	**0.680**	**0.807**	**0.764**	**0.113**	**0.659**	**0.775**	**0.719**	**0.067**	**0.681**	**0.838**	**0.775**

Table 2: Ablation study of variants with our GenSAM on camouflaged object detection.

Method's variant					settings on camouflaged object detection											
					CHAMELEON				CAMO				COD10K			
BLIP2 keyword	chain foreground	PMG	kkv self-attention	chain background	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
					0.180	0.557	0.710	0.637	0.206	0.466	0.666	0.573	0.187	0.448	0.672	0.601
✓					0.106	0.689	0.803	0.749	0.200	0.503	0.676	0.602	0.146	0.556	0.735	0.673
✓	✓				0.094	0.687	0.800	0.754	0.198	0.521	0.687	0.613	0.143	0.569	0.740	0.681
✓	✓	✓			0.098	0.659	0.779	0.741	0.161	0.554	0.719	0.642	0.086	0.616	0.797	0.731
✓	✓	✓	✓		**0.078**	**0.711**	**0.817**	**0.776**	0.147	0.583	0.746	0.666	0.069	0.660	0.820	0.760
✓	✓	✓	✓	✓	0.090	0.680	0.807	0.764	**0.113**	**0.659**	**0.775**	**0.719**	**0.067**	**0.681**	**0.838**	**0.775**

Table 3: Ablation study on COD10K.

(a) Number of chains.

Chains	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
1	0.069	0.671	0.827	0.767
2	**0.066**	0.679	0.832	0.772
3	0.067	0.681	**0.838**	**0.775**
4	**0.066**	0.680	0.834	0.773
5	**0.066**	**0.683**	0.833	**0.775**

(b) Heatmap upsample factor.

Factor	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
8	0.110	0.496	0.750	0.689
4	0.082	0.596	0.806	0.741
2	**0.067**	**0.681**	**0.838**	**0.775**
1	0.081	0.658	0.808	0.754
0.5	0.107	0.595	0.753	0.708

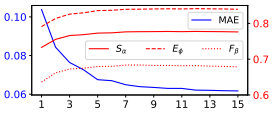
(c) Heatmap threshold.

Thr.	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
0.80	0.107	0.549	0.767	0.717
0.85	0.080	0.623	0.814	0.754
0.90	**0.067**	**0.681**	**0.838**	**0.775**
0.95	0.068	0.679	0.818	0.763

(d) Post processing.

Post processing	$M\downarrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
None	0.073	**0.683**	0.822	0.763
MaxBox	0.107	0.639	0.799	0.746
Mask	0.114	0.666	0.800	0.753
MaxIOUBox	**0.067**	0.681	**0.838**	**0.775**

(e) Iteration of adaptation.



(Zhang et al. 2020), and SCOD (He et al. 2023b). Three distinct levels of supervision are introduced, including scribble supervision and point supervision, alongside our proposed task-generic prompt settings. Following the previous weakly supervised segmentation setting (He et al. 2023a), in terms of scribble supervision, it involves providing foreground and background supervision by drawing the primary structure of objects and background areas. Point supervision refers to the provision of separate points as supervision for the foreground and background. In our newly proposed generic task prompt, we provide a unified prompt description "the camouflaged animal" for all images. The model is required to independently convert this unified description into specific supervision to guide the segmentation process based on the characteristics of each image. Regarding SAM, we follow the suggested setup by (He et al. 2023a) that involves comparing two variants of SAM: SAM-S and SAM-P. They finetune the mask decoder of SAM through scribble and point supervision training data respectively. Both variants employ partial cross-entropy loss for training. Note that all the comparative methods we employ in our study are trained on camouflage segmentation datasets and tested on a separate test set. However, GenSAM does not require training data at all, while directly utilize the test set for test-time adaptation. Based on the approaches used in previous studies (Fan et al. 2021a, 2020a), we employ four commonly used metrics for evaluation, include Mean Absolute Error (M), adaptive F-measure (F_β) (Margolin, Zelnik-Manor, and Tal 2014), mean E-measure (E_ϕ) (Fan et al. 2021b), and structure measure (S_α) (Fan et al. 2017). It is worth noting that a smaller value of M or larger values of F_β, E_ϕ, and S_α indicate better segmentation performance.

Implementation Details. For image-to-caption model, we use the BLIP-2 ViT-g OPT$_{6.7B}$ verison of BLIP2. For CLIP, we apply the CS-ViT-B/16 pretrained model. As for obtaining the point prompt from the heatmap, we use threshold=0.9 to filter out all the positive points and get the same amount of points with the lowest scores as the negative point prompt. In each iteration, we use the output mask generated from the last iteration as an auxiliary prompt in addition to the point prompts to guide SAM in the current iteration, to ensure consistency in each iteration. We totally apply 6 iterations to get our best results. We use PyTorch framework and conduct experiments on a single NVIDIA A100 GPU.

