GraCo: Granularity-Controllable Interactive Segmentation

Yian Zhao[1,3] Kehan Li[1,3] Zesen Cheng[1,3] Pengchong Qiao[1,2,3] Xiawu Zheng[4]
Rongrong Ji[4] Chang Liu[5] Li Yuan[1,2,3] Jie Chen[1,2,3] ✉

[1]School of Electronic and Computer Engineering, Peking University, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China
[4]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China
[5]Department of Automation and BNRist, Tsinghua University, Beijing, China
zhaoyian@stu.pku.edu.cn jiechen2019@pku.edu.cn

Abstract

*Interactive Segmentation (IS) segments specific objects or parts in the image according to user input. Current IS pipelines fall into two categories: single-granularity output and multi-granularity output. The latter aims to alleviate the spatial ambiguity present in the former. However, the multi-granularity output pipeline suffers from limited interaction flexibility and produces redundant results. In this work, we introduce **Gra**nularity-**Co**ntrollable Interactive Segmentation (**GraCo**), a novel approach that allows precise control of prediction granularity by introducing additional parameters to input. This enhances the customization of the interactive system and eliminates redundancy while resolving ambiguity. Nevertheless, the exorbitant cost of annotating multi-granularity masks and the lack of available datasets with granularity annotations make it difficult for models to acquire the necessary guidance to control output granularity. To address this problem, we design an any-granularity mask generator that exploits the semantic property of the pre-trained IS model to automatically generate abundant mask-granularity pairs without requiring additional manual annotation. Based on these pairs, we propose a granularity-controllable learning strategy that efficiently imparts the granularity controllability to the IS model. Extensive experiments on intricate scenarios at object and part levels demonstrate that our GraCo has significant advantages over previous methods. This highlights the potential of GraCo to be a flexible annotation tool, capable of adapting to diverse segmentation scenarios. The project page: https://zhao-yian.github.io/GraCo.*

1. Introduction

Interactive Segmentation (IS) aims to segment specific objects or parts according to user interactions, providing a pixel-level interactive AI system that follows human intent. Recently, remarkable progress has been achieved in IS, resulting in various applications such as controllable image generation [42, 49], image editing [4, 20], and the well-known pixel-level annotation. Extensive research has been undertaken on various types of interactive information, such as bounding boxes [23, 48], scribbles [1, 11, 31], and clicks [6, 7, 19, 35, 36, 38, 45, 47]. Among them, the click-based interaction becomes mainstream due to its simplicity and well-established training and evaluation protocols.

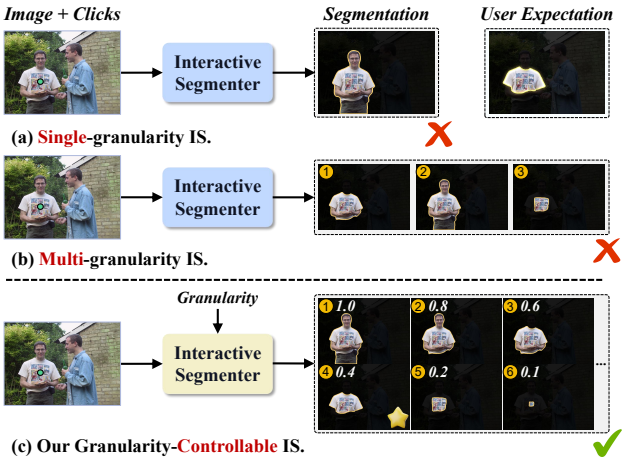

Figure 1. **(a)**: Single-granularity IS ignores spatial ambiguity. **(b)**: Multi-granularity IS is limited in the number of outputs and produces redundant results. **(c)**: Our Granularity-Controllable IS allows precise control of output granularity to match user expectations by attaching additional parameters to the input.

The current click-based IS methods are based on deep learning technology. Xu *et al*. [47] first introduces this technology to formulate IS and establishes training and evaluation protocols. Specifically, clicks are typically encoded into distance maps and then combined with the image to send the semantic segmentation model for interactive association training between clicks and GT masks. The emer-

✉ Corresponding author.

gence of SAM [22] strengthens the advancement of IS and proposes multi-granularity output pipelines to alleviate spatial ambiguity. The ambiguity refers to the concept that, given an interaction click, the desired segmentation region for the user may be the concept of objects with different parts nearby. However, this multi-granularity output pipeline suffers from limited scalability and produces redundant results, requiring the selection of the optimal mask based on confidence or user expectations.

Intuitively, the spatial ambiguity arises from the sparse clicks information supplied by the user, which fails to impose sufficient constraints for the model to establish a distinctive dense mask. To address this, we aim to achieve **Gra**nularity-**Co**ntrollable Interactive Segmentation (**GraCo**), which introduces a granularity control parameter to the input to explicitly constrain the model. For instance, the granularity can be controlled by a value ranging from 0 to 1, where a lower value corresponds to a finer granularity and vice versa, as shown in Figure 1. This approach allows precise control of prediction granularity, thereby enhancing the customization of pixel-level AI systems for human-machine interaction and eliminating redundancy while resolving ambiguity. However, the exorbitant cost of annotating multi-granularity masks and the lack of available datasets with granularity annotations corresponding to the masks make it difficult for models to acquire the necessary guidance to control output granularity.

