AlignDiff: Aligning Diffusion Models for General Few-Shot Segmentation

Ri-Zhao Qiu[1,2], Yu-Xiong Wang[1†], and Kris Hauser[1†]

[1] University of Illinois Urbana-Champaign
[2] University of California San Diego
{rizhaoq2, yxw, kkhauser}@illinois.edu
[†]equal advising
http://github.com/RogerQi/AlignDiff

Abstract. Text-to-image diffusion models have shown remarkable success in synthesizing photo-realistic images. Apart from creative applications, can we use such models to synthesize samples that aid the few-shot training of discriminative models? In this work, we propose AlignDiff, a general framework for synthesizing training images and masks for few-shot segmentation. We identify two crucial misalignments that arise when utilizing pre-trained diffusion models in segmentation tasks, which need to be addressed to create realistic training samples and align the synthetic data distribution with the real training distribution: 1) instance-level misalignment, where generated samples of rare categories are often misaligned with target tasks) and 2) annotation-level misalignment, where diffusion models are limited to generating images without pixel-level annotations. AlignDiff overcomes both challenges by leveraging a few real samples to guide the generation, thus improving novel IoU over baseline methods in few-shot segmentation and generalized few-shot segmentation on Pascal-5^i and COCO-20^i by up to 80%. Notably, AlignDiff is capable of augmenting the learning of out-of-distribution uncommon categories on FSS-1000, while naïve diffusion model generates samples that diminish segmentation performance.

Keywords: Semantic Segmentation · Text-to-Image Diffusion · Data Synthesis

1 Introduction

Few-shot semantic segmentation has recently attracted increasing attention [17, 20, 28], given that it copes with the scarcity of (pixel-level) densely annotated data in practical scenarios. Existing efforts have primarily focused on either designing specialized architectures in low-data regimes [5, 29] or employing data augmentation to produce variations of the provided data [20]. However, these methods struggle to improve performance, as they ultimately rely on the small number of support samples that often do not faithfully represent the real data distribution. A promising approach is to leverage general-purpose text-to-image

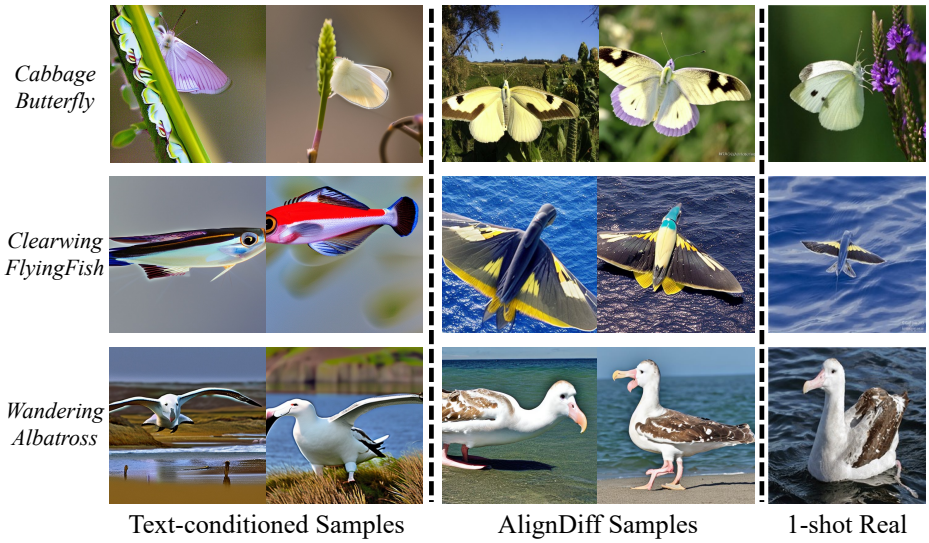

Fig. 1: While it is tempting to use images synthesized from text-to-image diffusion models such as Stable Diffusion [23] to train segmentation models, text-conditioned generated images often fail to represent the desired distributions of uncommon classes. Naively using these samples leads to training distributions that are **misaligned** to the target data distribution, where samples are often degenerate for rare categories and there is no pixel-level annotation. AlignDiff gracefully addresses both issues by conditioning the generative process using a few real training samples, leading to an aligned synthetic distribution for training.

diffusion models to synthesize additional examples that are *not biased* to the support sample distribution, thereby improving few-shot segmentation models.