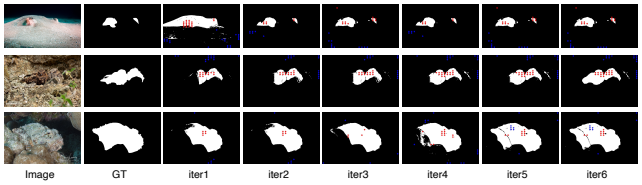


Figure 3: Iterative qualitative results of GenSAM. The visualized results indicate that as test-time adaptation progresses, the segmentation results consistently improve.

Experiment Results and Analysis

Experiment Results. As shown in Tab. 1, we compare GenSAM with approaches that use different supervision methods, including scribble supervision, point supervision, and our newly proposed generic task prompt supervision. Overall, due to varying levels of supervision signals, scribble supervision outperformed point supervision. However, our GenSAM, despite having only one generic task prompt as the universal supervision for the entire dataset, achieved superior performance compared to point supervision in terms of the M, S_α, and E_ϕ metrics on CHAMELEON. GenSAM approaches the performance level of scribble supervision. This trend is even more pronounced on the more challenging CAMO and COD10K datasets. Remarkably, on the most challenging COD10K dataset, our method even achieves better results in terms of the S_α and E_ϕ metrics under the less supervised generic task prompt supervision, surpassing both scribble supervision and point supervision approaches. This further demonstrates the superiority of GenSAM. Additionally, our method consistently outperforms SAM, SAM-P, SAM-S and CLIP Surgery+SAM, indicating that the improvements of our method stem from its own merits rather than relying solely on the superior segmentation capabilities of SAM. The qualitative results are shown in Fig. 3.

Component Analysis We further analyze the impact of components in Tab. 2. When all modules are removed, the model became CLIP Surgery+SAM. We use the general task description "the camouflaged animal" for each sample. The performance is weak across various datasets in this case. This result emphasizes the effectiveness of our complete GenSAM approach. In the third row, we add heatmaps as visual prompts during the iterative process. It is noticed that there is a notable performance improvement. It shows the significance of setting heatmaps as visual prompts, although there is still performance gap compared to GenSAM. In the second and fifth rows, we add consensus heatmaps for foreground and background using the chain-of-thought prompting. The comparison with other experiments emphasizes the importance of chain-of-thought for achieving consensus. The last two rows emphasize the importance of using a chain-of-thought to remove background interference. The unusual results on the CHAMELEON dataset are due to its small size, resulting in greater randomness, while the results from the other two larger and more complex datasets match our expectations. In the forth row, we replace our k-k-v self-attention with the v-v-v self-attention from CLIP Surgery. A notable decrease in performance is observed compared to using k-k-v, indicating the impact of k-k-v self-attention. As shown in Fig. 3e, the performance of the model's test-time adaptation shows a significant boost within the first 1-6 iterations. Then it gradually stabilizes thereafter. Although the best performance is achieved at the 8th iteration, the improvement compared to the 6th iteration is not substantial, and it incurs additional time loss. Therefore, we set the number of iterations to 6.

Table 4: Results on Polyp Image Segmentation and Shadow Detection with generic task prompt.

Datasets	Methods	M↓	$F_\beta\uparrow$	$E_\phi\uparrow$	$S_\alpha\uparrow$
ETIS (Silva et al. 2014)	CLIP Surgery+SAM	0.537	0.047	0.296	0.272
(Polyp Image Segmentation)	GenSAM	**0.205**	**0.090**	**0.554**	**0.430**
SBU(Vicente et al. 2016)	CLIP Surgery+SAM	0.331	0.336	0.517	0.442
(Shadow Detection)	GenSAM	**0.215**	**0.421**	**0.621**	**0.529**

Generalization of GenSAM. In Tab. 4, we assess GenSAM's performance on two segmentation tasks: Polyp Image Segmentation and Shadow Detection. We employ the generic prompts "Polyp" and "Shadow" for them respectively. Experiments demonstrate the significant improvement achieved by our method compared to the baseline.

Number of chains. We also evaluate the impact of the number of chains in the chain-of-thought prompting in Tab. 3a. When the number of chains is equal to or less than 3, performance gradually improves with an increasing number of chains. However, when the number of chains exceeded 3, although the inference time increases, there is no significant improvement in performance or even a decline.

From heatmap to point prompt. In Tab. 3b-3c we performe a parameter scan of heatmap upsample factor, the threshold to get the point prompts from heatmap. As the original heatmap SI is of relatively low resolution (14×14), we upsample SI to obtain H and obtain the point prompt using a threshold. We finally set the heatmap upsample factor as 2 and the threshold as 0.9

Generating box prompt. In Tab. 3d, we try different methods to transform the mask output from the previous iteration into an auxiliary mask or box prompt in addition to the point prompt to guide the current iteration, which ensures the consistency of mask outputs in different iterations. We try different transformation methods, including directly using the last mask output as the mask prompt (Mask), using the maximum surrounding box (MaxBox) and the box that has the largest Intersection over Union of the mask (MaxIOUBox). Results show that MaxIOUBox outperforms other strategies

Conclusion

In this paper, we present GenSAM, which automatically generates image-specific consensus prompts for WSCOD with only a generic task description via our test-time progressive mask generation framework. Experiments on various COD datasets show GenSAM's superiority.