To acquire the any-granularity masks and granularity annotations at a low cost, we design an Any-Granularity mask Generator (AGG) that is fully automated and does not require any additional manual annotation. Specifically, AGG consists of two key components: a mask engine and a granularity estimator. For the mask engine, we observe that object-level pre-trained IS models (*e.g.*, SimpleClick [38]) demonstrate the semantic property in delineating local concepts and object parts via appropriate interaction signals, which has the potential to generate proposals of any granularity, shape and intricacy. Based on this observation, we propose the multi-granularity loop simulation to automatically simulate the human-in-the-loop mechanism and generate diverse interaction signals to drive the mask engine. To estimate the granularity of the masks, we design the granularity estimator and establish computational rules from both the scale and semantic perspectives to ensure that the model behaviour is consistent with human cognition. Based on the mask-granularity pairs generated by AGG, we develop a simple yet efficient granularity-controllable learning (GCL) strategy, which incorporates the granularity embedding into the input and employ LoRA [18] technology. This enables the IS model to efficiently possesses granularity controllability while maintaining the original IS performance without requiring extensive computational cost.

To evaluate the performance of the IS models in multi-granularity scenarios, we follow standard protocols [47] and conduct experiments on both object and part level benchmarks. For the object-level, we perform evaluation on four commonly used datasets including GrabCut [43], Berkeley [40], SBD [14], and DAVIS [41]. For the part-level, we employ the part segmentation datasets PascalPart [5] and PartImageNet [15]. Thanks to the abundant mask-granularity pairs generated by AGG and the GCL strategy, the pre-trained IS model efficiently grasps the granularity controllability, achieving inspiring performance across all benchmarks on both levels. Specifically, our GraCo surpasses the state-of-the-art single-granularity IS methods on all benchmarks, especially on part-level benchmarks. Furthermore, GraCo outperforms the multi-granularity IS approach SAM [22] on all benchmarks and achieves comparable performance on SA-1B [22].

The main contributions can be summarized as: (i). We propose granularity-controllable interactive segmentation, which allows precise control of prediction granularity, thereby enhancing the flexibility of IS models and eliminating redundancy while resolving ambiguity; (ii). We explicitly exploit the semantic property of the pre-trained IS models and design a fully automated any-granularity mask generator to generate abundant mask-granularity pairs; (iii). We propose granularity-controllable learning strategy that enables the IS model to achieve inspiring performance on all benchmarks at both object and part levels.

2. Related Work

Single-granularity Interactive Segmentation. Interactive Segmentation (IS) is a thriving field due to its adaptability and broad applications. Early studies for IS typically utilize the low-level features and build optimization-based methods, including graph cut with max-flow algorithm [3], random walk [11], geodesic distance [2], and star-convexity [12]. These methods usually suffer from unsatisfactory performance when processing complex surroundings. DIOS [47] first introduces deep learning for IS, which proposes a click sampling strategy and establishes training and evaluation protocols. Based on this framework, researchers propose a range of optimization schemes from the perspectives of global segmentation and local refinement. FCA-Net [35] highlights the significance of first click. RITM [46] propose an iterative sampling strategy in training. BRS [19, 45] introduces online optimization to correct mislabeled pixels. CDNet [6] designs a conditional diffusion module to optimize segmentation. FocusCut [36] and FocalClick [7] focus on local refinement to improve the mask quality. GPCIS [50] formulates IS as a Gaussian process classification to fully propagate click information. SimpleClick [38] and iCMFormer [27] achieve superior performance using a Transformer-based architecture that has made brilliant achievements in the field of computer

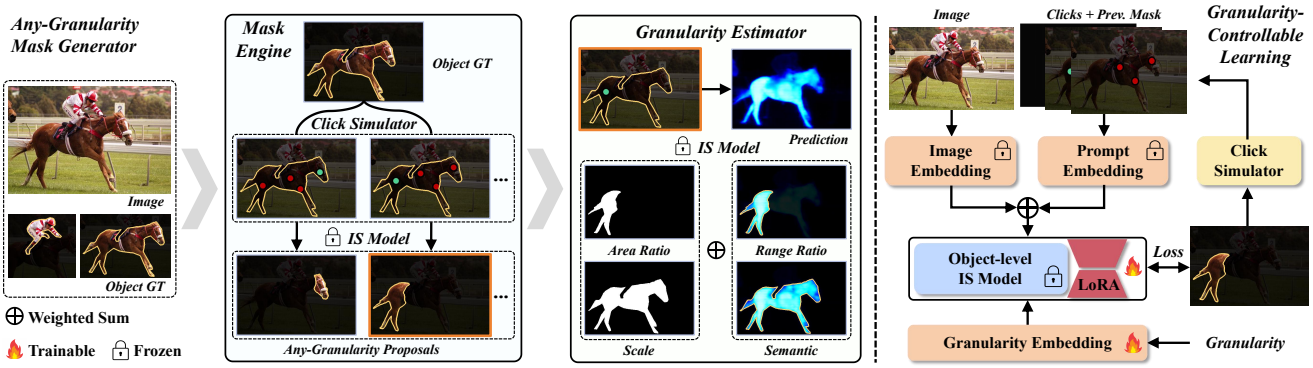

Figure 2. **Illustration of our granularity-controllable interactive segmentation.** Our GraCo consists of two stages. For the first stage, the Any-Granularity mask Generator (AGG) is designed to automatically generate any-granularity proposals (mask engine) and granularity annotations (granularity estimator) based on the object GT, without requiring additional manual annotation. For the second stage, the mask-granularity pairs generated by AGG are utilized to perform Granularity-Controllable Learning (GCL) on the object-level pre-trained IS model, enabling the model to efficiently possesses granularity controllability.

vision [26, 28, 39]. These methods are all single-granularity output pipelines, ignoring spatial ambiguity.

