INSID3: Training-Free In-Context Segmentation with DINOv3

Claudia Cuttano [1,2] Gabriele Trivigno [1] Christoph Reich [2,3,5,6]
Daniel Cremers [3,5,6] Carlo Masone [1] Stefan Roth [2,4,5]
[1]Politecnico di Torino [2]TU Darmstadt [3]TU Munich [4]hessian.AI [5]ELIZA [6]MCML

https://github.com/ClaudiaCuttano/insid3

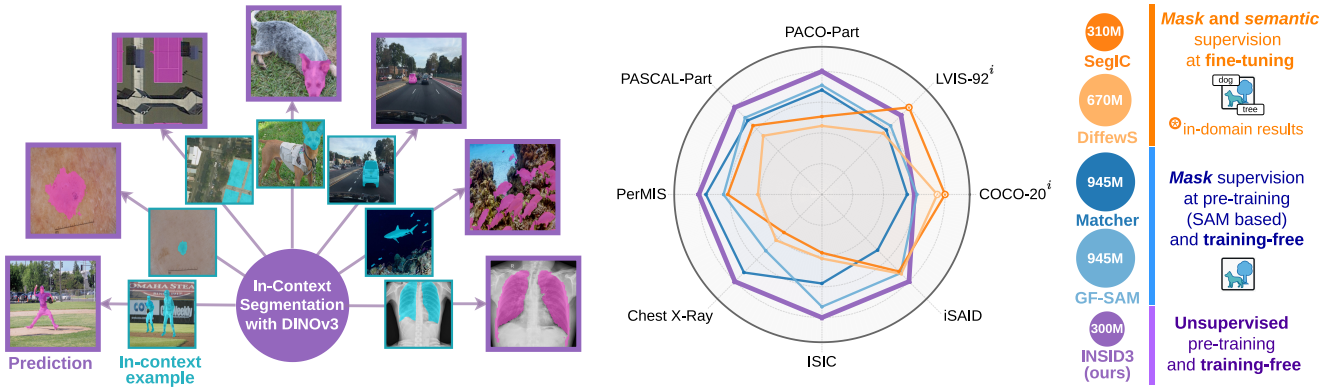


Figure 1. **Results and overview of INSID3, our training-free in-context segmentation approach**. INSID3 performs in-context segmentation directly from DINOv3 [9] features, without any decoder, fine-tuning, or model composition. *(left)* A single annotated example guides the model to segment any concept, from object parts to medical images and aerial views. *(right)* Comparing generalization across datasets and segmentation granularities: fine-tuned methods *(orange)* excel in-domain ( ) but degrade out of distribution, while SAM-based pipelines *(blue)* generalize better but rely on large, multi-stage architectures. INSID3 *(purple)* achieves the strongest generalization with a single backbone, revealing that robust segmentation can emerge directly from the dense self-supervised representations of DINOv3.

Abstract

In-context segmentation (ICS) aims to segment arbitrary concepts, e.g., *objects, parts, or personalized instances, given a few annotated visual examples. Existing work relies on* (i) *fine-tuning vision foundation models (VFMs), which improves in-domain results but limits generalization, or* (ii) *combines multiple frozen VFMs, which preserves generalization but yields architectural complexity and fixed segmentation granularities. We revisit ICS from a minimalist perspective and ask: Can a single self-supervised backbone support both semantic matching and segmentation, without any supervision or auxiliary models? We show that scaled-up dense self-supervised features from DINOv3 exhibit strong spatial structure and semantic correspondence. We introduce INSID3, a training-free approach that* segments concepts *at varying granularities only from frozen DINOv3 features, given an in-context example. INSID3 achieves state-of-the-art results across one-shot semantic, part, and personalized segmentation, outperforming previous work by +7.5 % mIoU, while using 3× fewer parameters and without any mask or category-level supervision.*

