Toward Robust In-Context Segmentation via Concept Guidance

Zhigang Chen[1], Xiawu Zheng[1], and Rongrong Ji[1⋆]

Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R. China
chenzg11@stu.xmu.edu.cn, {zhengxiawu, rrji}@xmu.edu.cn

Abstract. In-context segmentation (ICS) requires a model to segment target regions in a query image using only a few reference images and their corresponding masks, without updating any parameters. Despite recent progress, prior ICS studies have largely overlooked a critical aspect: *system robustness*, *i.e.*, whether the model can produce stable segmentation results for the same query under different references. In this work, we revisit ICS from the robustness perspective and introduce a novel paradigm, Concept-Guided In-Context Segmentation (CG-ICS), which performs segmentation by extracting high-level semantic concepts from references rather than relying solely on low-level visual matching. Specifically, CG-ICS introduces a concept reasoning module that uses an MLLM to propose candidates and a SAM3-driven scoring function with tree-search refinement to select reliable textual concepts, together with a parallel visual exemplar route that provides query-side spatial grounding via a simple context construction. Both the textual concept and the visual exemplar are then used to activate the segmentation capability of a frozen SAM3 backbone. Extensive experiments on standard ICS benchmarks demonstrate that CG-ICS not only achieves state-of-the-art accuracy but also substantially improves robustness, yielding a more reliable ICS system with significantly reduced variance across diverse reference choices.

1 Introduction

Image segmentation stands as a fundamental task in computer vision. However, traditional fully supervised approaches are restricted by their heavy reliance on costly annotations and limited generalization to novel categories. Recently, inspired by the success of In-context Learning in Large Language Models (LLMs) [6], In-Context Segmentation (ICS) [5, 35, 36] has emerged to address these challenges. ICS aims to segment the target region in a query image using only a few reference examples, without updating any model parameters. It offers a flexible and efficient solution for open-world segmentation.

While recent ICS methods have made notable progress in segmentation accuracy [22–24, 36, 40, 41], they largely overlook an equally critical aspect: system

⋆ Corresponding author.

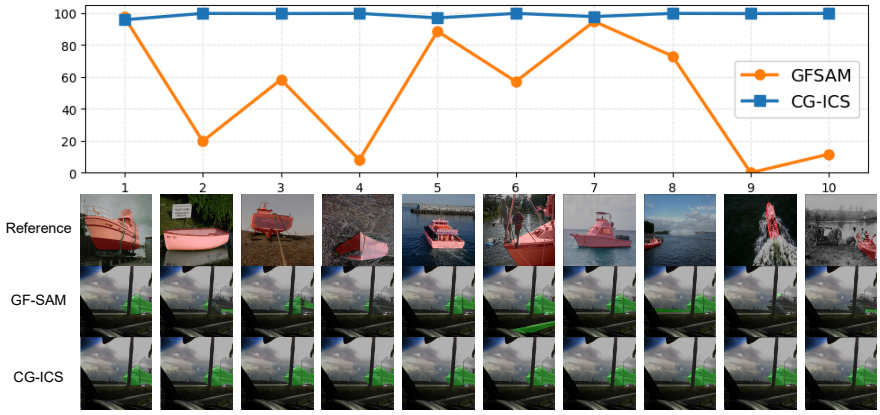


Fig. 1: We present the qualitative results of a query sample with multiple distinct reference examples and report the IoU achieved by GF-SAM [40] and our CG-ICS.

robustness, i.e., whether the model can produce stable segmentation results for the same query under different references. Due to the few-shot nature of ICS, where predictions are guided by only a handful of in-context examples, they are highly sensitive to the choice of reference instances. To make this issue concrete, we probe a state-of-the-art training-free method, GF-SAM [40], by varying the in-context references while keeping the query fixed. As shown in Fig. 1, GF-SAM exhibits pronounced variance across different references and can even fail catastrophically when paired with unfavorable reference instances. Recently, several works [30, 38, 42] have also exposed this sensitivity, but they mainly focus on selecting the best prompt from an available pool of candidate references. However, such a candidate pool is often absent in real deployments, where users provide arbitrary references on the fly, and the system has no opportunity to search for a better alternative. UNICL-SAM [28] also recognizes the robustness issue and attempts to incorporate the uncertainty of reference samples into the training process. However, it only focuses on the model's performance when the reference samples are corrupted, without considering references that are inherently low quality. Therefore, a fundamental requirement emerges: an ICS system should remain accurate and stable under any reasonable reference, rather than performing well only when paired with a fortunate one.

In this paper, we take robustness as a first-class objective for ICS, and aim to reduce the prediction variance induced by reference choice while preserving strong average segmentation quality. Our methodology is directly inspired by the rapid evolution of segmentation foundation models [7, 19, 26]. In particular, SAM3 [7] introduces a Promptable Concept Segmentation (PCS) paradigm, where the model can effectively interpret a textual concept prompt and produce accurate segmentation masks accordingly. We posit that such a concept-guided formulation offers a promising direction for advancing ICS, as textual concepts provide a higher-level and more invariant representation of the target than brit-

tle low-level visual correspondences. Consequently, concept-based guidance is expected to yield more robust and stable predictions under varying references.