Multi-granularity Interactive Segmentation. A few efforts have been made to tackle the ambiguity in IS. LD [33] proposes to overcome this challenge by using two convolutional networks to select from coarse to precise. Recently, the emergence of SAM [22] boosts the progress of IS. SAM provides a unified interface to support multiple types of interactions and utilizes the diversity training to attain multi-granularity masks. Semantic SAM [24] extends the multi-granularity output, but is limited to generating pre-defined segments and only supports a positive click. These models learn multiple possibilities [13] of sparse prompts to dense masks mapping from large-scale multi-granularity annotations [5, 15, 22, 34, 44], which requires expensive data and training costs. Although the multi-granularity output pipeline alleviates ambiguity, it results in excessive output redundancy and limited scalability. Unlike previous works, our GraCo resolves ambiguity without redundancy and allows flexible control of prediction granularity without additional manual annotation and extensive training.

Instance and Part Segmentation. Instance segmentation is a fundamental task in computer vision that aims to accurately detect and segment each instance. Instance segmentation has achieved remarkable results after decades of development, and representative works include [8, 16, 25]. Part segmentation is a sub-task of image segmentation that aims to segment instances into more fine-grained parts. By identifying the internal structure of objects, part segmentation provides a more comprehensive visual understanding, with typical works including [9, 30]. Although instance and part segmentation are oriented towards different granularities, both only support segmentation at a fixed granularity and cannot perform human-machine interaction. Our GraCo supports not only the segmentation of specific parts, but also the flexible manipulation of the granularity level.

3. The Proposed GraCo

3.1. Overall Approach

In this section, we elaborate how to construct the proposed GraCo. The process of implementing GraCo consists of two stages, *cf.* Figure 2. In the first stage, we design an Any-Granularity mask Generator (AGG), which includes the mask engine and the granularity estimator (*cf.* Section 3.2). The mask engine employs the multi-granularity loop simulation to automatically generate abundant part proposals, and the granularity estimator is responsible for quantifying the granularity of each proposal. In the second stage, the mask-granularity pairs generated by the previous stage are utilized to perform Granularity-Controllable Learning (GCL) on the object-level pre-trained IS model (*cf.* Section 3.3). The details are described as follows.

3.2. Any-Granularity Mask Generator

Mask Engine. The core of AGG is the automatic generation of abundant mask-granularity pairs. To achieve this goal, we exploit the semantic property of the pre-trained IS model to segment local concepts and object parts by simulating appropriate interaction clicks. Specifically, we first utilize the instance GT as the mask prompt, and randomly select a positive point within the mask to input into the model, marking the object to be parsed. To drive the mask engine, we design a multi-granularity loop simulation to generate diverse interaction clicks. At each loop iteration, the click simulator takes a negative click from the current mask and appends it to the click set (*cf.* Figure 3). The current mask is then updated with the model prediction. Formulaically, given an image $\boldsymbol{I} \in \mathbb{R}^{h \times w \times 3}$ and a click set $\boldsymbol{\mathcal{C}}$, the positive and negative clicks in set $\boldsymbol{\mathcal{C}}$ are transformed into the disk map $\boldsymbol{D} \in \mathbb{R}^{h \times w \times 2}$. The object GT is denoted as

Figure 3. **Illustration of the multi-granularity loop simulation and visualization of the mask proposals generated by AGG.**

$\boldsymbol{G} \in \{0,1\}^{h\times w}$, the IS model $\mathcal{F}(\cdot)$ outputs the probability for each pixel being foreground. The mask generation process is as follows:

$$\boldsymbol{Y}_0 = \mathcal{F}(\text{Fusion}(\boldsymbol{I}, \boldsymbol{D}_0, \boldsymbol{G})),\ \boldsymbol{Y}_0 \in [0,1]^{h\times w}, \quad (1)$$

$$\boldsymbol{Y}_t = \mathcal{F}(\text{Fusion}(\boldsymbol{I}, \boldsymbol{D}_t, \boldsymbol{Y}_{t-1})),\ t = 1, 2, \ldots, N, \quad (2)$$

where $\boldsymbol{Y}_t$ represents the output mask in the t-th simulation, N is the number of iterations, and $\text{Fusion}(\cdot)$ is a fusion operation (*e.g.*, addition) of all types of features. In each iteration, we check that the new click is not too close to existing clicks in $\boldsymbol{\mathcal{C}}$, to avoid confusion. After the loop simulation, the mask engine generates abundant part proposals with diverse granularity. Furthermore, considering that an entire object consists of multiple parts, we regard the complement within the object of each proposal also as effective parts to increase the diversity of proposals and improve the efficiency of the mask engine. All proposals are saved after post-processing, which involves morphological processing to eliminate mask holes and connected component filtering to select the connected part.

Granularity Quantification. The granularity refers to the level of detail in the segmentation of objects. Fine-grained masks furnish rich internal details and part boundaries, while coarse-grained masks provide more general object representations. To endow the IS model with rational granularity controllability, it is necessary to quantify the granularity consistent with human cognition for each proposal. Specifically, we consider the granularity quantification from both semantic and scale perspectives. Semantic granularity is estimated based on the image content covered by the mask, while scale granularity is based on the ratio of the mask in the area to the entire object. The rationality can be explained as follows. The head of the cat is larger in scale than the crane, as the head accounts for a larger proportion of the cat, but semantically, the two have similar granularity. On the contrary, the feline body can be divided into different granularities, such as individual limbs or specific left and right limbs. Although these two manners possess semantic equivalence, they differ in scale granularity.

