Explore In-Context Segmentation via Latent Diffusion Models

Chaoyang Wang[1], Xiangtai Li[2,3*], Henghui Ding[4], Lu Qi[5], Jiangning Zhang[6], Yunhai Tong[1], Chen Change Loy[3], Shuicheng Yan[2,3]

[1]School of Intelligence Science and Technology, Peking University, China
[2]Skywork AI, Singapore
[3]Nanyang Technological University, Singapore
[4]Institute of Big Data, Fudan University, China
[5]Wuhan University, China
[6]Zhejiang University, China
cywang@stu.pku.edu.cn, xiangtai94@gmail.com

Abstract

In-context segmentation has drawn increasing attention with the advent of vision foundation models. Its goal is to segment objects using given reference images. Most existing approaches adopt metric learning or masked image modeling to build the correlation between visual prompts and input image queries. This work approaches the problem from a fresh perspective – unlocking the capability of the latent diffusion model (LDM) for in-context segmentation and investigating different design choices. Specifically, we examine the problem from three angles: instruction extraction, output alignment, and meta-architectures. We design a two-stage masking strategy to prevent interfering information from leaking into the instructions. In addition, we propose an augmented pseudo-masking target to ensure the model predicts without forgetting the original images. Moreover, we build a new and fair in-context segmentation benchmark that covers both image and video datasets. Experiments validate the effectiveness of our approach, demonstrating comparable or even stronger results than previous specialist or visual foundation models. We hope our work inspires others to rethink the unification of segmentation and generation.

Project page —
https://wang-chaoyang.github.io/project/refldmseg

Introduction

In-context learning (Brown et al. 2020; Balažević et al. 2023; Bar et al. 2022) provides a new perspective for cross-task modeling for vision and natural language processing (NLP). It enables a model to learn and predict according to the prompts. GPT-3 (Brown et al. 2020) first introduced the concept of in-context learning, which refers to inferring solutions for unseen tasks by conditioning on input-output pairs provided as context. Subsequently, several studies (Bar et al. 2022; Wang et al. 2023a) explored in-context learning in the vision domain, where prompts are designed as visual task input-output pairs.

In the segmentation field (Li et al. 2024; Ding et al. 2023b,a; Liu, Ding, and Jiang 2023), in-context learning serves a similar purpose to few-shot segmentation (FSS) (Shaban et al. 2017; Ding, Zhang, and Jiang 2023). Most methods compute the matching distance between query images and support images, which act as visual prompts for in-context learning.

To address the strict constraints on data volume and category in FSS and enable generalization across different tasks, recent works (Bar et al. 2022; Wang et al. 2023a,b) have extended the concept to in-context segmentation, framing it as a mask generalization task (Fig. 1(b)). This approach fundamentally differs from matching or prototype-based discriminative models (Fig. 1(a)), as it directly generates masks through mask decoding. However, these methods typically require large datasets to learn such correspondences.

Latent diffusion models (LDMs) (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) have shown significant potential for generative tasks. Several studies (Zhang, Rao, and Agrawala 2023; Rombach et al. 2022) have demonstrated their strong performance in conditional image content creation. Although LDMs were originally designed for generative tasks, there have been attempts to apply them to perceptual tasks as well. Fig. 2(a) shows the mainstream pipeline (Zhao et al. 2023; Xu et al. 2023; Geng et al. 2023; Baranchuk et al. 2022) for LDM-based segmentation, which typically relies on textual prompts for semantic guidance and additional neural networks to support the LDM. However, textual prompts are not always available in real-world scenarios, and relying on additional networks limits the exploration of LDM's segmentation capabilities. This dependency can also degrade model performance when these auxiliary components are absent. Based on this analysis, we propose that in-context segmentation can be reframed as a conditional image mask generation process, fully leveraging the generative potential of LDMs.

In this paper, we explore, for the first time, the potential of diffusion models for in-context segmentation, as illustrated in Fig. 1(c). Our goal is to answer two key questions: First, can LDMs perform in-context segmentation? Second, what factors are crucial for performance, and how do they influence it? To address these questions, we introduce the **L**atent **D**iffusion-based **I**n-context **S**egmentation framework, or **LDIS**, shown in Fig. 2(b). LDIS leverages visual prompts for guidance, eliminating the need for additional neural networks. We focus our analysis on three

*Project Leader

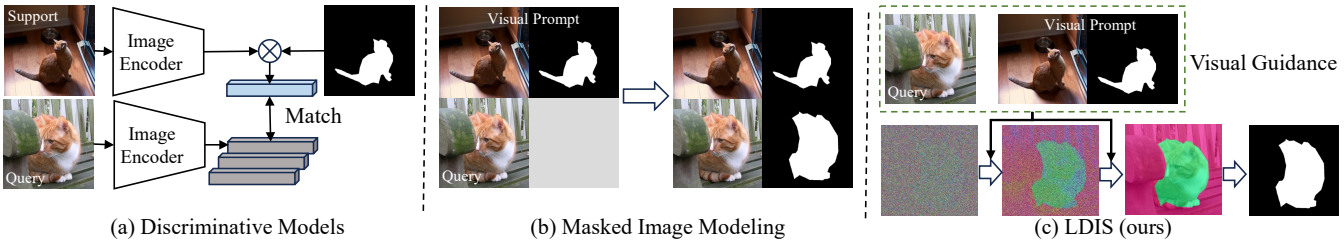


Figure 1: Method comparison. (a) Discriminative models match query images with support prototypes. (b) Masked image modeling methods adopt inpainting training. (c) Our LDM-based model generates segmentation masks guided by visual prompts.

critical factors: instruction extraction, output alignment, and meta-architectures.