However, adapting PCS to the ICS task presents a unique challenge: the standard ICS setting strictly provides only a reference image and a binary mask, lacking any prior category names or textual descriptions. In that case, our primary objective is to autonomously derive these concepts. The most straightforward solution is to leverage a model that can understand images and output textual descriptions. Therefore, we adopt a Multimodal Large Language Model (MLLM) [2–4, 32, 34], as the concept generator, which not only meets the above requirement but also exhibits strong open-world generalization capabilities to resolve visual ambiguities and improve robustness. Since the MLLM and SAM3 are decoupled, a key challenge lies in selecting the optimal concept that can best prompt SAM3 for accurate segmentation. Therefore, as illustrated in Fig. 2, we design a tree-search-based Concept Reasoning procedure. Specifically, the MLLM-driven Concept Generation module expands tree branches by producing multiple candidate concepts as tree nodes. For each node, a SAM3-based Concept Scoring module evaluates the candidate by jointly considering the reference and the query. The resulting scores determine whether the current concept should be expanded (to generate refined child nodes) or pruned. This iterative expand–score–prune process continues until a predefined stopping criterion is met, yielding the best concept for prompting SAM3. However, relying solely on linguistic concepts remains insufficient in challenging cases. SAM3 also highlights that providing visual exemplars (bounding boxes of the target) can help specify uncommon categories [7]. Therefore, we introduce a parallel Visual Exemplars Extraction route that extracts query-side visual exemplars from the concatenated reference-query image. Finally, CG-ICS fuses the selected textual concept with the visual exemplars in a single SAM3 inference to produce the query mask. We term this framework **Concept-Guided In-Context Segmentation (CG-ICS)**. Overall, CG-ICS effectively recasts ICS as a PCS problem, leading to not only improved segmentation accuracy but also substantially enhanced robustness to reference selection, as evidenced in Fig. 1.

Our contributions and the main findings are summarized as follows:

- We tackle a critical yet underexplored issue in ICS, namely the pronounced sensitivity to in-context references. To address this problem, we make the first attempt to build a robust ICS system that maintains stable performance under diverse and low-quality references, rather than focusing on selecting an optimal reference as in prior work.
- The proposed CG-ICS introduces a concept reasoning module to extract the most suitable textual concept, together with a parallel visual exemplar route that provides query-side spatial grounding. With this design, we, for the first time, successfully bring the powerful SAM3 into the ICS.
- Through extensive experiments on four ICS benchmarks under both standard and robust evaluation protocols, we demonstrate that CG-ICS achieves strong segmentation performance while substantially reducing variance across

diverse reference selections and quality, offering practical insights for building more reliable ICS systems.

2 Related work

In-Context Segmentation. In-context segmentation (ICS) has emerged as a practical variant of visual in-context learning, where a model segments the target region in a query image conditioned on a few reference image–mask pairs without updating parameters. Since Bar et al. [5] first introduced the in-context learning concept into vision, this paradigm has rapidly attracted extensive attention and led to a series of generalist segmentation systems. Existing ICS methods can be broadly categorized into training-based and training-free paradigms. Training-based methods explicitly learn the ICS paradigm by jointly feeding the in-context reference and the query into a unified model during training. Painter [35] and SegGPT [36] formulate ICS via masked image modeling [17] on the concatenated image of reference and query. Another line of works, including SegIC [24], SINE [22], and Iris [13], freeze a powerful vision foundation model [8, 25] to extract dense correspondences and train a lightweight decoder to predict masks. Diffusion-based [18, 27] approaches such as DiffewS [43] and LDIS [33] explore generative modeling as a vehicle for in-context segmentation. In addition, some methods adapt promptable segmenters by fine-tuning the SAM [19, 26] family for ICS, such as VRP-SAM [29] and SANSA [10]. Despite strong performance, training-based ICS typically requires substantial in-domain training data and may exhibit limited generalization when faced with novel domains, corrupted annotations, or distribution shifts. Training-free methods avoid parameter updates by extracting visual correspondences between reference and query and converting them into prompts for a frozen promptable segmenter (e.g., SAM). PerSAM [41], Matcher [23], and GF-SAM [40] follow this paradigm by deriving point or region cues from reference-query matching to drive mask prediction. While these approaches primarily focus on improving average segmentation accuracy, the robustness of ICS systems under diverse and imperfect reference choices remains underexplored. A few studies [30, 42] have noticed the sensitivity of ICS to reference selection, but they mainly aim to retrieve the best reference from an available pool rather than ensuring stable behavior under arbitrary references. In contrast, our work targets high performance and low variance across diverse reference examples, which better matches real-world usage where reference quality and choice cannot be guaranteed.

Robust Segmentation. Robust segmentation [1, 9, 11, 14, 37, 39] studies examine how dense prediction models behave under distribution shifts and real-world degradations. More recently, robustness has been explored in few-shot segmentation and in-context segmentation settings, where the prediction can be highly sensitive to the choice and quality of reference examples. Tang *et al.* [31] estimate the contribution of each reference image and then enhance or prune low-utility references to improve few-shot segmentation performance. UNICL-SAM [28] reduces the performance drop under corrupted references by injecting uncertainty

signals during training, which makes the model less brittle to degraded references. However, it does not explicitly evaluate stability across different references, and its training-based paradigm also incurs higher data and computation requirements. In contrast, our method adopts a training-free design that leverages advanced foundation models to perform ICS, offering a more flexible and lightweight solution while directly targeting robustness to reference variation.

3 Method

We present Concept-Guided In-Context Segmentation (CG-ICS), a framework that performs ICS without updating any model parameters. Built on MLLM and SAM3, CG-ICS derives textual concepts and visual exemplars and combines them as prompts for final mask prediction. We first introduce the ICS problem setup and the two foundation models in preliminaries, and then present CG-ICS in detail. The pipeline is described in a one-shot setting for clarity, and then extended to multiple reference examples.