1. Introduction

Understanding visual scenes is a fundamental task with widespread applications (*e.g.*, robotics [2]). In-context segmentation (ICS) [5, 6, 15] enables segmentation of arbitrary concepts from one or more annotated examples at inference time, providing an annotation-efficient, flexible framework for understanding visual scenes (*cf*. Fig. 1, *left*). A key requirement for ICS is robust correspondence between the annotated reference and the target image. Prior work has shown that such correspondences emerge in vision foundation models (VFMs) [10, 14]. Recent ICS methods build on this observation in two ways. The *first* line extends pre-trained VFMs with explicit segmentation capabilities, either by training a decoder on top of DINOv2 or by fine-tuning a diffusion model [4, 6, 16]. While effective in-domain, these methods fail to generalize well outside of their training domain and require task-specific supervision (*cf*. Fig. 1, *right*). The *second* line avoids task-specific training and instead combines multiple pre-trained VFMs, typically DINOv2, to establish correspondence, and SAM [3] for mask prediction [5, 13]. While offering generalization across do-

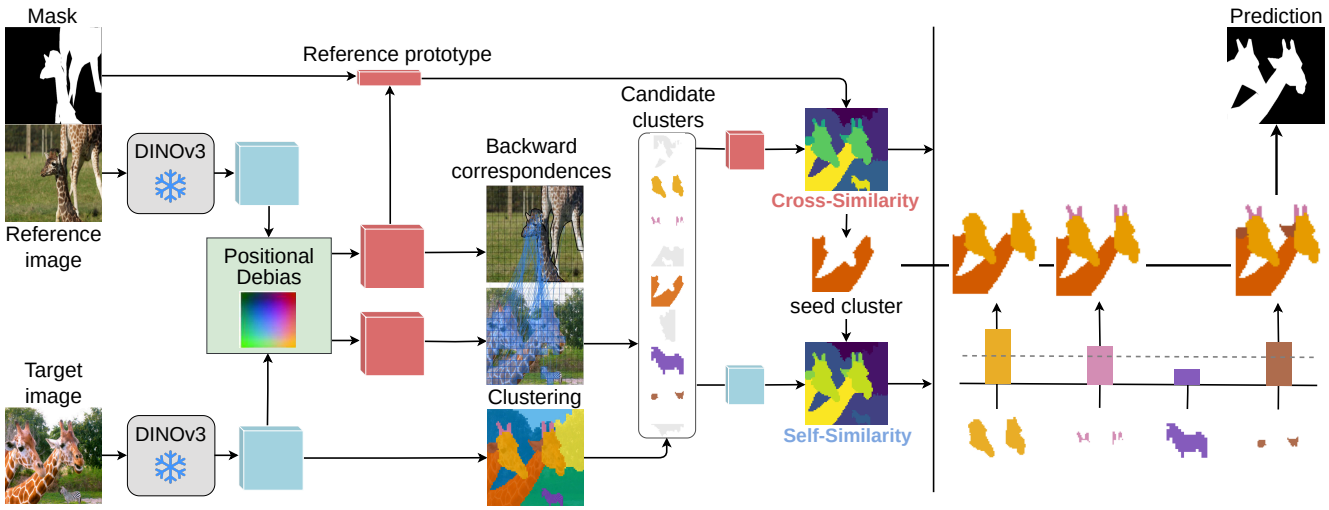


Figure 2. **Overview of INSID3.** We leverage the semantic and spatial structure of DINOv3 to perform in-context segmentation without training or model composition. Dense features from the reference and target images are first debiased to suppress positional bias, improving cross-image matching. The target is then decomposed into coherent regions through agglomerative clustering, providing a structured representation. We retain candidate clusters that match the reference through backward correspondence in the debiased space; a reference prototype derived from the annotated region anchors the *seed cluster* via cross-image similarity. Finally, we combine cross-image similarity, capturing semantic alignment, with self-similarity, measuring the affinity of each cluster to the seed, to form the final mask from the seed.

mains, these approaches introduce significant complexity by model composition and require large-scale mask supervision for pre-training (*cf*. Fig. 1, *right*).