First, we introduce a simple yet effective instruction extraction strategy. Experimental results show that these extracted instructions provide strong guidance, and our model remains robust even when instructions are incorrect. Next, to bridge the gap between binary segmentation masks and 3-channel images, we design a novel output alignment target using pseudo-masking modeling. We then propose two meta-architectures: LDIS-1 and LDIS-n. These differ in their input formulation, sampling steps, and optimization targets. Specifically, we design two optimization targets for LDIS-1, one in pixel space and the other in latent space. Experiments highlight the critical role of output alignment. Unlike existing methods (Wang et al. 2023b), our approach focuses more on the architectural impact than on the size of the training data. We aim to use a dataset that is larger than typical few-shot datasets but significantly smaller than the datasets used for foundation models. Finally, we introduce an in-context segmentation benchmark that covers image semantic segmentation, video object segmentation, and video semantic segmentation. We conduct extensive ablation studies and compare our method with previous works to demonstrate its effectiveness. Our contributions are summarized as follows:

- We unlock the in-context segmentation capabilities of latent diffusion models, enabling them to segment specified concepts using visual prompts alone, without relying on textual instructions or additional refinement networks.
- We investigate three key aspects, namely instruction extraction, output alignment, and meta-architectures, and highlighting the importance of accurate instructions, direct optimization targets, and expressive power.
- We propose an in-context learning benchmark, covering both image and video segmentation tasks, and show the effectiveness of our proposed LDIS on these tasks.

Related Work

Diffusion Model Design. Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021; Song et al. 2021) have shown remarkable performance on generation tasks, such as image generation (Rombach et al. 2022; Zhang, Rao, and Agrawala 2023; Ho and Salimans 2021; Dhariwal and Nichol 2021), image editing (Brooks, Holynski, and Efros 2023; Lugmayr et al. 2022; Saharia et al. 2022a; Meng et al. 2022; Hertz et al. 2023; Li et al. 2022), image super resolution (Saharia et al. 2022b; Ho et al. 2022), video generation (Harvey et al. 2022; Yang, Srivastava, and Mandt 2023), and point cloud (Luo and Hu 2021; Zeng et al. 2022). Although the diffusion model is initially designed for generation tasks, several works employ it for segmentation through two pipelines. The first pipeline treats the diffusion model as a feature extractor. These works (Li et al. 2023a; Zhao et al. 2023; Geng et al. 2023; Xu et al. 2023; Xie et al. 2023; Baranchuk et al. 2022; Khosravi et al. 2023; Li et al. 2023b; Wan et al. 2023) typically rely on a decoder head for post-processing. Conversely, the second pipeline (Chen et al. 2023a,b; Gu et al. 2024; Amit et al. 2021; Le et al. 2023) extracts features through a pre-trained backbone, then employs the diffusion model as the decoder head. Recently, several works (Qi et al. 2024) also explore using LDM to generate segmentation masks. However, these studies mainly focus on class-agnostic mask generation with no reference object as the context. These additional neural networks greatly influence our judgment of the true capabilities of diffusion model *itself* in the segmentation task. Moreover, many of these works require textual prompts for guidance. However, the textual prompts, such as categories or captions, are not always available in real-world scenarios.

In-context Learning. GPT-3 (Brown et al. 2020) firstly defines in-context learning, which is interpreted as inferring on unseen tasks conditioning on some input-output pairs given as contexts, also known as prompts. As a new concept in computer vision, in-context learning motivates several attempts (Alayrac et al. 2022; Bar et al. 2022; Lu et al. 2023; Wang et al. 2023a,b; Balažević et al. 2023; Zhang, Zhou, and Liu 2023). The work (Bar et al. 2022) is the first to adopt masked image modeling (MIM) (Bao et al. 2022; He et al. 2022; Xie et al. 2022) as a visual in-context framework. Painter (Wang et al. 2023a) and SegGPT (Wang et al. 2023b) follow the same spirit but scale up with massive training data. Different from MIM, we aim to explore in-context segmentation with the latent diffusion model to explore the potential of condition generation, where we are the first to carry out this study.

Few-shot Segmentation. This task aims to segment query images given support samples. The current works (Wang et al. 2019; Tian et al. 2020; Yang et al. 2020; Liu et al. 2022b,a) typically draw on the idea of metric learning by matching spatial location features with semantic centroids. Furthermore, two-branch conditional networks (Sha-

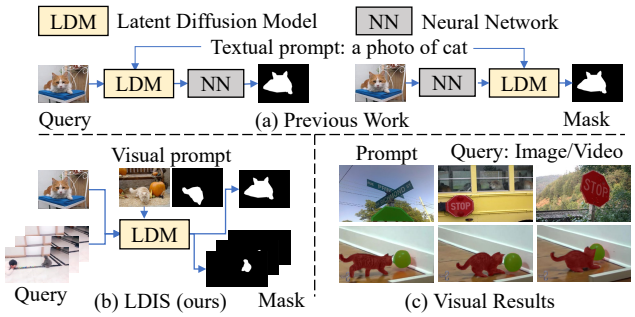


Figure 2: **Latent diffusion model for in-context segmentation.** (a) Previous works mainly rely on textual prompts and additional neural networks for segmentation. (b) Our proposed minimalist framework, LDIS. (c) Segmentation results on images and videos.

ban et al. 2017; Rakelly et al. 2018; Lu et al. 2021; He et al. 2023), 4D dense convolution (Hong et al. 2022; Min, Kang, and Cho 2021) and transformer-based architecture (Zhang et al. 2022; Shi et al. 2022; Xie et al. 2021; Kim et al. 2023) are also widely adopted by researchers. Although few-shot segmentation and in-context segmentation share a similar episode paradigm, the dataset used in few-shot segmentation is typically very small, making the model prone to overfitting, which may influence the evaluation of the generalization ability.