Acknowledgements. This work was supported by Veritone, the Alan Turing Institute Turing Fellowship, and the China Scholarship Council.

References

Bai, C.-Y.; Lin, H.-T.; Raffel, C.; and Kan, W. C.-w. 2021. On training sample memorization: Lessons from benchmarking generative modeling with a large-scale competition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2534–2542.

Chen, F.; Chen, L.; Han, H.; Zhang, S.; Zhang, D.; and Liao, H. 2023a. The ability of Segmenting Anything Model (SAM) to segment ultrasound images. *BioScience Trends*.

Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Zhang, S.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2023b. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, and More. *arXiv preprint arXiv:2304.09148*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2021a. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 6024–6042.

Fan, D.-P.; Ji, G.-P.; Qin, X.; and Cheng, M.-M. 2021b. Cognitive vision inspired object segmentation metric and loss function. *Scientia Sinica Informationis*, 6(6).

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2777–2787.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273. Springer.

He, C.; Li, K.; Zhang, Y.; Xu, G.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2023a. Weakly-Supervised Concealed Object Segmentation with SAM-based Pseudo Labeling and Multi-scale Feature Grouping. *arXiv preprint arXiv:2305.11003*.

He, R.; Dong, Q.; Lin, J.; and Lau, R. W. 2023b. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 781–789.

Hou, J. Y. Y. H. W.; and Li, J. 2011. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 15: 2201–2205.

Hu, J.; Tuo, H.; Wang, C.; Qiao, L.; Zhong, H.; and Jing, Z. 2019. Multi-Weight Partial Domain Adaptation. In *BMVC*, 5.

Hu, J.; Tuo, H.; Wang, C.; Qiao, L.; Zhong, H.; Yan, J.; Jing, Z.; and Leung, H. 2020. Discriminative partial domain adversarial network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, 632–648. Springer.

Hu, J.; Zhong, H.; Yang, F.; Gong, S.; Wu, G.; and Yan, J. 2022. Learning Unbiased Transferability for Domain Adaptation by Uncertainty Modeling. In *European Conference on Computer Vision*, 223–241. Springer.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106.

Ji, G.-P.; Fan, D.-P.; Xu, P.; Cheng, M.-M.; Zhou, B.; and Van Gool, L. 2023a. SAM Struggles in Concealed Scenes–Empirical Study on" Segment Anything". *arXiv preprint arXiv:2304.06022*.

Ji, W.; Li, J.; Bi, Q.; Li, W.; and Cheng, L. 2023b. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184: 45–56.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Bras, R. L.; Choi, Y.; and Hajishirzi, H. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255.

Mirza, M. J.; Micorek, J.; Possegger, H.; and Bischof, H. 2022. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14765–14775.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, 16888–16905. PMLR.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2160–2170.

Pérez-de la Fuente, R.; Delclòs, X.; Peñalver, E.; Speranza, M.; Wierzchos, J.; Ascaso, C.; and Engel, M. S. 2012. Early evolution and ecology of camouflage in insects. *Proceedings of the National Academy of Sciences*, 109(52): 21414–21419.

Pike, T. W. 2018. Quantifying camouflage and conspicuousness using visual salience. *Methods in Ecology and Evolution*, 9(8): 1883–1895.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sengottuvelan, P.; Wahi, A.; and Shanmugam, A. 2008. Performance of decamouflaging through exploratory image analysis. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, 6–10. IEEE.

Silva, J.; Histace, A.; Romain, O.; Dray, X.; and Granado, B. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9: 283–293.

Skurowski, P.; Abdulameer, H.; Błaszczyk, J.; Depta, T.; Kornacki, A.; and Kozieł, P. 2018. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6): 7.

Tang, L.; Xiao, H.; and Li, B. 2023. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*.

Vicente, T. F. Y.; Hou, L.; Yu, C.-P.; Hoai, M.; and Samaras, D. 2016. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *ECCV*, 816–832. Springer.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, W.; Zhong, Z.; Wang, W.; Chen, X.; Ling, C.; Wang, B.; and Sebe, N. 2023. Dynamically Instance-Guided Adaptation: A Backward-Free Approach for Test-Time Domain Adaptive Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24090–24099.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yu, S.; Zhang, B.; Xiao, J.; and Lim, E. G. 2021. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3234–3242.

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12546–12555.

Zhu, D.; Chen, J.; Haydarov, K.; Shen, X.; Zhang, W.; and Elhoseiny, M. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*.