Granularity Estimator. The granularity estimator is responsible for quantifying the granularity of each proposal $\boldsymbol{P}_j^i \in \{0,1\}^{h\times w}$, where $\boldsymbol{P}_j^i$ represents the i-th part of object j. We calculate the scale and semantic granularity for each proposal respectively. The former is directly calculated by dividing the area of the part proposal $\boldsymbol{P}_j^i$ by the corresponding instance mask $\boldsymbol{G}_j$, and the latter is calculated based on the probability map predicted by the pre-trained IS model. Specifically, IS model predicts the probability that each pixel belongs to the foreground, and then uses a preset threshold to obtain the binarized mask. As the threshold increases, the mask shrinks to parts of different scales. Therefore, we calculate the semantic granularity by the ratio of peak difference $(\max(\boldsymbol{M}_p) - \min(\boldsymbol{M}_p))/(\max(\boldsymbol{M}_g) - \min(\boldsymbol{M}_g))$, where $\boldsymbol{M}$ is the probability map obtained from the pre-trained IS model with a positive click at the center of the mask, $\boldsymbol{M}_p$ and $\boldsymbol{M}_g$ is the probabilities within the proposal $\boldsymbol{P}_j^i$ and the corresponding instance mask $\boldsymbol{G}_j$. Formally, the probability map is calculated by Eq. (3), and the calculation rules for scale and semantic granularity are shown in Eq. (4) and Eq. (5).

$$\boldsymbol{M}_j^i = \mathcal{F}(\text{Fusion}(\boldsymbol{I}, \boldsymbol{D}_j^i, \boldsymbol{G}_j)),\ \boldsymbol{M}_j^i \in \mathbb{R}^{h\times w}, \quad (3)$$

$$\mathcal{G}_{scale}^{i,j} = Area(\boldsymbol{P}_j^i)\ /\ Area(\boldsymbol{G}_j), \quad (4)$$

$$\mathcal{G}_{semantic}^{i,j} = \psi(\boldsymbol{M}_j^i, \boldsymbol{P}_j^i)\ /\ \psi(\boldsymbol{M}_j^i, \boldsymbol{G}_j), \quad (5)$$

where $Area(\cdot)$ represents the mask area, $\psi(\cdot,\cdot)$ represents the peak difference. Finally, the granularity of the proposal $\boldsymbol{P}_j^i$ is calculated as a linear combination as:

$$\mathcal{G}^{i,j} = (1-\lambda)\cdot\mathcal{G}_{scale}^{i,j} + \lambda\cdot\mathcal{G}_{semantic}^{i,j}, \quad (6)$$

where λ represents the weight coefficient, which is set to 0.5 in the experiments.

3.3. Granularity-Controllable Learning

Granularity Embedding. We transform the granularity into the learnable embedding as an additional prompt to the IS model. According to Equation (4) and Equation (5), it is apparent that the granularity fall within the range of [0,1]. Therefore, we discretize the interval from 0 to 1 into

Method	Backbone	GrabCut		Berkeley		SBD		DAVIS		PascalPart	PartImageNet
		NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@85
Single-granularity Interactive Segmentation											
DIOS [47]	FCN	5.08	6.08	-	-	9.22	12.80	9.03	12.58	-	-
LD [33]	VGG-19	3.20	4.79	-	-	7.41	10.78	5.05	9.57	-	-
BRS [19]	DenseNet	2.60	3.60	-	5.08	6.59	9.78	5.58	8.24	-	-
FocalClick [7]	MiT-B0	1.66	1.90	-	3.14	4.34	6.51	5.02	7.06	-	-
FocusCut [36]	ResNet-101	1.46	1.64	1.81	3.01	3.40	5.31	4.85	6.22	-	-
PseudoClick [37]	HRNet-32	-	1.84	-	2.98	-	5.61	4.74	6.16	-	-
CDNet [6]	ResNet-34	1.86	2.18	1.95	3.27	5.18	7.89	5.00	6.89	14.95	11.96
RITM [46]	HRNet-18	1.76	2.04	1.87	3.22	3.39	5.43	4.94	6.71	10.95	9.02
GPCIS [50]	ResNet-50	1.64	1.82	1.60	2.60	3.80	5.71	4.37	5.89	10.91	8.24
SimpleClick [38]	ViT-B	1.40	1.54	1.44	2.46	3.28	5.24	4.10	5.48	10.97	8.58
SimpleClick [38]	ViT-L	1.38	1.46	1.40	2.33	2.69	4.46	4.12	5.39	10.23	8.14
SimpleClick¶ [38]	ViT-B	3.56	4.04	4.05	5.28	5.46	7.83	7.09	8.94	7.98	6.45
SimpleClick¶ [38]	ViT-L	3.86	4.48	4.43	5.66	5.59	7.92	7.62	9.46	7.28	5.81
Multi-granularity Interactive Segmentation											
SAM [22]	ViT-B	2.42	2.72	2.21	2.96	7.22	11.05	6.13	7.88	13.89	13.32
SAM [22]	ViT-L	1.86	1.96	1.84	2.42	5.99	9.52	4.94	6.48	13.15	11.94
SAM⋆ [22]	ViT-B	1.56	1.68	1.35	1.91	6.53	10.38	4.81	6.44	13.68	12.98
SAM⋆ [22]	ViT-L	1.72	1.92	1.37	2.01	5.74	9.32	5.04	6.48	13.45	12.76
Granularity-Controllable Interactive Segmentation (ours)											
GraCo w/ GT	ViT-B	1.46	1.64	1.73	2.85	3.82	5.35	5.34	7.16	6.12	6.05
GraCo w/ AGG	ViT-B	1.34	1.46	1.37	2.21	3.44	4.89	4.44	5.72	6.38	6.01
GraCo w/ GT+AGG	ViT-B	**1.24**	**1.36**	**1.33**	**2.07**	**3.22**	**4.65**	**4.36**	**5.49**	**6.08**	**5.32**
GraCo w/ GT	ViT-L	1.74	1.88	1.71	2.70	3.49	4.90	5.65	7.13	**5.81**	5.34
GraCo w/ AGG	ViT-L	1.18	1.24	1.23	1.73	2.73	3.96	4.24	5.19	6.12	5.26
GraCo w/ GT+AGG	ViT-L	**1.18**	**1.20**	**1.17**	**1.61**	**2.69**	**3.96**	**3.87**	**4.83**	6.00	**4.92**