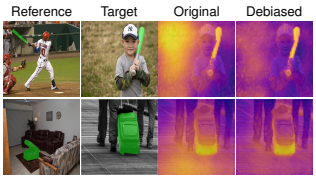
This raises a natural question: *can ICS emerge directly from a single self-supervised visual backbone, without decoders, pre-training supervision, fine-tuning, or model composition?* The recent DINOv3 model [9] demonstrated strong evidence that self-supervised features alone can solve arbitrarily dense prediction tasks. Unlike previous DINO models [1, 8], DINOv3 is designed to produce dense, localized, and semantically rich features, providing an expressive basis to establish correspondences and perform segmentation. We propose INSID3 (**In**-context **S**egmentation w**I**th **DINOv3**), a training-free method that relies solely on frozen DINOv3 features. INSID3 comprises three stages: *(i)* clustering target-image features into fine-grained region candidates, *(ii)* selecting a seed cluster through cross-image similarity with the annotated reference, and *(iii)* recovering the full extent of the prompted concept by aggregating clusters according to within-image feature affinity. In addition, we uncover a subtle but important limitation of DINOv3 features for cross-image reasoning: feature similarities exhibit a systematic positional bias, whereby patches at similar absolute image locations spuriously match even in the absence of semantic agreement. We mitigate this effect with a simple training-free correction, obtained by estimating a positional subspace from a noise image and performing matching in its orthogonal complement. This improves ICS and also benefits semantic correspondence beyond the segmentation setting.

Our contributions can be summarized as follows:

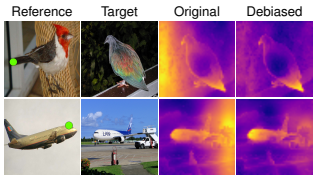
- We are the first to show that *a self-supervised VFM suffices for training-free in-context segmentation*, building on DINOv3's core strengths of robust correspondence and its dense, localized feature structure.
- Despite its simplicity, INSID3 *generalizes better across the board*, from traditional, challenging benchmarks to out-of-domain datasets and part segmentation (Fig. 1, *purple*), outperforming fine-tuned *and* training-free approaches relying on SAM by an average of +7.5 % mIoU.
- We unveil a *positional bias in DINOv3*, which impairs its effectiveness in matching features across images, and present a simple training-free correction that generalizes beyond ICS, achieving gains of up to +6.6 % PCK on the related task of semantic correspondence.

2. Related Work

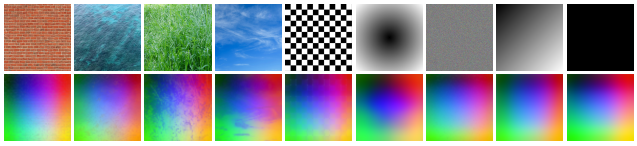
In-context segmentation. ICS transfers the idea of adapting models from contextual examples to segmentation. Early generalist models such as Painter [11] and SegGPT [12] rely on large-scale supervised training. More recent methods follow two main directions. One line injects segmentation capability into pretrained vision backbones, either by training a decoder on top of DINOv2 or by adapting a diffusion model [4, 6, 16]. Another combines frozen correspondence models with Segment Anything for mask generation [5, 13, 15]. In contrast, we address ICS with a single frozen self-supervised backbone, without fine-tuning, pre-training supervision, and model composition.



(a) Cross-image similarity map using an **object region** as reference.



(b) Cross-image similarity map using a **keypoint** as reference.



(c) **Positional subspace.** PCA on uniform and low-complexity images.

Figure 3. **Positional bias in DINOv3 features.** For both region *(a)* and keypoint *(b)* prompts, similarity maps computed with the original DINOv3 features show structured activations aligned with the reference coordinates, independent of semantics. Our debiased features mitigate this behavior. *(c)* PCA of features from images with low semantic complexity (*e.g.*, noise, flat textures) reveals a stable low-dimensional positional subspace underlying this bias.

3. In-context Segmentation with INSID3

Our goal is to segment arbitrary concepts, *i.e.*, objects, parts, or personalized instances, from a single[1] in-context example, using only a frozen DINOv3 encoder and without training or model composition. As illustrated in Fig. 2, INSID3 builds on two complementary properties of DINOv3 features: cross-image semantic matching and strong within-image self-similarity. We exploit the former to localize the reference concept in the target image, and the latter to recover its full spatial extent.