Parameter Efficient Tuning. Research works in this domain (Houlsby et al. 2019; Li and Liang 2021; Hu et al. 2022; Zaken, Ravfogel, and Goldberg 2022; Guo, Rush, and Kim 2021) aim to fine-tune only a tiny portion of parameters to adapt the pre-trained foundation models to various downstream tasks. They maintain the pre-trained knowledge of the foundation models. However, they suffer from inadequate expressive power. In our experiments, we adopt the low-rank adaptation (LoRA) (Hu et al. 2022), a tool widely used in diffusion models, to demonstrate the task gap between generation and segmentation. Although some works (Bahng et al. 2022; Gal et al. 2022; Khani et al. 2023) try to avoid the dilemma by fine-tuning the prompts only, it is time-costly to learn and restore a new embedding for each prompt. In contrast, we employ a prompt encoder to extract in-context instructions from prompts.

Method

This section first introduces the preliminaries of diffusion models and in-context segmentation, then analyzes the design of LDIS from three aspects, namely instruction extraction, output alignment, and meta-architecture, respectively. The notations are illustrated in Tab. 1.

Preliminaries

Diffusion Model. Diffusion models belong to probabilistic generative models that define a chain of forward and backward processes. In the forward process, the model gradually corrupts the data sample z_0 into a noisy latent z_t for $t \in \{1...T\}$: $q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\overline{\alpha}_t}z_0, (1-\overline{\alpha}_t)I)$, where

Notation	Definition	Notation	Definition
$\mathcal{E}$	VAE Encoder	$\mathcal{D}$	VAE Decoder
E_τ	Prompt Encoder	$\mathcal{M}$	Pseudo Mask
M_q	Query Mask	I_q	Query Image
M_s	Prompt Mask	I_s	Prompt Image
M	Ground Truth	τ	Instruction
z	Latent Input	$\tilde{z}$	Latent Prediction
z_q	Latent Query	z_p	Latent Pseudo Mask

Table 1: Illustration of some notations in the Method section.

$\overline{\alpha}_t = \prod_{i=0}^{T} \alpha_s = \prod_{i=0}^{T}(1-\beta_s)$ and β is determined by noise scheduler. During training, the model learns to predict the noise $\epsilon_\theta(z_t, t)$ under the supervision of L_2 loss as $\mathcal{L} = \frac{1}{2}||\epsilon_\theta(z_t, t) - \epsilon(t)||^2$. During inference, the model starts from a random noise $z_T \sim \mathcal{N}(0, 1)$, gradually predicting the noise. Then, it reconstructs the original data z_0 with T steps based on the estimated noise.

In-context Segmentation. Define query image I_q and context set $S = (I_i, M_i)_{i=1}^{K}$, where I_i is a prompt image belonging to a specific visual concept, M_i is the corresponding mask and K is the number of prompts. The model aims to learn a segmentation function $g(I_q, S) \mapsto M_q$ such that based on the context information, it accurately segments target regions in the query image that are similar to the context.

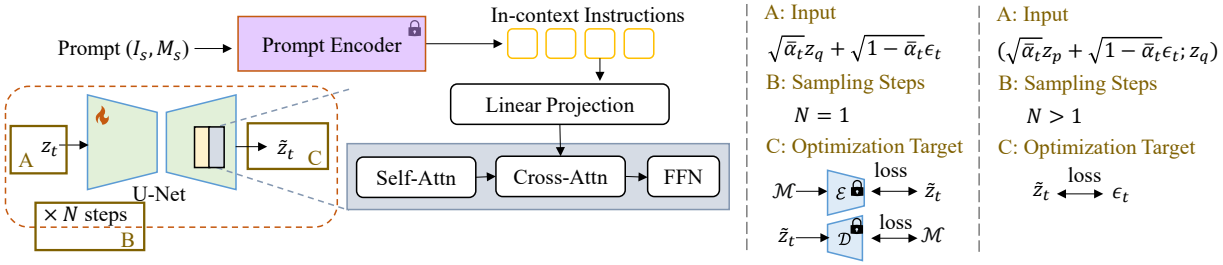
Framework

The LDM is orinally designed for generative tasks. Most existing works (Geng et al. 2023; Zhao et al. 2023) that apply LDMs to segmentation rely on task-specific decoders to process intermediate features or refine imperfect segmentation results. However, incorporating such decoders limits the generative potential of LDMs. To address this, we use Stable Diffusion (SD) as the base model with minimal modifications to fully explore its capabilities. In this subsection, we investigate three key factors that influence the process: instruction extraction, output alignment, and meta-architectures.

Instruction Extraction. Instructions play an essential role in LDM. They act as the compressed prompt representation and guide the generation process. To align with SD, it is intuitive to employ the CLIP vision encoder as the prompt encoder and use a binary mask to filter foreground information. However, this simple approach brings about the leakage of interference. Taking CLIP ViT L/14 as an example, it takes as input a 224×224 image I_s and down-samples it to 16×16. In this process, the information of background or irrelevant targets is distributed among 196 tokens. Moreover, the down-sampled binary mask M_s is ambiguous and cannot precisely represent the boundary of targets. To this end, we propose a two-stage masking strategy consisting of pre-masking and post-masking. In the pre-masking stage, the mask M_s is taken as inputs along with the prompt image I_s, described as Eq. 1.

$$\tau = F\left(E_\tau\left(I_s, M_s\right)\right), \tag{1}$$

where F indicates the linear projection for alignment. The binary mask M_s filters the foreground tokens in this post-



(a) Meta-architecture (b) LDIS-1 (c) LDIS-n

Figure 3: **Our proposed LDIS**. **Left:** Meta-architecture. Our model operates as a minimalist and generates the mask under the guidance of in-context instructions. **Right:** The two variants of our meta-architecture differ in input formulation, sampling steps, and optimization target. Notations are illustrated in Tab. 1.

masking stage. In practice, we employ it as an attention map in cross-attention layers.

Output Alignment. We employ LDM for the segmentation task, so the inconsistency between 1-channel masks and 3-channel images is non-negligible. A pseudo mask must be designed to align the gaps as an intermediate step toward the binary segmentation mask.