Table 1. **Comparison with previous methods on both object and part level benchmarks.** Single-granularity IS models listed and our GraCo are trained on SBD [14] dataset, and SAM is trained on SA-1B [22]. All models listed are from official source and use specific data pre-processing pipeline. ¶ represents fine-tuning the model utilizing the part annotation. ⋆ represents selecting the best matching result from multiple predictions. For GraCo, we select the optimal granularity for each instance from 0 to 1 with a step of 0.1 to report the average NoC. **Bold** indicates the best performance and underlined the second best.

B bins and establish a table that maps the discrete granularities to high-dimensional embeddings. The prompts, including granularity, clicks and mask, are integrated with the image embedding and jointly fed into the feature extractor.

Proposal Sampling and Training. Considering the uneven granularity distribution of mask-granularity pairs generated by AGG, we formulate the sampling probability of each mask as an inversely proportional function of the ratio of the corresponding granularity in the proposal database to improve the training stability. For training, the IS model utilizes the iterative sampling strategy [38, 46]. The segmentation of the previous iteration step serves as the mask prompt for the model and we feed an empty mask for the first iteration. The iterative sampling strategy achieves a high-level of consistency in simulating the user behaviour, thereby improving performance. We take the Normalized Focal Loss (NFL) following [27, 38] for training.

LoRA Technology. We utilize LoRA technology [18] to facilitate the object-level pre-trained IS model in efficiently comprehending granularity controllability while preserving its primary performance. For the feature extractor with a weight matrix $\boldsymbol{W} \in \mathrm{R}^{d\times d}$, we maintain the $\boldsymbol{W}$ frozen while learning a new weight matrix $\boldsymbol{BA}$. Formulaically, given a feature extractor $\mathcal{E}(\cdot)$ and input $\boldsymbol{x}$, the forward process is represented as:

$$\mathcal{E}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}, \tag{7}$$

where $\boldsymbol{B} \in \mathrm{R}^{d\times r}$ and $\boldsymbol{A} \in \mathrm{R}^{r\times d}$. The rank r is typically lower than the dimension d to reduce the computational cost. For implementation, $\boldsymbol{A}$ employs Gaussian initialization while $\boldsymbol{B}$ initializes with zero, ensuring that $\boldsymbol{BA}$ is a zero matrix at the start of fine-tuning. We apply LoRA to the projection layers of $\boldsymbol{Q}$ and $\boldsymbol{K}$ in each attention block.

4. Experiments

4.1. Experimental Settings

Dataset. To demonstrate the performance of the IS model in multi-granularity scenarios, we utilize object and part level benchmarks for evaluation. For the object-level, we conduct evaluation on four commonly used benchmarks: **GrabCut** [43], **Berkeley** [40], **SBD** [14], **DAVIS** [41]. For

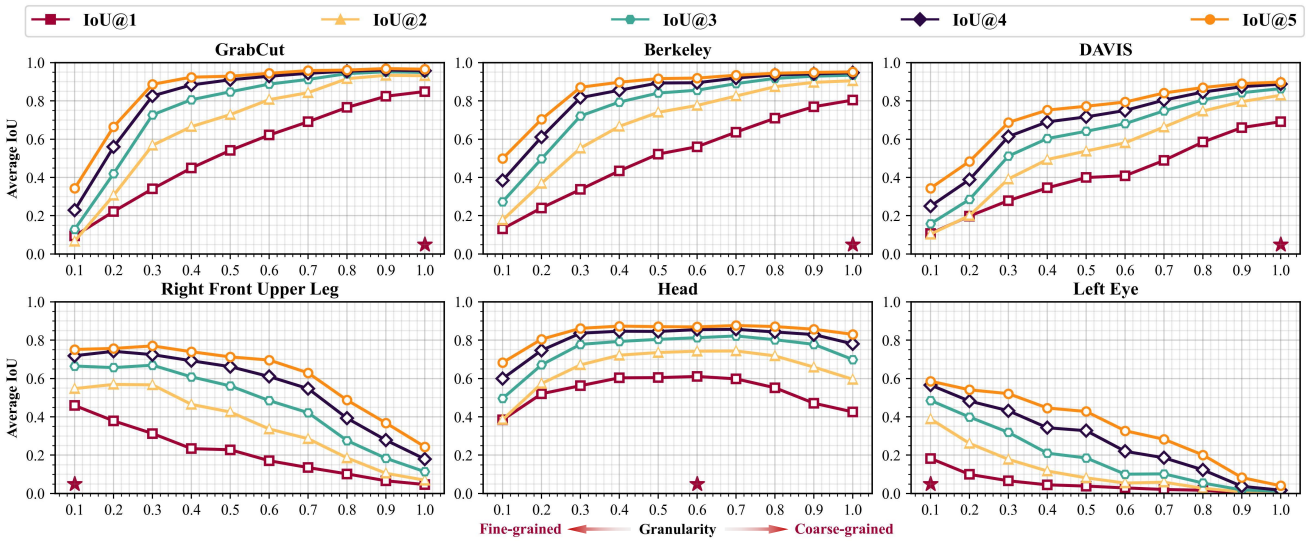

Figure 4. **Verification of the granularity controllability.** We calculate IoU@k under different granularities to plot IoU-granularity curves. The optimal granularity (marked by the red star) of the objects is about 1.0, while for the parts of the cow from PascalPart [5] it is different.

the part-level, we utilize two part segmentation datasets: **PascalPart** [5] and **PartImageNet** [15]. Note that we train our GraCo on SBD and remove samples from the PascalPart validation set that belong to the SBD training set. See the Appendix for a detailed description of these datasets.