3.1 Preliminaries

ICS. Given a single user-provided reference example (I_r, M_r) and a query image I_q, ICS aims to produce a mask M_q that captures the content in the query image. Only the reference image and mask are provided, and no category label or predefined class information is assumed.

MLLM. A MLLM is a generative model that jointly processes visual inputs and natural language, enabling unified perception and language reasoning. MLLMs are known to be robust in handling diverse scenes and open-vocabulary descriptions, and can generalize well under distribution shifts and noisy visual conditions. Formally, given an image (or a set of images) and a textual instruction prompt, an MLLM produces a textual response as

$$T = \mathcal{M}_{\mathrm{MLLM}}(I, P)\,, \tag{1}$$

where $\mathbf{I}$ denotes the visual input, $\mathbf{P}$ denotes the input prompt, and $\mathbf{T}$ is the generated text output.

SAM3. SAM3 is a foundation model that performs promptable concept segmentation. The target can be specified by a short noun phrase (Textual Concept), a visual exemplar bounding box, or both. Given a concept prompt, SAM3 predicts semantic and instance-level masks for all instances matching the concept. The model also includes a presence head to indicate whether the queried concept is present in the input image. Formally, we write SAM3 inference as

$$\left(M^{\mathrm{ins}},\, M^{\mathrm{sem}},\, S^{\mathrm{pres}}\right) = \mathcal{M}_{\mathrm{SAM3}}(I,\, T,\, V)\,, \tag{2}$$

where I is the image, T is a textual concept prompt, V is the visual exemplar boxes, and the outputs are instance masks M^{ins}, a semantic mask M^{sem}, and a presence score S^{pres}.

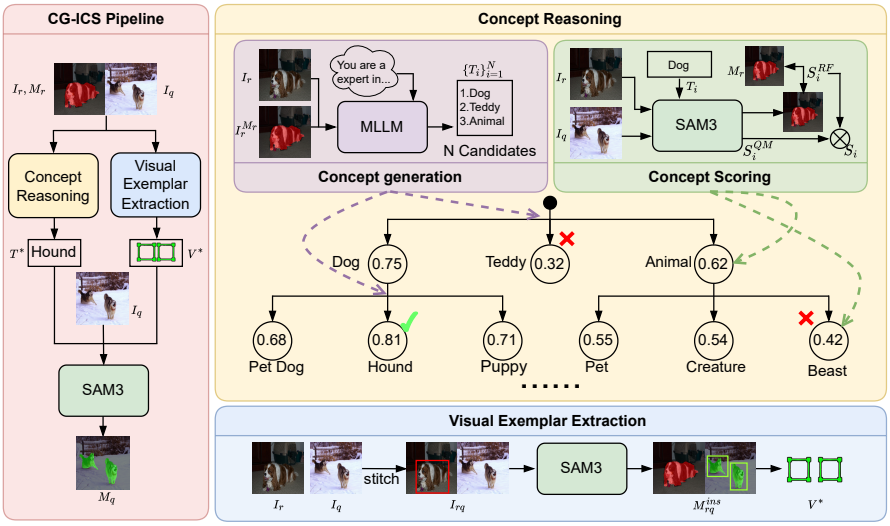


Fig. 2: Overview of the proposed CG-ICS framework. Given a reference image and mask (I_r, M_r) together with a query image I_q, CG-ICS performs concept reasoning and visual exemplar extraction in parallel. The concept reasoning formulates concept selection as a score-guided tree search, where the MLLM performs concept generation to expand branches and SAM3 serves as the scoring function to evaluate nodes, yielding the best-scoring concept for the target. Meanwhile, the visual exemplar extraction uses SAM3 to obtain a visual prior from the reference–query pair. Finally, SAM3 fuses the selected concept with the visual exemplars to predict the target mask.

3.2 Concept Reasoning

We propose a tree-search-based reasoning strategy to automatically extract the most suitable concept for SAM3 from the in-context references. The problem is formulated as searching over a concept tree, where each node corresponds to a candidate textual concept and each edge denotes a refinement from a parent concept to a more specific one.

Concept Generation. We extract candidate textual concepts from the reference image–mask pair using a frozen MLLM, which acts as the branch expansion operator in our concept tree. To preserve global scene context while highlighting target details, we provide the MLLM with a two-view visual input. The first view is the original reference image I_r. The second view is a highlighted image $I_r^{\mathrm{M_r}}$, obtained by overlaying the masked region in red on I_r. This two-view design retains full-image context from I_r while enforcing fine-grained understanding of the target object, producing more discriminative branches for subsequent search. The MLLM was then queried with an instruction prompt P_{gen} that follows three rules. (1) It requests a short noun phrase to match the textual concept input required by SAM3. (2) It explicitly asks the model to focus on the red-highlighted

masked region. (3) If a parent-node concept is provided, it instructs the model to propose either synonymous alternatives or more fine-grained refinements conditioned on that concept, enabling progressive expansion along the tree. Formally, we generate the candidates set as

$$\{T_i\}_{i=1}^{N} = \mathcal{M}_{\mathrm{MLLM}}\big(I_r,\, I_r^{\mathrm{M_r}},\, P_{\mathrm{gen}}\big)\,, \tag{3}$$

where $\{T_i\}_{i=1}^{N}$ denotes the resulting concept candidates.