An intuitive method follows a mapping rule that transforms the binary masks M to 3-channel pseudo masks $\mathcal{M}_v$:

$$\mathcal{M}_{vi} = \begin{cases} (b, a, (a+b)/2)\,, & M_i \in bg \\ (a, b, (a+b)/2)\,, & M_i \in fg \end{cases}, \tag{2}$$

where bg and fg indicate background and foreground, respectively. M_i is the value in position i. a, b are both scalar, indicating the value of a specific channel in the pseudo masks. We set $a < b$.

In the inference stage, the binary segmentation mask can be recovered with simple arithmetic operations:

$$\tilde{M} = \tilde{\mathcal{M}}_v[1] > \tilde{\mathcal{M}}_v[0]. \tag{3}$$

where $\tilde{M}$ and $\tilde{\mathcal{M}}_v$ indicate the predicted segmentation mask and vanilla pseudo masks, respectively. $[k]$ means the value in the k^{th} channel.

Beyond the vanilla design, we also propose an augmented strategy to fuse the information of images into pseudo masks. Denote the query image as I_q, and the augmented pseudo masks are formulated as follows:

$$\mathcal{M}_a = (1-\gamma)\mathcal{M}_v + \gamma I_q, \tag{4}$$

where γ controls the strength of the information of the image.

Meta-architectures. As shown in Fig. 3, we explore two representative meta-architectures, namely LDIS-1 and LDIS-n. They mainly differ in the input formats, sampling steps, and optimization targets.

LDIS-1 indicates **one-step sampling** and the optimization target is the segmentation mask itself. As shown in Fig. 3(a), a noise variant ϵ_t is added to the latent variant z_q. The model takes as input the noisy latent $z_t = \sqrt{\bar{\alpha}_t} z_q + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where ϵ_t is the output of the noise scheduler, t controls the noise strength.

We propose two optimization strategies that align the model outputs with the ground truth in pixel space or latent space, respectively. We employ the L2 loss, typically used in LDM, rather than any explicit segmentation loss.

$$\mathcal{L}_{fp} = \mathbb{E}_{z_t,\tau}\left[||\mathcal{M} - \tilde{\mathcal{M}}_t||_2^2\right], \tag{5}$$

$$\mathcal{L}_{fl} = \mathbb{E}_{z_t,\tau}\left[||z_p - \tilde{z}_t||_2^2\right], \tag{6}$$

where $\tilde{\mathcal{M}}_t = \mathcal{D}(\tilde{z}_t)$ and $z_p = \mathcal{E}(\mathcal{M})$.

In the inference stage, LDIS-1 conducts only one-time step and outputs the segmentation (pseudo) masks. The video is treated as a sequence of images. The first frame and its annotation are used as prompts, and subsequent frames are inferred conditioned on it. For videos containing multiple categories, we first calculate the probability of each category as a foreground in turn and then select the category with the highest probability:

$$\tilde{p}_c = \frac{\exp(\mathcal{M}[1])}{\exp(\mathcal{M}[0])}, \quad p_c = \frac{\tilde{p}_c}{1 + \sum_{i=1}^{C} \tilde{p}_i}, \tag{7}$$

where $\tilde{p}_c$ indicates the normalized foreground probability map for category c. C is the number of categories. $\mathcal{M}[i]$ means the value in channel i of pseudo masks. It is evident that $\tilde{p}_0$, as the background's probability, equals 1.

LDIS-n indicates **multi-step sampling** and employs an indirect optimization strategy. Unlike LDIS-1, it starts from Gaussian noise and gradually denoises to get the final segmentation mask.

A plain SD architecture is not suitable for LDIS-n. We make minimal but necessary modifications to the architecture by extending the input dimension from 4 to 8. Specifically, denote the latent expression of query image and pseudo mask as $z_q \in \mathbb{R}^{4\times H\times W}$ and $z_p \in \mathbb{R}^{4\times H\times W}$, we get $z_t \in \mathbb{R}^{8\times H\times W}$ by concatenating the noisy pseudo mask latent with z_q in the channel dimension. The noisy pseudo mask latent is obtained by adding noise ϵ_t to z_p:

$$z_t = \mathrm{CONCAT}((\sqrt{\bar{\alpha}_t} z_p + \sqrt{1-\bar{\alpha}_t}\epsilon_t); z_q), \tag{8}$$

where the noise scheduler determines t.

Similar to LDIS-1, the ICS model f inputs the latent variable z_t and the instructions τ, but outputs the estimation of the noise rather than the pseudo mask. We also adopt L2 loss as follows:

$$\mathcal{L}_n = \mathbb{E}_{z_t,t,\tau}\left[||\epsilon_t - \tilde{z}_t||_2^2\right]. \tag{9}$$

Dataset	Task	#Category	#Videos		#Images	
			Train	Val	Train	Val
PASCAL	ISS	20	-	-	10582	2000
COCO	ISS	80	-	-	82081	5000
DAVIS-16	VOS	-	30	16	(2064)	-
VSPW	VSS	58	1000	100	(16473)	-

Table 2: **Details of the combined datasets.** We choose two image semantic segmentation (ISS) datasets, one video object segmentation (VOS) dataset, and one video semantic segmentation (VSS) dataset. (N) means the equivalent number of images.

To reduce the randomness caused by initial noise, enhance the influence of in-context instructions, and ensure consistency between outputs and queries, we employ classifier-free guidance (CFG) (Ho and Salimans 2021). The query latent z_q and condition τ are randomly set to null embedding with probability $p = 0.05$ in the training stage.