Trained on a large corpus of image-text pairs, large-scale text-to-image diffusion image generation models [23, 25] have shown remarkable success in creative applications. However, the naïve use of diffusion models to generate examples for training segmentation models leads to **misalignments** between the generated sample distribution and the target data distribution. There are two dimensions of misalignment. 1) **Instance-level.** Simple text conditioning may fail to synthesize images of rare categories (illustrated in Fig. 1). We define this issue as *Out-Of-Distribution* (OOD) generation, which means generated samples are misaligned with real data. We will define OOD generation quantitatively in the experiment section. 2) **Annotation-level.** Segmentation models need accurate *pixel-level mask annotation*, whereas existing diffusion models [23] are limited to only generating images.

To address these challenges, we introduce the **Align**ment **Diff**usion (AlignDiff) framework which is based on the key insight that the few-shot support data *can be leveraged to align the generation process* with the real distribution. Specifically, to address the instance-level misalignment, we propose *normalized*

masked textual inversion. Our method learns an instance-specific word embedding from as few as a single image-mask pair, which ensures consistency with the given real examples. Empirically, our method works on OOD generation scenarios in which existing methods [6, 23] fail. To generate accurate pixel-level annotations, we design an efficient mask generation approach by taking inspiration from semi-supervised learning, where we perform few-shot conditioning of the mask generation process on few novel samples. We formulate the mask generation process as a process to refine noisy masks from a few examples of high-quality masks. Empirically, our approach is much more efficient than previous methods to extract masks from diffusion models [31] with on-par accuracy.

In sum, our contributions are as follows. **1)** To improve instance-level alignment and handle OOD generation, we propose normalized masked textual inversion that conditions the generation process using as few as one novel sample. To the best of our knowledge, we are the first method that uses Stable Diffusion to improve few-shot segmentation on the challenging FSS-1000 dataset [12]; while samples synthesized with plain text conditioning diminish segmentation performance by over 10%. **2)** To improve annotation-level alignment, we propose a novel mask-generation pipeline to guide mask generation conditioning on a few real samples. Compared to previous methods [31], our proposed method reduces the mask generation time from an average of 40 seconds per image [31] to 0.5 seconds with similar mask quality. **3)** We carry out extensive experiments and show that our method can be easily combined with existing few-shot segmentation methods as an augmented data source to achieve state-of-the-art performance on both few-shot segmentation and the more challenging generalized few-shot segmentation.

2 Related Work

Semantic Segmentation. Semantic segmentation is a dense vision task assigning a semantic label to each pixel in an image. Learning-based semantic segmentation methods can roughly be categorized into two paradigms: per-pixel classification and mask classification. Long *et al*. [16] proposed the diagram for treating semantic segmentation as a per-pixel classification problem for Convolutional Neural Networks (CNNs). Later works then investigated different architectural improvements [2, 3, 34] and applications in 3D [21, 35]. More recently, an alternative paradigm, mask classification, was proposed by [4]. This line of methods uses a detect-first-recognize-later paradigm. Since AlignDiff is a model-agnostic data synthesis method, works in designing network architectures for semantic segmentation are orthogonal to our method.

Few-Shot Semantic Segmentation. To allow segmentation models to operate in the low-data regime, Few-shot Semantic Segmentation (FSS) methods study how to predict segmentation masks of novel classes using only a few training examples of the novel class. Many methods [5, 17, 29, 30, 32] and even specialized datasets such as FSS-1000 [12] have been proposed to investigate this

problem. Besides metric learning [5, 29, 30], recent works also exploit test-time optimization [5, 17] for few-shot segmentation.