Implementation Details. We build our GraCo based on SimpleClick [38], which consists of two patch embedding modules for image and click map respectively (we introduce an extra granularity embedding for our GraCo), a ViT [10] backbone initialized with MAE [17], a simple feature pyramid [32], and an MLP segmentation head. The IS model employed in AGG is SimpleClick with ViT-Base. The multi-granularity loop simulation iterations for each instance are randomly selected from a range of 3 to 6. For LoRA [18], the rank is set to 8 and the discretization interval for granularity is set to 0.1. We set the maximum number of iterative clicks to 3 follow [38]. We train the GraCo for 55 epochs using the Adam [21] optimizer with a learning rate of 5e-5, which decays by a factor of 10 at 50 epochs. For inference, we set the threshold for binarizing the prediction to 0.5 and use the same data augmentation as [29].

Evaluation Protocol. We conduct the evaluation following the standard protocol of previous click-based IS methods [6, 7, 36, 38, 46, 47]. Specifically, the first positive click is sampled in the center of the object, while the subsequent clicks are derived from the largest error region by comparing the current mask with the GT. For the metrics, we adopt the Number of Click (NoC) to evaluate the performance, which counts the average number of clicks required to achieve a fixed Intersection over Union (IoU), with lower values indicating better performance. We set two commonly used target IoU thresholds (85% and 90%, denoted as NoC@85 and NoC@90 respectively) and 20 clicks as the upper bound for interaction, which are same with previous works [27, 38, 46]. Moreover, the IoU-granularity curves are drawn to verify the granularity controllability of our GraCo. We also calculate the average IoU of the first click, and the results are shown in the Appendix 2.1.

4.2. Main Results and Analysis

Comparison with Previous Method. We compare our results with previous single and multiple granularity IS methods on four object-level benchmarks and two part-level benchmarks. Note that we report NoC@85 and NoC@90 for the object-level benchmarks and only NoC@85 for the part-level benchmarks. The reason is that multi-granularity parts are more difficult to segment than objects. As a result, it is challenging to achieve an IoU of up to 90% within 20 clicks. The experimental results are shown in Table 1. We present the results of single-granularity models equipped with different backbones trained on SBD [14], alongside the results of the multi-granularity model (*i.e.*, SAM) trained on SA-1B [22]. We utilize the official models and retain their specific data pre-processing pipeline for evaluation. For our GraCo, we present the performance using the mask proposals generated by our AGG (denoted as GraCo w/ AGG in Table 1). Based on the results, single-granularity IS methods show satisfactory performance in object-level benchmarks, but poor performance in handling the part-level, and the multi-granularity method perform poorly at both levels. In contrast, our GraCo w/ AGG achieves superior performance on all benchmarks at both levels.

In addition, we fine-tune SimpleClick [38] and our GraCo utilizing the training set of SBD [14] with part annotations from PascalPart (denoted as SimpleClick¶ and GraCo w/ GT). The results of SimpleClick¶ indicate that

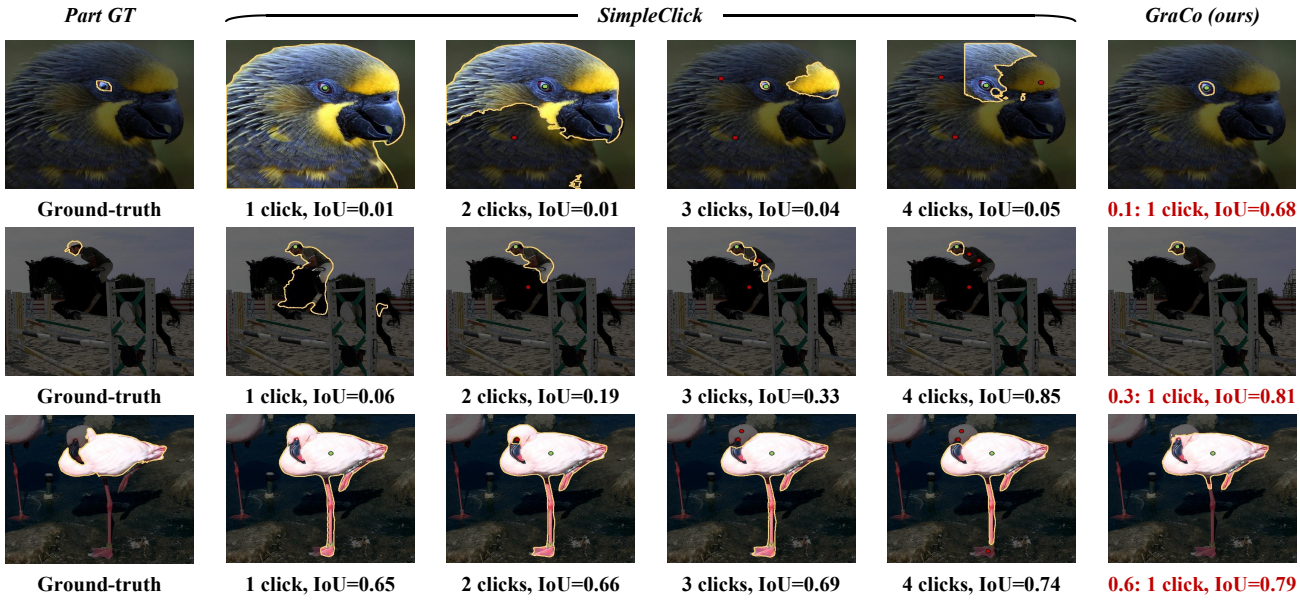

Figure 5. **Visualization of interactive segmentation on part GT using SimpleClick [38] and our GraCo.** We note the input granularity for our GraCo, which is roughly estimated based on human cognition.

fine-tuning the model with part annotations not only weakens the object-level segmentation performance, but also achieves a marginal improvement at the part-level. However, our GraCo w/ GT using the proposed GCL strategy achieves significant performance improvements over vanilla SimpleClick, demonstrating the effectiveness of GCL.