We also adopt CFG in the inference stage. Specifically, LDIS-n outputs the $\tilde{z}_t(z_q, \tau)$ on the basis of three conditional outputs $\tilde{z}_t(z_q, \tau)$, $\tilde{z}_t(\varnothing, \varnothing)$, $\tilde{z}_t(z_q, \varnothing)$ (Eq. 10).

$$\begin{aligned}\tilde{z}_t(z_q, \tau) = &\tilde{z}_t(\varnothing, \varnothing) \\ &+ \gamma_q \cdot (\tilde{z}_t(z_q, \varnothing) - \tilde{z}_t(\varnothing, \varnothing)) \\ &+ \gamma_\tau \cdot (\tilde{z}_t(z_q, \tau) - \tilde{z}_t(z_q, \varnothing)),\end{aligned} \tag{10}$$

where γ_q and γ_τ control the guidance of query and in-context instruction, respectively.

Experiments

We begin by outlining the experimental settings, followed by a comprehensive comparison of our methods with existing specialist and generalist models. Finally, we perform ablation studies to evaluate the effectiveness of our design choices.

Experimental Setting

Benchmark Details. The in-context segmentation model aims to solve multiple tasks with *one* model, regardless of data type and domain. To this end, we adopt several popular datasets as part of our benchmark (Tab. 2), including PASCAL (Everingham et al. 2010), COCO (Lin et al. 2014), DAVIS-16 (Perazzi et al. 2016) and VSPW (Miao et al. 2021). The training data for VSPW is sampled every four frames. All 'stuff' categories are annotated as background.

Implementation Details. We utilize the SD 1.5 model as the initialization and set the resolution as 256×256. Our model is jointly trained on the combined dataset for 160K iterations with a batch size of 64. We employ an AdamW optimizer. Alpha CLIP (Sun et al. 2024) ViT-L is adopted as the prompt encoder. We set the CFG coefficient for query and instructions as 1.5 and 7, respectively. Our training follows the spirit of episodic learning. For the image dataset, images with the same semantic labels are considered a pair of queries and prompts. The video dataset follows the image dataset but requires that the query and prompt come from the same video.

Evaluation Metrics. We adopt the class mean intersection over union as our evaluation metric in empirical study, which is formulated as $\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C} \mathrm{IoU}_i$. C is the number of classes except background. We also report the foreground-background IoU for other image-level tasks in our benchmark following (Tian et al. 2020). For video-level tasks, we respectively adopt their evaluation metrics in DAVIS-16 (Perazzi et al. 2016) and VSPW (Miao et al. 2021).

Benchmark Results

Baselines. We report specialist models and generalist models as baselines. We take PFENet, SVF, VTM, DCAMA, and VPD as specialist baselines. Specifically, the first four methods are proposed for few-shot tasks, which aligns closely with our setting. VPD exploits the features extracted by SD UNet. All the specialist models are jointly trained on the combined datasets. As for generalist models, we take Painter (Wang et al. 2023a), SegGPT, PerSAM (Zhang et al. 2024) and Prompt Diffusion (Wang et al. 2023c) into comparisons. They represent several classic modeling approaches, respectively. In detail, Painter and SegGPT are based on masked image modeling. Prompt Diffusion adopts a ControlNet-like architecture. PerSAM leverages the robust segmentation capabilities from SAM (Kirillov et al. 2023).

Quantitative Results. We make comprehensive comparisons in Tab. 3. For image tasks, LDIS-1 achieves the best results compared with these methods. LDIS-n also achieves decent performance of 76.7 mIoU on PASCAL and 52.6 mIoU on COCO. Our approaches perform comparablely to the generalist models for video tasks. Two factors explain this phenomenon. One is the data volume. The training data of the other generalist models is far beyond ours. The other is the spatial prior — the content of the first and subsequent frames is highly similar in video segmentation tasks. The other generalist models use spatial information by taking the pixel-level prompts as inputs. However, our model only relies on the visual embedding extracted by the prompt encoder without spatial knowledge. Achieving state-of-the-art on all metrics is not the primary goal of this work, and we leave it for future work.

Qualitative Results. Fig. 4 shows the visual comparison between our model and previous works on the COCO dataset. It is evident from these results that our model successfully segments the target region. In contrast, other methods fail to establish the semantic connection between prompt and query, leading to missing and false-positive predictions.

Ablation Study and More Analysis

In this subsection, we conduct comprehensive ablation experiments to study the subtle effects of different designs. COCO dataset with fixed prompt-query pairs is used for evaluation. The experiments are conducted from three aspects, namely instruction extraction, output alignment, and meta-architectures, respectively.

Instruction Extraction. It is crucial to provide accurate visual instructions to guide the model in segmenting the specified concepts. However, the quality of instructions is affected by three factors. The initial factor to be considered

Modeling	Method	Backbone / Initialization	PASCAL mIoU	PASCAL FB-IoU	COCO mIoU	COCO FB-IoU	DAVIS-16 $\mathcal{J}\&\mathcal{F}$	VSPW mIoU	VSPW fwIoU
Discriminative Modeling	PFENet (Tian et al. 2020)	R50	76.4	88.1	49.4	76.9	-	-	-
	SVF (Sun et al. 2022)	R50	77.0	88.5	49.2	77.0	-	-	-
	VTM (Kim et al. 2023)	R50	73.9	85.4	52.1	76.2	-	-	-
	DCAMA (Shi et al. 2022)	R50	71.4	85.2	43.5	73.2	-	-	-
	VPD (Zhao et al. 2023)	SD 1.5	75.3	86.9	48.6	75.6	-	-	-
MIM	Painter (Wang et al. 2023a)	ViT-L	53.7	69.7	30.8	58.3	76.7	12.2	76.7
	SegGPT (Wang et al. 2023b)	Painter	75.6	86.7	50.7	76.9	**80.5**	**61.7**	**93.3**
	SegGPT† (Wang et al. 2023b)	Painter	65.2	81.4	39.5	71.2	73.4	50.4	90.0
SAM	PerSAM (Zhang et al. 2024)	SAM	47.6	69.4	25.5	57.5	68.7	42.6	79.1
LDM	Prompt Diffusion (Wang et al. 2023c)	SD 1.5	9.0	40.1	5.9	40.7	-	-	-
	LDIS-n (ours)	SD 1.5	76.7	87.2	52.6	76.3	64.7	30.5	79.7
	LDIS-1 (ours)	SD 1.5	**85.3**	**93.1**	**62.6**	**84.5**	67.8	41.6	87.8

Table 3: **Benchmark results.** We compare our method with several representative specialist models and generalist models. The **bold** entries represent the best performance. † indicates the number of reproductions with a resolution of 256.