Similar to conventional semantic segmentation, AlignDiff is orthogonal to work in FSS as it can be used to augment any proposed architecture in FSS.

Generalized Few-Shot Semantic Segmentation. Recently, generalized few-shot semantic segmentation (GFSS) had been proposed [28] as a more challenging task setting than vanilla few-shot segmentation methods. Compared to few-shot segmentation which produces novel-class-only binary masks, generalized few-shot segmentation tasks models to segment both base and novel classes within query images. Among recent works [1, 18, 20, 22, 28], Tian *et al.* [28] proposed to approach this problem using the test-time optimization scheme from few-shot segmentation, while later works [1, 18, 20] found that fine-tuning the models with few-shot continual learning techniques attain promising performance. We apply AlignDiff to this challenging task setting by combining it with GAPS [20], a recent work on GFSS. Closely related to our work, AnomalyDiffusion [11] learns embedding of defects to generate defective samples with masks of defects, but it requires abundant normal samples, does not consider background context, and generate defects only at certain locations, which is not applicable for general segmentation settings.

Text-to-Image-Mask Generation. Some recent works [14, 19, 31] attempt to modify text-to-image synthesis models into text-to-image-*mask* synthesis models for training segmentation models. Li *et al.* [14] was first in the line and proposed grounded diffusion (GD), a zero-shot segmenter for stable diffusion model [23]. However, the mask quality from GD [14] is not ideal, as we quantitatively verified in experiments. Wu *et al.* [31] proposed to use the intermediate attention maps in the diffusion models to generate coarse masks, which are then refined with a noise learning process. Though DiffuMask [31] generates masks of good quality, it requires heavy manual prompt engineering and its noise learning process is very time-consuming. Specifically, since DiffuMask requires full training of a segmentation model for cross-validation on every category, which results in amortized cost of 40 seconds per generation of an image-mask pair. In stark contrast, AlignDiff generates high-quality masks with on-par accuracy as DiffuMask and much better efficiency - averaging 0.5 seconds per image. Finally, Nguyen *et al.* [19] proposed to use diffuison to distill datasets, but requires significant amount of data for training. In addition, these works [14, 19, 31] are vulnerable to the drawback in Fig. 1 and may generate unfaithful samples for out-of-distribution rare categories.

3 Method

3.1 Preliminary: Text-to-Image Diffusion Models

Our method is based on Stable Diffusion [23], which is an instance of the latent diffusion models that performs diffusion in the latent space with text conditioning. Given a text description $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, where v_i are token embedding

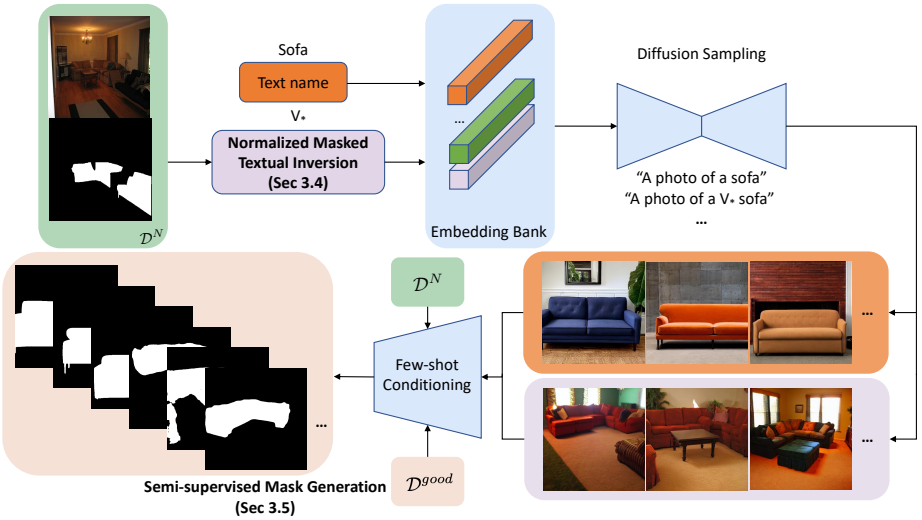

Fig. 2: System overview of AlignDiff. We propose a Normalized Masked Textual Inversion method (Sec. 3.4) to condition the generative process based on as few as a single image, which inverses an image-mask pair to an instance-specific textual embedding. For pixel-level annotation generation, we propose a semi-supervised process that uses both synthetic samples and real samples to generate high-quality masks (Sec. 3.5).