Concept Scoring. Given a candidate concept T_i, we design a SAM3-based scoring function that serves as the *node evaluation* module in our concept tree. This score quantifies how well T_i (i) explains the annotated reference target and (ii) transfers to the query image. We first define Reference Fidelity (RF) score to measure whether the concept accurately describes the reference target indicated by M_r. For each T_i, we prompt it to SAM3 on the reference image I_r and obtain the semantic mask $M_{r,i}^{\mathrm{sem}}$:

$$M_{r,i}^{\mathrm{sem}} = \mathcal{M}_{\mathrm{SAM3}}(I_r,\, T_i,\, \varnothing)^{\mathrm{sem}}\,. \tag{4}$$

It was then compared with the ground-truth reference mask M_r by computing their IoU. We use the IoU as the RF score:

$$S_i^{\mathrm{RF}} = \mathrm{IoU}\big(M_{r,i}^{\mathrm{sem}},\, M_r\big)\,. \tag{5}$$

Besides, the concept should also match the to-be-segmented part in the query image. We prompt SAM3 with the same T_i on the query image I_q and use the presence score as the Query Matchability (QM) score:

$$S_i^{\mathrm{QM}} = \mathcal{M}_{\mathrm{SAM3}}(I_q,\, T_i,\, \varnothing)^{\mathrm{pres}}\,. \tag{6}$$

Finally, we combine the two criteria with a multiplicative rule to obtain the concept score:

$$S_i = S_i^{\mathrm{RF}} \cdot S_i^{\mathrm{QM}}. \tag{7}$$

Tree Search for Concept Refinement. Based on Concept Generation and Concept Scoring, we perform a score-guided tree search to refine textual concepts. We maintain a frontier $\mathcal{F}^{(k)}$ at round k, and initialize the search by expanding the root node with Concept Generation to obtain N first-layer candidates. For each node concept $T_i^k \in \mathcal{F}^{(k)}$, we compute its score S_i using Concept Scoring. We prune nodes with $S_i^k < \tau_{\mathrm{pruned}}$, and only expand the remaining ones. Each surviving concept node will be passed into Concept Generation again to produce N children concepts, including near-synonyms and finer-grained refinements. We maintain a concept buffer $\mathcal{B}$ that stores all visited concepts to prevent duplicate generation during expansion. The resulting children form the next frontier $\mathcal{F}^{(k+1)}$, and the expand–score–prune loop repeats up to K rounds. The looping will be early stopped if any node reaches $S_i^k \geq \tau_{\mathrm{stop}}$. If all nodes in a round

are fully pruned, the expansion will restart by returning to the parent node to generate totally different concepts. Finally, we select the best concept with the highest score in $\mathcal{B}$ by

$$T^* = \arg\max_{\mathcal{B}} S_i. \tag{8}$$

3.3 Visual Exemplar Extraction

Relying on textual concepts alone can be unreliable, since MLLMs may hallucinate or output over-specific phrases that do not correspond to the true target. Moreover, SAM3 highlights that providing visual exemplars (bounding boxes of the target) can help specify uncommon categories. However, the standard SAM3 exemplar prompt is defined within a single image and does not directly support cross-image exemplar transfer. Inspired by SegGPT [36], which stitches the reference and query in a single input, we propose a simple yet effective way to derive a visual exemplar of the query image.

We first form a stitched image by concatenating the reference and query images horizontally:

$$I_{rq} = \mathrm{Stitch}(I_r,\, I_q). \tag{9}$$

The reference exemplar boxes can be extracted from the reference mask.

$$V_r = \mathrm{BBox}(M_r). \tag{10}$$

Next, we use V_r as a visual exemplars prompt to invoke SAM3 on the stitched image.

$$M_{rq}^{\mathrm{ins}} = \mathcal{M}_{\mathrm{SAM3}}(I_{\mathrm{rq}},\, \varnothing,\, V_r)_{\mathrm{ins}}\,. \tag{11}$$

Since SAM3 can segment all matched instances under a prompt, it can return masks on both the reference side and the query side within I_{rq}. We take the instance mask on the query half, denoted as $M_q^{\mathrm{ins}} = \mathrm{RightHalf}(M_{rq}^{\mathrm{ins}})$, and convert it into coarse target exemplar boxes:

$$V^* = V_q = \mathrm{BBox}\big(M_q^{\mathrm{ins}}\big)\,. \tag{12}$$

The resulting V^* serves as the visual exemplars, which can be subsequently used together with the textual concept to stabilize the final segmentation.

3.4 Segmentation with Joint Prompts

After selecting the best textual concept T^* and deriving the target exemplar boxes V^*, we feed them jointly to SAM3 for the final segmentation. The textual concept provides category-level semantics, while the visual exemplars constrain the spatial reference on the query image. Formally, we perform the final inference on the query image as

$$\Big(M_q^{\mathrm{ins}},\, M_q^{\mathrm{sem}},\, S_q^{\mathrm{pres}}\Big) = \mathcal{M}_{\mathrm{SAM3}}(I_q,\, T^*,\, V^*)\,. \tag{13}$$

We take the semantic mask M_q^{sem} as the final prediction M_q.

3.5 Multiple-Reference Setting

When multiple reference examples $\{(I_r^n, M_r^n)\}_{n=1}^m$ are available, we extend CG-ICS to the multiple-shot setting. For textual concepts, due to the MLLM's limited ability to reliably reason over many images at once, we perform concept reasoning on each reference independently and obtain one final concept per reference, denoted as $\{T^*\}_{n=1}^m$. We then compare these m concepts and select the best one by re-scoring each of them on all reference image-mask pairs to measure cross-reference consistency. For visual exemplar extraction, we stitch each reference with the query separately to obtain query-side bboxes from each reference, and then collect all such bboxes as visual exemplars.