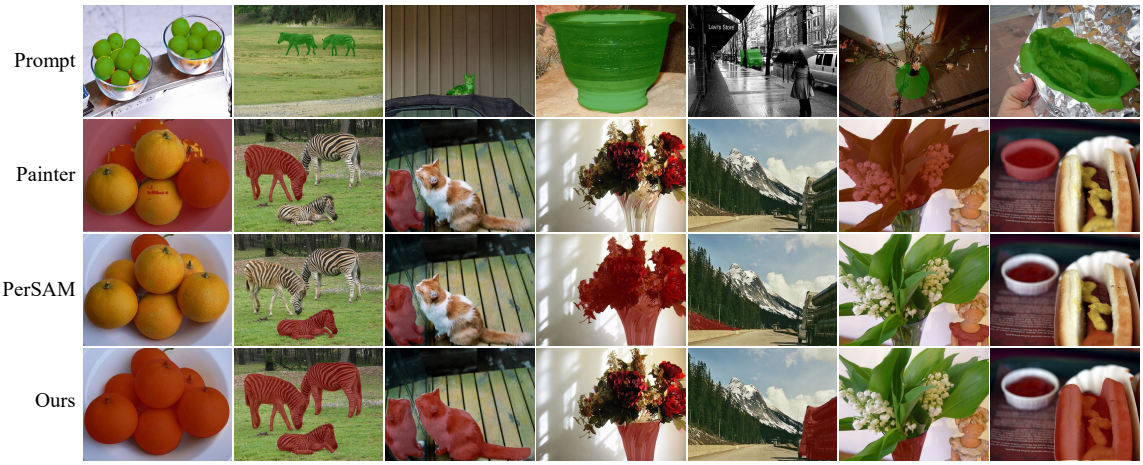


Figure 4: Visualization of segmentation results. We compare our LDIS-1 with Painter (Wang et al. 2023a) and PerSAM (Zhang et al. 2024) on the COCO dataset.

Figure 5: **Visualizations at different time steps.** LDIS-n captures low-frequency components at the beginning and then generates high-frequency information as the denoising process approaches completion. The number of denoising steps is 20.

is the number of visual prompts. As shown in Tab. 4, the performance of LDIS-n first grows significantly as the number of visual prompts increases from 1 to 5 and then saturates around 10. Multiple visual prompts provide different perspectives on the same concept, leading to more accurate estimations. However, too many prompts may cause information interference. The second factor is how instructions are extracted. Tab. 5 provides the ablation results of a two-stage masking strategy. Instructions without pre-masking or post-masking bring about unnecessary information leakage, and these models only achieve 43.8 and 49.7 mIoU. Models with a two-stage masking strategy get the best performance of 52.6 mIoU.

Number of Prompt	1	3	5	10	20
mIoU	49.7	55.3	56.2	56.7	55.7

Table 4: LDIS-n with different number of prompts.

In addition, we study the combination of instructions and their influences on performance. Fig. 6 illustrates some examples. Our model accurately segments the target region in

Pre-Mask	Post-Mask	mIoU
✓	-	43.8
-	✓	49.7
✓	✓	52.6

Table 5: Instruction extraction strategy on LDIS-n.

Method	mIoU
$\mathcal{M}_v$	39.4
$+\epsilon$	48.7
+I ($\mathcal{M}_a$)	49.7

Table 6: Pseudo masking strategy on LDIS-n.

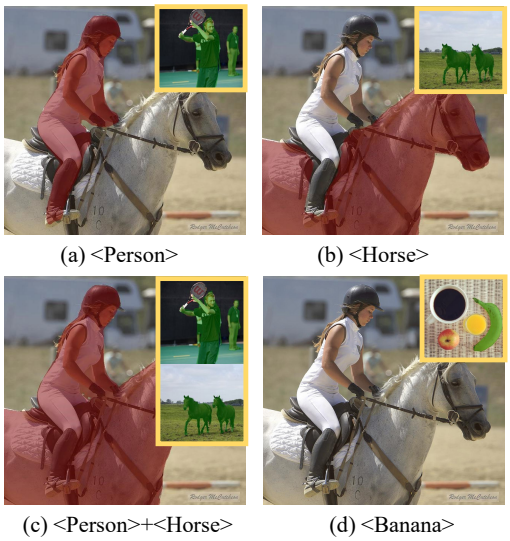
Meta-architecture	Full Train	Rank	PO	mIoU
LDIS-1	-	1	-	53.0
	-	4	-	53.2
	-	4	✓	55.6
	✓	-	-	61.3
LDIS-n	-	4	-	23.2
	✓	-	-	49.7

Table 7: Meta-architecture strategy. PO means pixel space optimizaiton.

the first and second cases based on the single instruction provided. The third case shows the model can accept several instructions without performance degradation. In the fourth case, the provided instruction becomes ineffective when it conflicts with the query.

Output Alignment. We study the effects of output alignment from two aspects. One of them is the format of the optimization target. We argue that a more challenging learning target benefits the training process and helps the model avoid learning some shortcuts. In Tab. 6, LDIS-n can only achieve 39.4 mIoU with vanilla pseudo masks. We then apply a certain perturbation intensity on masks, forcing the model to differentiate the foreground and background from the statistics rather than remembering the constant value. The perturbation follows a uniform distribution and is added to $\mathcal{M}_v$ as Eq. 4. Surprisingly, the perturbation brings about a significant improvement of 9.3 mIoU. Afterward, we replace the perturbation with the query image and get $\mathcal{M}_a$, which introduces the semantics. In this way, LDIS-n with $\mathcal{M}_a$ achieves a mIoU of 49.7. Another factor is the optimization method. In the second and third lines of Tab. 7, LDIS-1 improves 2.4 mIoU when optimized in the pixel space compared to the latent space. As pixel space directly aligns with the segmentation target, optimization methods like 'PO' help reduce the transformation error introduced by VAE and achieve better performance.