encoded by a text encoding network from texts, starting from $X_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the diffusion sampling process at each step is given by,

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t, \mathbf{v}) \right) + \sigma_t \mathbf{z}\,, \tag{1}$$

where α_t and $\bar{\alpha}_t$ are constants regulating the denoising schedule, ϵ_θ is a denoising U-Net [9] parameterized by θ, σ_t is a constant standard deviation, and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Let $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ be a noisy version of original image x_0. The training process is then given by,

$$\mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t, \mathbf{v}} \left[\left\| \epsilon - \epsilon_\theta(x_t, t, \mathbf{v}) \right\|_2^2 \right]. \tag{2}$$

3.2 Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ be the set of RGB images, $\mathcal{C} \subset \mathbb{N}$ be a set of indexed categories, $\mathcal{M}_y$ be the name of the category, and $\mathcal{Y}^{\mathcal{C}} \subset \mathbb{R}^{H \times W \times |\mathcal{C}|}$ be a set of label masks. Following existing work on general few-shot semantic segmentation [17, 20, 28], we split the set of possible categories into $\mathcal{C}^B$, a set of base categories, and $\mathcal{C}^N$, a set of novel categories. We further assume that the base image-mask dataset $\mathcal{D}^B$ has abundant examples and the novel image-mask dataset $\mathcal{D}^N$ has only a few examples per category (*i.e.*, $|\mathcal{D}^B| \gg |\mathcal{D}^N|$).

The goal of this paper is to synthesize training examples of novel categories to augment few-shot learning. More formally, we want to design a synthesis function $\Phi(\mathcal{M}_y, \mathcal{D}^N) \rightarrow \mathcal{X} \times \mathcal{Y}^{\mathcal{C}}$ that generates synthetic samples to augment $\mathcal{D}^N$ using both texts $\mathcal{M}_y$ and real samples $\mathcal{D}^N$. Note that this is different from previous text-to-image-mask works [14, 31], which focuses on using texts ($\mathcal{M}_y$) only. In summary, AlignDiff focuses on *conditioning* the generative process using $\mathcal{D}^N$ to *align* the synthetic data distribution to the real data distribution.

3.3 Method Overview

The insights and contributions of AlignDiff are centered at **how to condition the generative process using available real samples to align generated data distribution with real distribution**, where we make *full utilization* of both images and masks of novel samples. Fig. 2 provides a high-level overview of AlignDiff. In Sec. 3.4, we discuss how to handle out-of-distribution generation for rare categories using a few novel samples via normalized masked textual inversion. In Sec. 3.5, we describe how we relate mask generation to semi-supervised learning and describe a novel technique to use $\mathcal{D}^N$ to bootstrap such a process, which is shown to be much more efficient than existing method [31] with on-par accuracy.

3.4 Aligned Image Generation via Normalized Masked Textual Inversion

Vanilla large-scale text-to-image diffusion models are not appropriate to be directly used to generate training samples. The reasons are two-fold: 1) text-to-image synthesis models may completely fail for OOD generation of uncommon categories (*e.g.,* samples in Fig. 1), and 2) text-conditioned synthetic samples may not be diverse. For instance, images generated with text prompts 'a photo of a sofa' share a similar pattern of straight upfront views of sofas (upper right corner of Fig. 2), which fails to capture the viewport variations and occlusion like the real-world samples (upper left corner of Fig. 2).