4 Experiments

4.1 Datasets and Protocol

Datasets. To demonstrate semantic ICS capability, we evaluate CG-ICS on four representative ICS benchmarks, including Pascal-5^i, COCO-20^i, LVIS-92^i, and FSS-1000. **Pascal-5^i** is built upon Pascal VOC 2012 [12] and SDS [16]. It contains 20 object categories that are split into 4 folds, with 5 classes per fold. **COCO-20^i** is derived from MS COCO [21] and includes 80 categories in total. The 80 classes are partitioned into 4 folds, with 20 classes per fold. **LVIS-92^i** [23] is built based on LVIS [15] which is more challenging. It selects 920 categories with more than two images per class and divides them into 10 folds. **FSS-1000** [20] contains 1000 categories that are split into training, validation, and test containing 520, 240, and 240 categories, respectively.

Protocol. Three types of experiments are conducted. **(1) Standard ICS** evaluates overall segmentation performance as in prior ICS work, and thus we follow the setting of Matcher [23] for fair comparison. For Pascal-5^i and COCO-20^i, we sample 1,000 queries per fold from the validation split, for LVIS-92^i we sample 2,300 queries per fold, and for FSS-1000 we use all queries in the official test split. We evaluate both 1-shot and 5-shot settings on all datasets and report the mean Intersection-over-Union *(mIoU)* as the metric. **(2) Robustness to Reference Selection.** For each fold on Pascal-5^i, COCO-20^i, and LVIS-92^i, we sample 500 query images and pair each query with 50 different reference examples from the same class. In addition to the mean IoU, the Standard Deviation *(Std)* and the Coefficient of Variation $CV = \frac{Std}{mIoU}$ across these references are also reported. The lower values of Std and CV indicate better relative stability. **(3) Robustness to Corrupted References.** Following UNICL-SAM [28], we apply diverse corruptions to the reference examples and report the $mIoU$ with clean data as well as the performance drop under corruptions. This evaluation is conducted on 1500 sampled instances from COCO-20^i.

Table 1: Comparison with state-of-the-art methods under the **Standard ICS** protocol. Gray indicates methods trained on in-domain datasets that include the test categories. **Bold** denotes the best result, and underline denotes the second best.

Methods	Pascal-5^i		COCO-20^i		LVIS-92^i		FSS-1000	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Training-based								
SegGPT [36]	83.0	**89.8**	56.1	67.9	18.6	25.4	85.6	89.3
SINE [22]	85.4	86.2	64.5	66.1	31.2	35.5	–	–
DiffwS [43]	88.3	87.8	71.3	72.2	31.4	35.4	87.8	88.0
UNICL-SAM [28]	-	-	**77.8**	**78.7**	34.1	37.1	84.0	86.3
SANSA [10]	-	-	75.6	78.6	50.3	**59.0**	90.0	**91.0**
Training-free								
Matcher [23]	68.1	74.0	52.7	60.7	33.0	40.0	87.0	89.6
GF-SAM [40]	72.1	82.6	58.7	66.8	35.2	44.2	88.0	88.9
CG-ICS (ours)	**89.3**	89.6	72.3	72.6	**55.4**	56.2	**90.2**	89.3

4.2 Implementation Details.

We adopt Qwen3-VL-4B [3] as the MLLM backbone for concept generation. For scoring and mask prediction, the official SAM3 [7] release is adopted as the promptable segmentation backbone. All model parameters are frozen. For the tree search, each successful node expands to N=5 refined concept candidates generated by the MLLM. A node is pruned if its score is smaller than τ_{pruned}=0.5. We enable early stopping when any node reaches τ_{stop}=0.8 to reduce computation. The search runs for at most K=3 refinement rounds.

4.3 Comparison with state-of-the-art.

Standard ICS. We compare CG-ICS with state-of-the-art training-free and training-based generalist ICS methods in Table 1. Among training-free approaches, CG-ICS delivers a clear and consistent lead across all benchmarks, e.g., +17.2 $mIoU$ on Pascal-5^i , +13.6 on COCO-20^i , and +20.2 on LVIS-92^i in 1-shot setting over the strongest training-free baseline GF-SAM [40]. Notably, despite being training-free, CG-ICS remains competitive with training-based methods trained with in-domain data that includes the test categories. CG-ICS still achieves state-of-the-art results on Pascal-5^i (89.3), LVIS-92^i (55.4), and FSS-1000 (90.2) in 1-shot setting. In the remaining settings, CG-ICS only slightly trails the best training-based competitors, e.g., 0.2 $mIoU$ on Pascal-5^i 5-shot and 3.3/6.0 $mIoU$ on COCO-20^i 1-shot/5-shot. Overall, strong performance across diverse datasets indicates that our CG-ICS is a powerful paradigm.

Robustness to Reference Selection. Table 2 reports robustness under varying in-context references. Across all three benchmarks, CG-ICS obtains the lowest CV, indicating the most stable segmentation relative to its mean performance. On Pascal-5^i and COCO-20^i, CG-ICS achieves the lowest CV (7.8% and

Table 2: Comparison results under the Robustness to Reference Selection protocol. Gray indicates methods trained on in-domain datasets that include the test categories. **Bold** denotes the best result, and underline denotes the second best.