Meta-architecture. We investigate both meta-architectures by applying different training strategies. Generally, the model with all parameters trainable performs best, with LDIS-1 and LDIS-n getting mIoU of 61.3 and 49.7, respectively. Furthermore, we study the effects of parameter-efficient tuning on our model with LoRA. Performance degradation is observed on both architectures, 8.1 on LDIS-1 and 26.5 on LDIS-n, respectively. As we further reduce the rank from 4 to 1, LDIS-1's performance slightly degrades, reaching 53.0 mIoU. This phenomenon indicates that since SD was initially designed for generative tasks, the restriction of expressive power hinders its transfer to segmentation tasks. LDIS-n is more sensitive to this characteristic.



(a) <Person> (b) <Horse> (c) <Person>+<Horse> (d) <Banana>

Figure 6: **Combination of instructions.** The output of our model varies based on the instructions located in the top right corner of the query images. (a) and (b) Single instruction. (c) Multiple instructions. (d) Incorrect instruction.

Method	Fold-0	Fold-1	Fold-2	Fold-3	Mean
RePRI (Boudiaf et al. 2021)	32.0	38.7	32.7	33.1	34.1
BAM (Lang et al. 2022)	43.4	50.6	47.5	43.4	46.2
FPTrans (Zhang et al. 2022)	44.4	48.9	50.6	44.0	47.0
PerSAM (Zhang et al. 2024)	21.8	24.1	20.8	22.6	22.3
LDIS-1	59.0	64.8	59.6	57.8	60.3

Table 8: 1-shot segmentation on COCO-20^i using mIoU.

Results on One-shot Segmentation. We also test our model under the few-shot segmentation setting. As shown in Tab. 8, our model achieves decent performance on all folds of COCO-20^i dataset. Specifically, it outperforms some recently proposed generalists, such as PerSAM (Zhang et al. 2024), by a remarkable margin.

Conclusion

For the first time, we explore and unlock the in-context segmentation capability of latent diffusion models. We empirically study the influential factors of an LDM-based segmentation framework, including instruction extraction, output alignment, and meta-architectures, and highlight the importance of precise instructions, direct optimization targets, and expressive power. We also propose an in-context segmentation benchmark and achieve comparable or even better results than specialists or vision foundation models.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2023YFC3807600). This project was supported by NSFC under Grant No. 62472104.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Balažević, I.; Steiner, D.; Parthasarathy, N.; Arandjelović, R.; and Hénaff, O. J. 2023. Towards In-context Scene Understanding. In *NeurIPS*.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. Beit: Bert pre-training of image transformers. In *ICLR*.

Bar, A.; Gandelsman, Y.; Darrell, T.; Globerson, A.; and Efros, A. 2022. Visual prompting via image inpainting. In *NeurIPS*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2022. Label-efficient semantic segmentation with diffusion models. In *ICLR*.

Boudiaf, M.; Kervadec, H.; Masud, Z. I.; Piantanida, P.; Ben Ayed, I.; and Dolz, J. 2021. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *CVPR*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023a. Diffusiondet: Diffusion model for object detection. In *ICCV*.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2023b. A generalist framework for panoptic segmentation of images and videos. In *ICCV*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023a. MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions. In *ICCV*, 2694–2703.

Ding, H.; Liu, C.; He, S.; Jiang, X.; Torr, P. H.; and Bai, S. 2023b. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 20224–20234.

Ding, H.; Zhang, H.; and Jiang, X. 2023. Self-regularized prototypical network for few-shot semantic segmentation. *Pattern Recognition*, 133: 109018.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Geng, Z.; Yang, B.; Hang, T.; Li, C.; Gu, S.; Zhang, T.; Bao, J.; Zhang, Z.; Hu, H.; Chen, D.; et al. 2023. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*.

Gu, Z.; Chen, H.; Xu, Z.; Lan, J.; Meng, C.; and Wang, W. 2024. DiffusionInst: Diffusion Model for Instance Segmentation. In *ICASSP*.

Guo, D.; Rush, A. M.; and Kim, Y. 2021. Parameter-efficient transfer learning with diff pruning. In *ACL*.

Harvey, W.; Naderiparizi, S.; Masrani, V.; Weilbach, C.; and Wood, F. 2022. Flexible diffusion modeling of long videos. In *NeurIPS*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *CVPR*.

He, S.; Jiang, X.; Jiang, W.; and Ding, H. 2023. Prototype adaption and projection for few-and zero-shot 3d point cloud semantic segmentation. *IEEE Transactions on Image Processing*, 32: 3199–3211.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-prompt image editing with cross attention control. In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*.

Ho, J.; and Salimans, T. 2021. Classifier-free diffusion guidance. In *NeurIPS Workshops*.

Hong, S.; Cho, S.; Nam, J.; Lin, S.; and Kim, S. 2022. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *ECCV*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Khani, A.; Taghanaki, S. A.; Sanghi, A.; Amiri, A. M.; and Hamarneh, G. 2023. SLiMe: Segment Like Me. *arXiv preprint arXiv:2309.03179*.

Khosravi, B.; Rouzrokh, P.; Mickley, J. P.; Faghani, S.; Mulford, K.; Yang, L.; Larson, A. N.; Howe, B. M.; Erickson, B. J.; Taunton, M. J.; et al. 2023. Few-shot biomedical image segmentation using diffusion models: Beyond image generation. *Computer Methods and Programs in Biomedicine*.

Kim, D.; Kim, J.; Cho, S.; Luo, C.; and Hong, S. 2023. Universal few-shot learning of dense prediction tasks with visual token matching. In *ICLR*.