Existing methods [31] approach these two issues via prompt engineering, where hundreds of intra-class vocabularies are manually added to increase data diversity. However, such a method does not scale and it may not handle OOD generation. To address these challenges without handcrafted engineering, we propose *normalized masked textual inversion*. Given an image-mask pair (x, y) and its class name $\mathcal{M}_y$, AlignDiff optimizes for instance-specific textual embeddings for provided instances, which serve as an implicit language description of properties of the novel object in x (*e.g.,* color, the environment it is placed in, orientation, occlusion, etc.). Compared to existing methods for personalizing diffusion models [6, 24], which requires at least 5 close-up images of objects, our normalized masked inversion takes advantage of masks and works with as few as a single image where the object may occupy a small region.

More concretely, we denote the instance-specific embedding as v_* to describe the instance in x and use it as an adjective (*e.g.,* text prompt 'A photo of a v_*

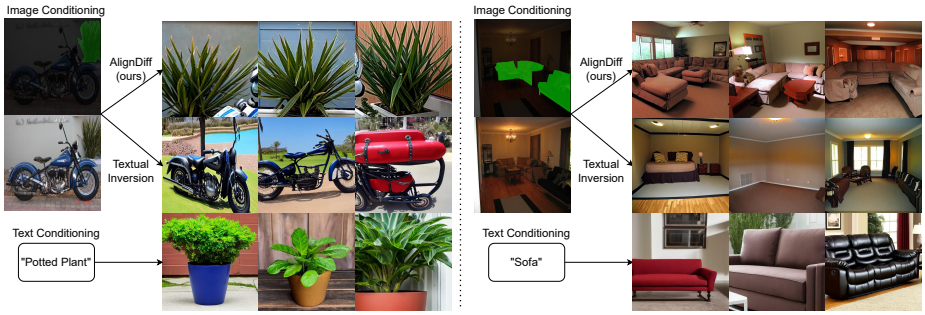

Fig. 3: Comparison between our proposed Normalized Masked Textual Inversion in AlignDiff, Naïve Textual Inversion, and plain text conditioning. Text-conditioned samples exhibit variance in texture and lighting conditioning, but may lack geometric variations and realistic scene layout. Naïve textual inversion captures the realistic scene layout, but often leads to undesired attention of unrelated background when the novel object occupies only a small portion of the image. Note how our proposed Normalized Masked Textual Inversion can successfully generates synthetic samples in such scenario.

sofa' would generate a specific variant of sofa of v_*). To learn the instance-specific embedding v_*, we first use the pre-trained text encoder to map prompt 'A photo of a $\mathcal{M}_y$' to get the language embedding $\mathbf{V} = (v_1, \ldots, v_{j-1}, v_j, v_{j+1}, \ldots, v_n)$. Here, v_j denotes the embedding vector that the determiner 'a' maps to. We then modify this vector to insert the instance embedding to create a new vector, $\mathbf{V}_* = (v_1, \ldots, v_{j-1}, v_j, v_*, v_{j+1}, \ldots, v_n)$, where v_* serves as the adjective description. This is similar to adjective token learning in DreamBooth [24] and is different from textual inversion [6] where the trainable embedding is the noun. We empirically found that treating the learnable embedding as an adjective leads to a faster and more stable training convergence (illustrated in Fig. 3).

The optimization goal of the vanilla textual inversion is given by a modification of Eq. 2, where the only trainable parameter is the embedding v_*. However, this is inappropriate for few-shot segmentation because the loss is distributed evenly across the entire image. For training samples where the objects of interest occupy only a small portion of the image, using simple textual inversion results in the generation of images with an unwanted focus on background. To amend this issue, we propose to mask and balance the loss of foreground and background using the provided mask,

$$v_* = \underset{v_*}{argmin} \mathbb{E}_{(x,y),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,\mathbf{v}} \left[\frac{\sum_{y_i=1} \left\| \epsilon_i - \epsilon_\theta(x_t, t, \mathbf{v})_i \right\|_2^2}{\sum_{y_i=1} 1} + \lambda \cdot \frac{\sum_{y_i=0} \left\| \epsilon_i - \epsilon_\theta(x_t, t, \mathbf{v})_i \right\|_2^2}{\sum_{y_i=0} 1} \right], \tag{3}$$

where i is the index of pixels in the image, $y_i = 1$ denotes foreground pixels, $y_i = 0$ denotes background pixels, and λ is a hyperparameter used to balance the

foreground and background loss. Compared to naïvely masking the loss with foreground masks, Eq. 3 also captures background context such as the surrounding environment and occlusion (an illustration is given in Fig. 2, where the image-conditioned sample shows similar structure to the provided sample). After v_* is optimized, we store v_* to a bank of embedding and sample it along with plain text encodings to generate samples, as illustrated in Fig. 2.