Methods	Pascal-5^i			COCO-20^i			LVIS-92^i		
	mIoU	*Std*	*CV*	*mIoU*	*Std*	*CV*	*mIoU*	*Std*	*CV*
Training-based									
SegGPT [36]	75.1	17.3	23.0%	45.7	22.7	49.7%	19.4	**13.5**	69.6%
SINE [22]	81.7	15.8	19.3%	68.3	16.1	23.6%	30.9	20.3	65.6%
SANSA [10]	83.9	9.3	11.1%	**74.3**	13.1	17.6%	50.8	18.2	35.8%
Training-free									
Matcher [23]	65.3	20.6	31.5%	53.9	18.7	34.7%	38.8	19.2	49.4%
GF-SAM [40]	70.6	19.3	27.3%	58.3	16.4	28.1%	42.2	18.7	44.3%
CG-ICS(ours)	**88.4**	**6.9**	**7.8%**	72.1	**9.3**	**12.9%**	**57.1**	17.2	**30.1%**

Table 3: Comparison results under the robustness to Corrupted References protocol. Gray indicates methods trained on in-domain datasets that include the test categories. **Bold** denotes the best result, and underline denotes the second best.

Methods	Clean	Color	Blurriness	Compression	Space	Domain	Deformation	Mean	Ratio
Training-based									
SegGPT [36]	62.1	-2.7	-3.0	**-2.7**	-7.8	-6.2	-8.8	-5.2	-8.4%
SINE [22]	70.0	-0.7	-1.7	-5.9	-2.8	-6.5	-12.4	-5.0	-7.1%
SANSA [10]	70.5	-1.2	-4.0	-11.1	-5.8	-12.9	-1.4	-6.1	-8.6%
UNICL-SAM [28]	**79.8**	-1.0	-1.8	-3.0	-3.3	-4.8	-5.0	-3.1	-3.9%
Training-free									
Matcher [23]	41.6	-0.5	-2.5	-6.5	-3.8	-5.3	-6.2	-4.1	-9.9%
GF-SAM [40]	61.5	**-0.1**	-0.8	-5.3	**-1.9**	**-3.8**	-3.0	-2.5	-4.1%
CG-ICS(ours)	74.6	**-0.1**	**-0.3**	-4.2	-3.4	-5.2	**-0.5**	**-2.3**	**-3.1%**

12.9%), markedly improving over GF-SAM (27.3% and 28.1%) and remaining more stable than training-based baselines such as SANSA (11.1% and 17.6%). On LVIS-92^i, CG-ICS still yields the best robustness with 30.1% *CV*, outperforming SANSA (35.8%) and largely reducing the instability of SegGPT/SINE (69.6%/65.6%). Overall, the consistently smallest *CV* indicates that CG-ICS maintains strong accuracy while being the least sensitive to reference selection.

Robustness to Corrupted References. Table 3 compares robustness under six corruptions by reporting the mIoU on clean data and the mIoU drops. CG-ICS attains the best decline Ratio across all compared generalists, reaching only -3.1%, which is better than the strongest training-based baseline UNICL-SAM (-3.9%). This advantage is not achieved by sacrificing accuracy, as CG-ICS also delivers a high clean mIoU of 74.6. In contrast, other training-free methods suffer noticeably larger relative drops, e.g., GF-SAM and Matcher degrade to -4.1%

Table 4: Ablation study of the each component on COCO-20^i. The first row represents directly feeding a single concept prompt generated by the MLLM into SAM3 for segmentation. We then incrementally add candidate sampling, RF/QM scoring, tree-search, and the visual branch.

Concept Reasoning				Visual	COCO-20^i		
Candidates	RF-Score	QM-Score	Search		*mIoU*	*Std*	*CV*
					67.6	15.1	22.3%
✓					66.5	15.3	23.0%
✓	✓				69.3	13.9	20.5%
✓		✓			62.8	18.2	29.0%
✓	✓	✓			70.1	11.3	16.1%
✓	✓	✓	✓		71.2	10.5	14.7%
✓	✓	✓	✓	✓	72.1	9.3	12.9%

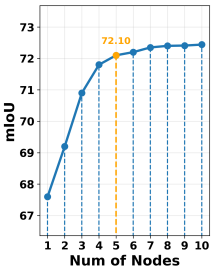
and −9.9%, respectively, while training-based SegGPT/SINE/SANSA are even more sensitive, with Ratio ranging from −7.1% to −8.6%. Across individual corruptions, CG-ICS remains nearly invariant to color and blur (both ≤ 0.3 drop) and is particularly resilient to deformation (−0.5), which jointly contributes to its lowest overall Ratio. Overall, the consistently smallest Ratio demonstrates that CG-ICS provides state-of-the-art robustness against distribution shifts, even outperforming training-based generalists.

4.4 Ablation Studies.

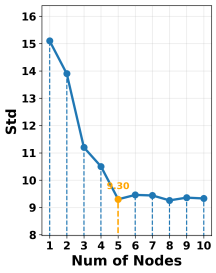
We conduct ablation studies under the 1-shot setting of Robustness to Reference Selection on COCO-20^i, which are the most widely used ICS benchmarks.

Ablation study of key components. As shown in Table 4, directly using a single MLLM-generated concept prompt yields limited performance and robustness, with 67.6 mIoU and 22.3% *CV*. Introducing multiple candidates alone brings marginal changes, while adding the proposed RF-Score and QM-Score consistently improves both accuracy and stability. In particular, combining RF and QM reduces the relative variation to 16.1% *CV* and boosts performance to 70.1 *mIoU*, indicating that concept selection is critical under varying references. The search-based selection further strengthens robustness (14.7% *CV*), and enabling the visual exemplar branch achieves the best overall trade-off, reaching 72.1 *mIoU* with the lowest dispersion (9.3 Std, 12.9% *CV*). Overall, each component contributes incrementally, and the full design is necessary to obtain both high accuracy and reliable reference-insensitive behavior.