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

Lang, C.; Cheng, G.; Tu, B.; and Han, J. 2022. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*.

Le, M.-Q.; Nguyen, T. V.; Le, T.-N.; Do, T.-T.; Do, M. N.; and Tran, M.-T. 2023. MaskDiff: Modeling Mask Distribution with Diffusion Probabilistic Model for Few-Shot Instance Segmentation. *arXiv preprint arXiv:2303.05105*.

Li, X.; Ding, H.; Zhang, W.; Yuan, H.; Cheng, G.; Jiangmiao, P.; Chen, K.; Liu, Z.; and Loy, C. C. 2024. Transformer-Based Visual Segmentation: A Survey. *IEEE TPAMI*.

Li, X.; Lu, J.; Han, K.; and Prisacariu, V. 2023a. SD4Match: Learning to Prompt Stable Diffusion Model for Semantic Matching. *arXiv preprint arXiv:2310.17569*.

Li, X.; Zhang, W.; Pang, J.; Chen, K.; Cheng, G.; Tong, Y.; and Loy, C. C. 2022. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*.

Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023b. Open-vocabulary Object Segmentation with Diffusion Models. In *ICCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, C.; Ding, H.; and Jiang, X. 2023. GRES: Generalized Referring Expression Segmentation. In *CVPR*, 23592–23601.

Liu, J.; Bao, Y.; Xie, G.-S.; Xiong, H.; Sonke, J.-J.; and Gavves, E. 2022a. Dynamic prototype convolution network for few-shot semantic segmentation. In *CVPR*.

Liu, Y.; Liu, N.; Cao, Q.; Yao, X.; Han, J.; and Shao, L. 2022b. Learning non-target knowledge for few-shot semantic segmentation. In *CVPR*.

Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2023. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*.

Lu, Z.; He, S.; Zhu, X.; Zhang, L.; Song, Y.-Z.; and Xiang, T. 2021. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *ICCV*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*.

Miao, J.; Wei, Y.; Wu, Y.; Liang, C.; Li, G.; and Yang, Y. 2021. VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild. In *CVPR*.

Min, J.; Kang, D.; and Cho, M. 2021. Hypercorrelation squeeze for few-shot segmentation. In *ICCV*.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*.

Qi, L.; Yang, L.; Guo, W.; Xu, Y.; Du, B.; Jampani, V.; and Yang, M.-H. 2024. Unigs: Unified representation for image generation and segmentation. In *CVPR*.

Rakelly, K.; Shelhamer, E.; Darrell, T.; Efros, A. A.; and Levine, S. 2018. Conditional Networks for Few-Shot Semantic Segmentation. In *ICLR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *SIGGRAPH*.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *TPAMI*.

Shaban, A.; Bansal, S.; Liu, Z.; Essa, I.; and Boots, B. 2017. One-Shot Learning for Semantic Segmentation. In *BMVC*.

Shi, X.; Wei, D.; Zhang, Y.; Lu, D.; Ning, M.; Chen, J.; Ma, K.; and Zheng, Y. 2022. Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In *ECCV*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *ICLR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Sun, Y.; Chen, Q.; He, X.; Wang, J.; Feng, H.; Han, J.; Ding, E.; Cheng, J.; Li, Z.; and Wang, J. 2022. Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. In *NeurIPS*.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. In *CVPR*.

Tian, Z.; Zhao, H.; Shu, M.; Yang, Z.; Li, R.; and Jia, J. 2020. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*.

Wan, Q.; Huang, Z.; Kang, B.; Feng, J.; and Zhang, L. 2023. Harnessing Diffusion Models for Visual Perception with Meta Prompts. *arXiv preprint arXiv:2312.14733*.

Wang, K.; Liew, J. H.; Zou, Y.; Zhou, D.; and Feng, J. 2019. Panet: Few-shot image semantic segmentation with prototype alignment. In *ICCV*.

Wang, X.; Wang, W.; Cao, Y.; Shen, C.; and Huang, T. 2023a. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. In *CVPR*.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023b. SegGPT: Segmenting Everything In Context. In *ICCV*.

Wang, Z.; Jiang, Y.; Lu, Y.; Shen, Y.; He, P.; Chen, W.; Wang, Z.; and Zhou, M. 2023c. In-context learning unlocked for diffusion models. In *NeurIPS*.

Xie, G.-S.; Xiong, H.; Liu, J.; Yao, Y.; and Shao, L. 2021. Few-shot semantic segmentation with cyclic memory network. In *ICCV*.

Xie, J.; Li, W.; Li, X.; Liu, Z.; Ong, Y. S.; and Loy, C. C. 2023. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *arXiv preprint arXiv:2309.13042*.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *CVPR*.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*.

Yang, B.; Liu, C.; Li, B.; Jiao, J.; and Ye, Q. 2020. Prototype mixture models for few-shot semantic segmentation. In *ECCV*.

Yang, R.; Srivastava, P.; and Mandt, S. 2023. Diffusion probabilistic modeling for video generation. *Entropy*.

Zaken, E. B.; Ravfogel, S.; and Goldberg, Y. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *ACL*.

Zeng, X.; Vahdat, A.; Williams, F.; Gojcic, Z.; Litany, O.; Fidler, S.; and Kreis, K. 2022. LION: Latent point diffusion models for 3D shape generation. In *NeurIPS*.

Zhang, J.-W.; Sun, Y.; Yang, Y.; and Chen, W. 2022. Feature-proxy transformer for few-shot segmentation. In *NeurIPS*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Dong, H.; Gao, P.; and Li, H. 2024. Personalize segment anything model with one shot. In *ICLR*.

Zhang, Y.; Zhou, K.; and Liu, Z. 2023. What Makes Good Examples for Visual In-Context Learning? In *NeurIPS*.

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing text-to-image diffusion models for visual perception. In *ICCV*.