3.5 Aligning Mask Generation via Few-shot Conditioning

A critical component of adapting diffusion models for segmentation training sample synthesis is to generate accurate masks. To exploit the internal representations of Diffusion models for such a purpose, Hertz *et al*. [8] proposed to extract image-text cross-attentions and use the response to category names as masks. However, such attention maps are often imprecise (illustrated in the supplementary material), necessitating mask refinement and filtering techniques. Recently, Wu *et al*. [31] attempted to address this issue by introducing a noise learning process. Though this approach generates masks of good quality, it requires *full training* of segmentation models for every category, which is prohibitively expensive and empirically takes 1 GPU day for a few thousand synthetic samples.

AlignDiff builds upon the work of Hertz *et al*. [8] and proposes a novel technique for refining coarse masks from diffusion models. We investigate the problem from a novel perspective, where we relate the task setting to semi-supervised learning. The key insight of AlignDiff is that the expensive noise learning process can be *largely avoided* if we bootstrap the process with few-shot conditioning with a few accurate image-mask pairs.

Intuitively, we are given two sets initially: D_{good}, a set of a few reference image-mask pairs, and D_{bad}, a set of many synthetic images with *coarse* masks. The goal is to refine masks in D_{bad} using the high-quality masks from D_{good}. This is very similar to the task setting in semi-supervised learning, in which a widely adopted paradigm is to use knowledge from a small set of good labeled data (comparable to D_{good}) to augment the learning on a larger set of data with no label or noisy labels (comparable to D_{bad}).

Algorithm 1 AlignDiff Mask Generation

Require: Coarse samples $D_{bad} = \{I_i, M_i\}_{i=1}^{N}$
Require: Given samples $D_{good} = \{I_i, M_i\}_{i=1}^{M}$
Require: FSS model f_θ
Require: IoU consensus threshold α
  $\theta \leftarrow FSSCond(f, D_{good})$ // Condition FSS
  for i from 1 to N **do** // Scoring
    $\hat{M}_i \leftarrow f_\theta(I_i), I_i \in D_{bad}$
    if $IoU(\hat{M}_i, M_i) \geq \alpha$ **then**
      $D_{good} \leftarrow D_{good} \cup (I_i, M_i)$
      $D_{bad} \leftarrow D_{bad} \setminus (I_i, M_i)$
    end if
  end for
  $\theta \leftarrow FSSCond(f, D_{good})$ // Re-condition
  for i from 1 to $|D_{bad}|$ **do** // Re-estimate
    $\hat{M}_i \leftarrow f_\theta(I_i), I_i \in D_{bad}$
    $D_{good} \leftarrow D_{good} \cup (I_i, \hat{M}_i)$
  end for

We design a mask generation process that iteratively migrates samples from D_{bad} to D_{good}. The detailed algorithm is given in Algo. 1.

More specifically, the algorithm is split into two stages: a scoring stage and a re-estimation stage. We train a few-shot segmentation (FSS) model on the base dataset $\mathcal{D}^B$. During the scoring stage, we condition the FSS model using the initial D_{good}, consisting of a few provided real samples. The conditioned FSS model is then used to predict masks for all samples in D_{bad}. If the IoU between the coarse masks and the pseudo annotation predicted by the FSS model exceeds a certain threshold α, then AlignDiff deems the coarse masks as high-quality and moves it to D_{good}. In the re-estimation stage, AlignDiff reconditions the FSS model using the expanded D_{good} for a more faithful representation. The updated FSS model is then used to generate pseudo labels for all remaining samples in D_{bad}. Our semi-supervised mask generation process is much more efficient than the previous noise learning paradigm [31], as both the FSS conditioning and inference require no optimization.