Ablation on the search nodes and rounds. On COCO-20^i, increasing the number of search nodes (concept candidates) N steadily improves the average segmentation accuracy and enhances robustness across different references. As shown in

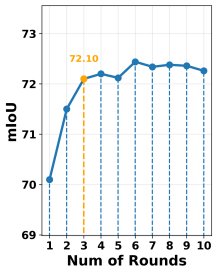


(a) Performance

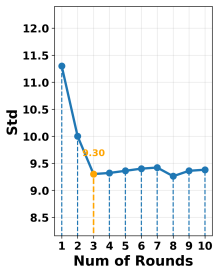


(b) Robustness

Fig. 3: Effect of the number of search nodes N on performance and robustness.



(a) Performance



(b) Robustness

Fig. 4: Effect of the number of search rounds K on performance and robustness.

Fig. 3a, mIoU increases from 67.6 at N=1 to 71.9 at N=4, and then quickly saturates around 72.1–72.4 for larger N (with 72.10 at N=5). Meanwhile, Fig. 3b shows that *Std* drops from 15.1 at N=1 to 10.5 at N=4, reaches 9.30 at N=5, and stays nearly unchanged afterwards. Overall, both curves become almost flat once $N > 5$, suggesting diminishing returns from generating more candidates. Therefore, we set N=5 in all experiments to balance performance, robustness, and inference cost. For the search rounds K, Fig. 4 exhibits the same trend, where both mIoU and Std change marginally once $K > 3$.

4.5 Qualitative Results.

We visualize qualitative comparisons between GF-SAM and CG-ICS under varying in-context references in Fig. 5. GF-SAM exhibits strong sensitivity to the reference choice, and its predictions can change noticeably as the reference varies from "worst" to "best." When the reference viewpoint shifts significantly (the first column of the top panel), it incorrectly segments the airplane tail instead of the whole target, indicating a tendency to match salient subparts rather than preserving the full object extent. When the object is partially occluded (the first column in the bottom panel), it is easily distracted by occluders and co-occurring instances, and thus tends to segment the glass rather than the bottle. Moreover, even for moderately better references, GF-SAM may still produce incomplete masks or leak to nearby regions with similar textures, suggesting inconsistent correspondence under subtle appearance changes. This also implies that GF-SAM often requires carefully selected "good" references to avoid failure cases, limiting its practicality in real-world usage. Such errors reflect the limitation of pure visual matching: correspondence is largely driven by local appearance similarity, which becomes unreliable under viewpoint changes and occlusions, leading to semantic drift and unstable boundaries. In contrast, CG-ICS remains notably stable across the same reference range and consistently outputs high-quality masks, demonstrating stronger robustness to imperfect in-context references.

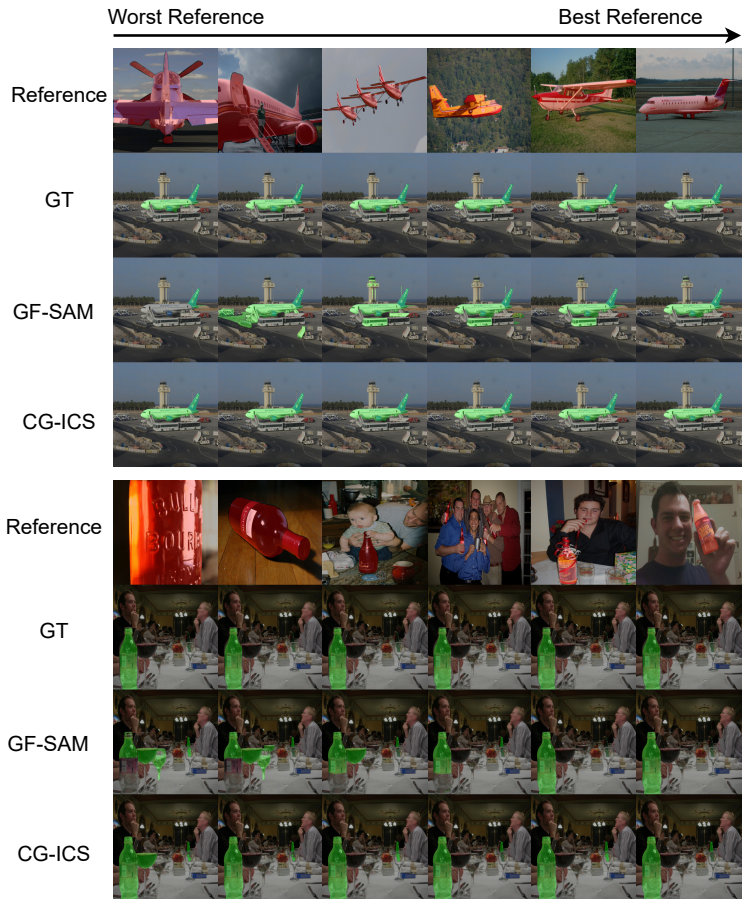


Fig. 5: Qualitative comparison of GF-SAM and CG-ICS under varying in-context references ranked from worst to best according to GF-SAM performance.