Task definition. Given a reference image $\mathbf{I}^r$ with binary mask $\mathbf{M}^r$ and a target image $\mathbf{I}^t$, we extract dense patch features with a frozen DINOv3 encoder Φ:

$$\mathbf{F}^r = \Phi(\mathbf{I}^r), \qquad \mathbf{F}^t = \Phi(\mathbf{I}^t), \tag{1}$$

where $\mathbf{F}^r, \mathbf{F}^t \in \mathbb{R}^{P \times D}$. Let $\Omega = \{1, \dots, P\}$ be the patch indices and $\mathcal{R} = \{j \in \Omega \mid \mathbf{M}^r_j = 1\}$ the reference patches.

Debiased matching. We observe that raw DINOv3 similarities across images exhibit a positional bias (*cf*. Fig. 3), leading to spurious matches between patches at similar absolute image locations. To mitigate this effect, we estimate a low-dimensional positional subspace from a noise image $\mathbf{I}^{\text{noise}}$ and project features onto its orthogonal complement:

$$\mathbf{F}^{\text{noise}} = \Phi(\mathbf{I}^{\text{noise}}), \qquad \tilde{\mathbf{F}} = \mathbf{F}(\mathbf{1}_D - \mathbf{B}\mathbf{B}^\top), \tag{2}$$

where $\mathbf{B}$ contains the top-s right singular vectors of $\mathbf{F}^{\text{noise}}$. We use debiased features for all cross-image comparisons, and retain the original features for within-image grouping.

[1]INSID3 can be extended to *multiple* in-context examples by simple majority-vote filtering during correspondence search.

Clustering and mask extraction. We first partition the target image into coherent regions by agglomerative clustering on the original target features $\mathbf{F}^t$, obtaining clusters $\{\mathcal{G}_1, \dots, \mathcal{G}_K\}$. To localize the reference concept, we compute backward correspondences in the debiased space,

$$\text{NN}(i) = \arg\max_{j \in \Omega} \langle \tilde{\mathbf{F}}^t_i, \tilde{\mathbf{F}}^r_j \rangle, \tag{3}$$

and retain target patches whose nearest reference neighbor lies inside the support mask, yielding a set of candidate clusters $\mathcal{C}_{\text{cand}}$. This backward matching acts as a filter that suppresses semantically related but irrelevant regions.

We then score each candidate cluster by cross-image similarity to the reference prototype:

$$\tilde{\mathbf{p}}^r = \frac{1}{|\mathcal{R}|} \sum_{j \in \mathcal{R}} \tilde{\mathbf{F}}^r_j, \qquad \tilde{\mathbf{p}}^t_k = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \tilde{\mathbf{F}}^t_i, \tag{4}$$

$$s^{\text{cross}}_k = \langle \tilde{\mathbf{p}}^t_k, \tilde{\mathbf{p}}^r \rangle, \qquad \mathcal{G}^* = \arg\max_{\mathcal{G}_k \in \mathcal{C}_{\text{cand}}} s^{\text{cross}}_k. \tag{5}$$

The resulting seed cluster typically captures the most discriminative region of the concept. To recover its full extent, we aggregate candidate clusters according to their similarity to the seed in the original feature space, which preserves the strong self-similarity structure of DINOv3:

$$S_k = s^{\text{cross}}_k \, s^{\text{intra}}_k, \quad \mathcal{M}_{\text{final}} = \mathcal{G}^* \cup \{\mathcal{G}_k \in \mathcal{C}_{\text{cand}} \mid S_k \geq \alpha\}. \tag{6}$$

This yields the final segmentation mask $\mathcal{M}_{\text{final}}$.

4. Experiments

We evaluate INSID3 on one-shot semantic, part, and personalized segmentation. In all settings, a single annotated reference mask is provided at inference time. We use six semantic segmentation benchmarks (LVIS-92^i, COCO-20^i, ISIC2018, SUIM, iSAID-5^i, and Chest X-Ray), two part segmentation benchmarks (PASCAL-Part and PACO-Part), and the PerMIS benchmark for personalized segmentation. We report the mean Intersection-over-Union (mIoU) and use the DINOv3 Large encoder, while following standard one-shot evaluation protocols from prior work [5, 13, 15].