Failure Analysis of SAM. SAM [22], a representative of multi-granularity IS methods, does not achieve ideal results in Table 1, which is below our expectations. Upon our analysis, we find that SAM has a bias towards segmenting small components on object-level benchmarks even when producing multiple masks. This factor causes SAM to require more clicks to reach the IoU thresholds, resulting in unsatisfactory NoC. Furthermore, the mask distribution of the selected part-level benchmarks deviates from its training set, exposing its limited generalization. To substantiate this claim, we evaluate the performance of SAM, SimpleClick [38], and our GraCo using the first 1000 images from the SA-1B [22] as a dedicated test subset in Table 2. Considering that each image in SA-1B contains an average of 100 masks, covering diverse granularities and overlapping, we select five non-overlapping masks for each image (selecting 4987 masks in total) for evaluation. We conclude that SimpleClick performs poorly on such a multi-granularity benchmark, while SAM achieves excellent performance because it is a subset of its training set, which is in line with our expectations. Our GraCo achieves comparable NoC@90 metrics to SAM, while significantly outperforming SimpleClick. This demonstrates the robust generalization and excellent performance of GraCo in multi-granularity segmentation. Furthermore, we also calculate

Method	Backbone	SA-1B [22]		
		NoC@85↓	NoC@90↓	IoU@1↑
SimpleClick [38]	ViT-B	5.56	7.29	0.22
SAM [22]	ViT-B	2.93	5.19	0.78
SAM⋆ [22]	ViT-B	**2.46**	4.42	**0.88**
GraCo w/ AGG	ViT-B	3.39	**4.29**	0.61
SimpleClick [38]	ViT-L	4.98	6.74	0.29
SAM [22]	ViT-L	1.99	3.31	0.81
SAM⋆ [22]	ViT-L	**1.77**	**2.97**	**0.91**
GraCo w/ AGG	ViT-L	3.10	3.96	0.65

Table 2. **Experimental results on the first 1000 images of SA-1B [22].** ⋆, **Bold** and underlined are the same as Table 1.

the IoU@1 on all benchmarks. We find that SAM achieves superior performance when producing multiple masks, providing an excellent user experience. The detailed results are shown in the Appendix 2.1.

Gains from AGG. We utilize part annotations, mask proposals generated by AGG, and the combination of both to perform the GCL strategy, corresponding to GraCo w/ GT, GraCo w/ AGG, and GraCo w/ GT+AGG in Table 1. Taking advantage of the any-granularity part proposals generated by AGG, GraCo w/ AGG performs better than GraCo w/ GT on all benchmarks except PascalPart [5]. We argue that this is due to the limited number of manual annotations and the existence of granularity variance, which cannot cover arbitrary granularities, resulting in sub-optimal generalization. In contrast, AGG automatically generates abundant any-granularity masks, thereby facilitating the IS model in capturing granularity controllability. Moreover, the results of GraCo w/ GT+AGG are superior to both GraCo w/ GT

LoRA	Granularity Embedding	GrabCut			Berkeley			SBD			PascalPart	
		NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	NoC@90↓	IoU@1↑	NoC@85↓	IoU@1↑
-	-	3.56	4.04	0.47	4.05	5.28	0.43	5.46	7.83	0.42	7.98	0.48
✓	-	3.24	3.68	0.45	3.67	4.97	0.41	4.66	6.68	0.48	8.68	0.43
-	✓	2.14	2.52	0.79	1.90	**2.78**	0.79	4.20	5.99	0.64	**5.84**	**0.59**
✓	✓	**1.46**	**1.64**	**0.86**	**1.73**	2.85	**0.80**	**3.82**	**5.35**	**0.66**	6.12	0.52

Table 3. **Results of ablation study on GCL.** We utilize SimpleClick [38] with ViT-B to train on SBD [14] with part annotations.

and GraCo w/ AGG, further demonstrating that the proposals generated by AGG offer a greater level of granularity abundance than GT and serve as an effective supplement.

Granularity Controllability Analysis. To verify the granularity controllability of our GraCo, we calculate the IoU@k at different granularities and plot the IoU-granularity curves (*cf.* Figure 4). Based on the granularity definition, 1.0 represents object-level segmentation, and the closer to 0, the finer the prediction granularity. For three object-level benchmarks, IoU@k increases with increasing granularity, especially IoU@1, which is as expected. For the part-level scenario, we randomly select three part categories belonging to the cow category for validation. For highly detailed parts such as the right front upper leg and left eye, GraCo performs optimally at a granularity of 0.1. For coarse-grained parts such as the head, the optimal granularity for GraCo is around 0.6. The part-level results further demonstrate that our GraCo possess granularity controllability consistent with human cognition.