5 Conclusion

In this work, we study ICS with robustness as a first-class objective and establish evaluation protocols that measure both average performance and reference-induced variance. We introduce Concept-Guided In-Context Segmentation (CG-ICS), a training-free framework that performs concept selection through a score-guided tree search, where the MLLM performs concept generation to expand branches and SAM3 serves as the scoring function to evaluate nodes to identify a reliable textual prompt, complemented by a parallel visual exemplar route for query-side grounding. Across standard ICS benchmarks and robust evaluation protocols, CG-ICS achieves strong accuracy while substantially reducing variance, moving ICS toward more reliable real-world deployment.

Acknowledgements

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0122701), the National Natural Science Foundation of China (No. 62576299), and Fundamental Research Funds for the Central Universities and Xiaomi Young Talents Program.

References

1. Aakanksha, Rajagopalan, A.N.: Improving robustness of semantic segmentation to motion-blur using class-centric augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10470–10479 (June 2023)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., Efros, A.: Visual prompting via image inpainting. Advances in Neural Information Processing Systems **35**, 25005–25017 (2022)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2025)
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (October 2021)
9. Chen, W.T., Vong, Y.J., Kuo, S.Y., Ma, S., Wang, J.: Robustsam: Segment anything robustly on degraded images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4081–4091 (June 2024)
10. Cuttano, C., Trivigno, G., Averta, G., Masone, C.: Sansa: Unleashing the hidden semantics in sam2 for few-shot segmentation. Advances in Neural Information Processing Systems **38**, 118923–118957 (2026)

11. Endo, K., Tanaka, M., Okutomi, M.: Semantic segmentation of degraded images using layer-wise feature adjustor. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3205–3213 (January 2023)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)
13. Gao, Y., Liu, D., Li, Z., Li, Y., Chen, D., Zhou, M., Metaxas, D.N.: Show and segment: Universal medical image segmentation via in-context learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20830–20840 (June 2025)
14. Guo, D., Pei, Y., Zheng, K., Yu, H., Lu, Y., Wang, S.: Degraded image semantic segmentation with dense-gram networks. IEEE Transactions on Image Processing **29**, 782–795 (2020)
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5356–5364 (June 2019)
16. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 991–998 (2011)
17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (June 2022)
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
20. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2869–2878 (June 2020)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
22. Liu, Y., Jing, C., Li, H., Zhu, M., Chen, H., Wang, X., Shen, C.: A simple image segmentation framework via in-context examples. Advances in Neural Information Processing Systems **37**, 25095–25119 (2024)
23. Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. In: International Conference on Learning Representations. vol. 2024, pp. 17904–17925 (2024)
24. Meng, L., Lan, S., Li, H., Alvarez, J.M., Wu, Z., Jiang, Y.G.: Segic: Unleashing the emergent correspondence for in-context segmentation. In: European Conference on Computer Vision. pp. 203–220. Springer (2024)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)

26. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollar, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations. vol. 2025, pp. 28085–28128 (2025)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
28. Sheng, D., Chen, D., Tan, Z., Liu, Q., Chu, Q., Gong, T., Liu, B., Han, J., Tu, W., Xu, S., Yu, N.: Unicl-sam: Uncertainty-driven in-context segmentation with part prototype discovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20201–20211 (June 2025)
29. Sun, Y., Chen, J., Zhang, S., Zhang, X., Chen, Q., Zhang, G., Ding, E., Wang, J., Li, Z.: Vrp-sam: Sam with visual reference prompt. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23565–23574 (June 2024)
30. Suo, W., Lai, L., Sun, M., Zhang, H., Wang, P., Zhang, Y.: Rethinking and improving visual prompt selection for in-context learning segmentation. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)
31. Tang, W., Yang, B., Heng, P.A., Liu, Y.H., Fu, C.W.: Overcoming support dilution for robust few-shot semantic segmentation. arXiv preprint arXiv:2501.13529 (2025)
32. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
33. Wang, C., Li, X., Ding, H., Qi, L., Zhang, J., Tong, Y., Loy, C.C., Yan, S.: Explore in-context segmentation via latent diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7545–7553 (2025)
34. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
35. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6830–6839 (June 2023)
36. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1130–1140 (October 2023)
37. Xia, W., Cheng, Z., Yang, Y., Xue, J.H.: Cooperative semantic segmentation and image restoration in adverse environmental conditions. arXiv preprint arXiv:1911.00679 (2019)
38. Xu, C., Liu, C., Wang, Y., Yao, Y., Fu, Y.: Towards global optimal visual in-context learning prompt selection. Advances in Neural Information Processing Systems **37**, 74945–74965 (2024)
39. Yang, Z., Dong, W., Li, X., Wu, J., Li, L., Shi, G.: Self-feature distillation with uncertainty modeling for degraded image recognition. In: European Conference on Computer Vision. pp. 552–569. Springer (2022)
40. Zhang, A., Gao, G., Jiao, J., Liu, C., Wei, Y.: Bridge the points: Graph-based few-shot segment anything semantically. Advances in Neural Information Processing Systems **37**, 33232–33261 (2024)

41. Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Qiao, Y., Peng, G., Li, H.: Personalize segment anything model with one shot. In: International Conference on Learning Representations. vol. 2024, pp. 18250–18279 (2024)
42. Zhang, Y., Zhou, K., Liu, Z.: What makes good examples for visual in-context learning? Advances in Neural Information Processing Systems **36**, 17773–17794 (2023)
43. Zhu, M., Liu, Y., Luo, Z., Jing, C., Chen, H., Xu, G., Wang, X., Shen, C.: Unleashing the potential of the diffusion model in few-shot semantic segmentation. Advances in Neural Information Processing Systems **37**, 42672–42695 (2024)