4.1. Main results

Table 1 compares INSID3 with both fine-tuned and training-free ICS approaches. Fine-tuned methods perform strongly on datasets whose train split matches the test distribution. Training-free approaches generalize better, yet depend on model composition, typically combining DINOv2 and SAM. In contrast, INSID3 uses only a *single*, frozen backbone (*i.e.*, DINOv3) and *no supervision*, except for the reference example. On average, across eight datasets, INSID3 achieves the best accuracy (55.1 % in mIoU), outperforming both fine-tuned and training-free approaches.

Table 1. **Comparison of INSID3 (mIoU in %, ↑) on one-shot semantic, part, and personalized segmentation.** State-of-the-art methods are grouped into task-specific fine-tuning and training-free approaches. Previous training-free methods rely on SAM, pretrained with mask-level supervision, whereas INSID3 uses only frozen self-supervised DINOv3 features. Gray indicates the model was trained on the corresponding train split of the dataset; best results **bold**, 2nd best underlined. † denotes a GF-SAM variant using DINOv3 features.

			Semantic						Part		Personalized	
Method	**Encoder**	**#Param**	**LVIS-92^i**	**COCO-20^i**	**ISIC**	**SUIM**	**iSAID**	**X-Ray**	**PASCAL**	**PACO**	**PerMIS**	**Avg**
Task-specific fine-tuning: *Semantic + mask supervision*												
Painter [11]	ViT	354 M	10.5	33.1	–	–	–	–	30.4	14.1	–	–
SegGPT [12]	ViT	354 M	18.6	56.1	37.5	34.9	30.9	**87.5**	35.8	13.5	18.7	37.1
SINE [4]	DINOv2	373 M	31.2	64.5	25.8	50.7	38.3	39.8	36.2	23.3	42.5	39.1
DiffewS [16]	Stable Diffusion	890 M	31.4	71.3	27.8	48.9	47.5	41.6	34.0	22.8	35.2	40.1
SegIC [6]	DINOv2	310 M	44.6	76.1	25.3	52.5	46.1	34.5	39.9	25.9	51.8	44.1
SegIC [6]	DINOv2	310 M	35.7	75.6	22.5	52.9	40.8	30.8	38.6	25.1	44.9	40.8
Training free: *Mask-supervised pre-training*												
PerSAM [15]	SAM	640 M	11.5	23.0	23.9	28.7	19.2	31.7	32.5	22.5	48.6	26.8
Matcher [5]	DINOv2 + SAM	945 M	33.0	52.7	38.6	44.1	33.3	70.8	42.9	34.7	63.8	46.0
GF-SAM [13]	DINOv2 + SAM	945 M	35.2	**58.7**	48.7	53.1	47.1	51.0	44.5	36.3	54.1	47.6
GF-SAM† [13]	DINOv3 + SAM	945 M	31.8	54.8	50.9	50.5	46.7	56.1	44.9	34.4	52.6	47.0
↳ + *our debias*	DINOv3 + SAM	945 M	34.6	55.9	51.8	52.9	47.6	60.0	46.2	36.1	54.5	48.8
Training free: *Unsupervised pre-training*												
INSID3 (ours)	DINOv3	**304 M**	**41.8**	57.6	**54.4**	**54.9**	**52.1**	78.8	**50.5**	**38.7**	**67.0**	**55.1**

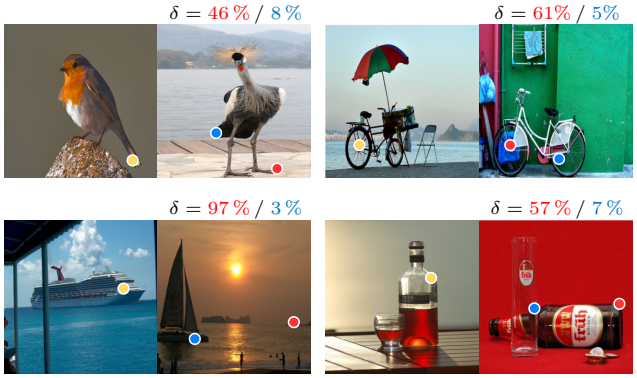
Table 2. **Semantic correspondence on SPair-71k** (PCK@T in %, ↑). Comparison across DINOv3 backbones, w/ and w/o debiasing.