Qualitative Results. Figure 5 shows the qualitative results using SimpleClick [38] and our GraCo on some segmentation examples. We randomly select several parts from PascalPart [5] annotations and automatically generate the next click according to the evaluation protocol. We find that SimpleClick requires multiple clicks to segment the desired mask in multi-granularity scenarios. In contrast, our GraCo requires only a single click to match expectations well based on roughly estimated input granularity. This demonstrates the flexibility of our GraCo to adapt to diverse scenarios.

4.3. Ablation Study

Granularity-Controllable Learning. To demonstrate the effectiveness of the GCL strategy, we evaluate the contributions of its two key components, *i.e.*, granularity embedding and low-rank adaptation. Specifically, we conduct experiments including removing granularity embedding, removing LoRA (*i.e.*, full parameter fine-tuning), and removing both simultaneously (*cf.* Table 3). We conclude that incorporating granularity embedding effectively enhances the performance, whereas the LoRA technology preserves the original performance of the pre-trained model. More detailed ablation studies are provided in Appendix 2.2.

Granularity Definition. To demonstrate the necessity of both semantic and scale granularity, we conduct experiments in two settings: with scale granularity only, and with both scale and semantic granularity. We plot a histogram and line graph to display the frequency distribution of optimal granularity on two object-level benchmarks, *i.e.*, DAVIS [41] and SBD [14], Figure 6. We conclude that the optimal granularity tends to be skewed to 1.0 when employing both scale and semantic granularity. This aligns with the granularity definition for the whole instance. Moreover, we quantitatively evaluate the performance of the two settings on part-level benchmarks in Appendix 2.2, which demonstrates the necessity of the two types of granularity.

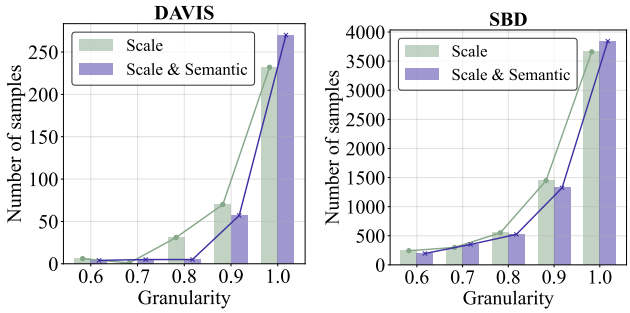

Figure 6. **Frequency distribution of optimal granularity.**

5. Conclusion

In this work, we propose a novel paradigm for interactive segmentation that allows users to control the segmentation granularity to resolve ambiguity. Our GraCo fine-tunes the pre-trained IS model to endow it with granularity controllability without requiring additional manual annotation, providing a non-redundant, low-cost and highly flexible solution to address spatial ambiguity. Excellent experimental results demonstrate the effectiveness and generalization of our method, and the granularity controllability analysis confirms the consistency of the model with human cognition. We hope that our exploration will open up new avenues for resolving ambiguity in pixel-level interactive AI systems.

Acknowledgements. This work was supported in part by the National Key R&D Program of China (No. 2022ZD0118201), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465), and the Shenzhen Medical Research Funds in China (No. B2302037).

References

[1] Junjie Bai and Xiaodong Wu. Error-tolerant scribbles based interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 392–399, 2014. 1

[2] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[3] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004. 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[5] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014. 2, 3, 6, 7, 8

[6] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7345–7354, 2021. 1, 2, 5, 6

[7] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. FocalClick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 1, 2, 5, 6

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[9] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 3

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 6

[11] Leo Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006. 1, 2

[12] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3129–3136. IEEE, 2010. 2

[13] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in Neural Information Processing Systems*, 25, 2012. 3

[14] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 2, 5, 6, 8

[15] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 2, 3, 6

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 5, 6

[19] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 1, 2, 5

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 5, 6, 7

[23] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 277–284. IEEE, 2009. 1

[24] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 3

[25] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 3041–3050, 2023. 3

[26] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *IEEE Transactions on Image Processing*, 2023. 3

[27] Kun Li, George Vosselman, and Michael Ying Yang. Interactive image segmentation with cross-modality vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 762–772, 2023. 2, 5, 6

[28] Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, and Jie Chen. Acseg: Adaptive conceptualization for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7162–7172, 2023. 3

[29] Kehan Li, Yian Zhao, Zhennan Wang, Zesen Cheng, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Multi-granularity interaction simulation for unsupervised interactive segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

[30] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-partformer: Learning a unified model for panoptic part segmentation. In *European Conference on Computer Vision*, pages 729–747. Springer, 2022. 3

[31] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Lazy snapping. *ACM Transactions on Graphics (ToG)*, 23 (3):303–308, 2004. 1

[32] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022. 6

[33] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018. 3, 5

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[35] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13339–13348, 2020. 1, 2

[36] Zheng Lin, Zheng-Peng Duan, Zhao Zhang, Chun-Le Guo, and Ming-Ming Cheng. FocusCut: Diving into a focus view in interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 1, 2, 5, 6

[37] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. Pseudoclick: Interactive image segmentation with click imitation. In *European Conference on Computer Vision*, pages 728–745. Springer, 2022. 5

[38] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22290–22300, 2023. 1, 2, 5, 6, 7, 8

[39] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023. 3

[40] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010. 2, 5

[41] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016. 2, 5, 8

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[43] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3): 309–314, 2004. 2, 5

[44] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019. 3

[45] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 1, 2

[46] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 2, 5, 6

[47] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 373–381, 2016. 1, 2, 5, 6

[48] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. In *Procedings of the British Machine Vision Conference*. British Machine Vision Association, 2017. 1

[49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[50] Minghao Zhou, Hong Wang, Qian Zhao, Yuexiang Li, Yawen Huang, Deyu Meng, and Yefeng Zheng. Interactive segmentation as gaussion process classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19488–19497, 2023. 2, 5