	Small		Base		Large	
T	*original*	*debias*	*original*	*debias*	*original*	*debias*
0.05	26.8	**27.9**	29.2	**32.3**	32.7	**33.6**
0.10	43.8	**45.7**	45.0	**50.0**	50.6	**52.0**
0.15	53.2	**55.6**	54.0	**59.9**	60.3	**62.1**
0.20	59.8	**62.6**	59.8	**66.4**	66.4	**68.6**

In comparison to GF-SAM, the recent training-free state-of-the-art approach, INSID3 outperforms on seven of the eight datasets, including LVIS-92^i (+6.6 %), ISIC (+5.7 %), PASCAL-Part (+6.0 %), PACO-Part (+2.4 %), and PerMIS (+12.9). Accuracy gains over GF-SAM are especially significant in challenging domains such as Chest X-Ray, where INSID3 improves over GF-SAM by +27.8 % pts.

4.2. Semantic correspondence

To isolate the effect of our feature debiasing, we evaluate semantic correspondence on SPair-71k [7]. Following prior work, for each source keypoint, we identify the target patch with the highest cosine similarity and report PCK@T. As shown quantitatively in Tab. 2 and qualitatively in Fig. 4, removing positional biases consistently improves accuracy across all DINOv3 sizes and evaluation thresholds, with gains of up to +6.6 % in PCK, demonstrating effectiveness.



● source keypoint; ● prediction (original DINOv3); ● prediction (debiased). δ is the relative error w.r.t. GT.

Figure 4. **Qualitative examples on SPair-71k** with DINOv3-L.

5. Conclusion

We introduced INSID3, a training-free framework for in-context segmentation built solely on DINOv3. By leveraging the dual nature of DINOv3 features, *i.e.* semantic alignment and spatial coherence, INSID3 performs correspondence estimation and segmentation within a single backbone. While existing methods rely on either fine-tuning or training-free pipelines grounded in mask-supervised pre-training, INSID3 remains fully *unsupervised*, relying solely on the in-context example for guidance. This suggests that reducing supervision may foster more robust and transferable representations, marking a concrete step toward more scalable and general-purpose visual understanding.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.*, 32(11):1231–1237, 2013. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pages 4015–4026, 2023. 1

[4] Yang Liu, Chenchen Jing, Hengtao Li, Muzhi Zhu, Hao Chen, Xinlong Wang, and Chunhua Shen. A simple image segmentation framework via in-context examples. In *NeurIPS*, pages 25095–25119, 2024. 1, 2, 4

[5] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. In *ICLR*, 2024. 1, 2, 3, 4

[6] Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M. Alvarez, Zuxuan Wu, and Yu-Gang Jiang. SegIC: Unleashing the emergent correspondence for in-context segmentation. In *ECCV*, volume 38, pages 203–220, 2024. 1, 2, 4

[7] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv:1908.10543 [cs.CV]*, 2019. 4

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2

[9] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv:2508.10104 [cs.CV]*, 2025. 1, 2

[10] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, pages 1363–1389, 2023. 1

[11] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, pages 6830–6839, 2023. 2, 4

[12] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Towards segmenting everything in context. In *ICCV*, pages 1130–1140, 2023. 2, 4

[13] Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Bridge the points: Graph-based few-shot segment anything semantically. In *NeurIPS*, pages 33232–33261, 2024. 1, 2, 3, 4

[14] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable Diffusion complements DINO for zero-shot semantic correspondence. In *NeurIPS*, pages 45533–45547, 2023. 1

[15] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. In *ICLR*, 2024. 1, 2, 3, 4

[16] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. In *NeurIPS*, pages 42672–42695, 2024. 1, 2, 4