4 Experiments

To demonstrate the efficacy of our method, we apply AlignDiff to generate data for three commonly used few-shot segmentation datasets to augment state-of-the-art models' segmentation in a plug-and-play manner. We mainly use the FSS-1000 [12] dataset, for its rich class diversity that allows us to evaluate AlignDiff on *uncommon* categories.

4.1 Evaluation setup

Datasets. We follow previous literature in few-shot segmentation [29, 30, 32] and generalized few-shot segmentation [1, 20, 28] to use the Pascal-5^i and the COCO-20^i dataset in both settings. Since the FSS-1000 [12] dataset focuses on class diversity and contain only individual class per image, it is appropriate only for the FSS setting, following existing works [17, 28, 29].

Synthesizer Baselines. We use Grounded Diffusion (GD) [14] as the main synthesis baseline. Another recent work, DiffuMask [31], (1) performs heavy prompt engineering on the Pascal and the COCO dataset, which violates the few-shot setting that the class information is not known beforehand; (2) requires an expensive noise-learning process for every category, which makes comparisons on COCO prohibitively expensive (we approximate $>$ 80 GPU days). Given this practical limitation, we compare the efficiency and the mask accuracy of AlignDiff with DiffuMask only on a subset of the Pascal dataset in Table. 5.

Evaluation Protocol. The experiment on FSS follows the standard episode-based protocol [17, 30]. In both the base training and the few-shot testing stage, the model is presented with episodes that contain a few supporting examples and a query example. The model is tasked to perform binary segmentation of the query sample. In our experiments, we do not modify the base training stage, but we augment the support set $\mathcal{D}^N$ during few-shot testing by supplying extra synthetic samples from AlignDiff, which is conditioned on $\mathcal{D}^N$. Results are average across 1,000 runs. For GFSS, we follow existing work [1, 20] and first

Table 1: Results on FSS-1000 [12] over all 240 testing categories under the 1-shot setting. The OOD (out-of-distribution) classes are determined as classes where samples synthesized by GD [14] diminish the performance (detailed in supplementary). Support set source indicates how the support sets are augmented (*e.g.,* 1R+20S means 1 real sample with 20 synthetic samples). Using simple samples from GD [14] diminishes the final performance due to OD generation. In contrast, our proposed normalized masked textual inversion (T.Inv.) and semi-supervised masking (S.Mask.) in AlignDiff consistently improve novel IoU, indicating the OOD generation capability of AlignDiff.

Method	Support Set Source	Synthesis Setup		Overall IoU	OOD IoU
		Conditioning	Mask		
HSNet [17]	1R	N/A	N/A	86.5	87.5
	1R+20S	Text	GD [14]	81.4	80.2
	1R+20S	Text	S.Mask. **(Ours)**	87.2	86.9
	1R+20S	T.Inv. + Text **(Ours)**	S.Mask. **(Ours)**	**88.3**	**88.2**
VAT [10]	1R	N/A	N/A	90.0	90.5
	1R+20S	Text	GD [14]	84.7	83.5
	1R+20S	Text	S.Mask. **(Ours)**	89.8	88.6
	1R+20S	T.Inv. + Text **(Ours)**	S.Mask. **(Ours)**	**90.8**	**90.6**

train the models on $\mathcal{D}^B$ excluding any novel samples. During few-shot learning, $\mathcal{D}^N$ is presented to the model for adaptation. We apply AlignDiff to synthesize 1,000 training samples per novel class to augment $\mathcal{D}^N$ and report metrics on both the base and the novel classes. Note that the evaluation of the validation set is done in a single pass, which is different from the usual episode-based FSS evaluation scheme [5,17]. The reported results are averaged across multiple folds in a cross-validating fashion. For each fold, we average results from 5 random runs.

4.2 Main Result - Out-of-Distribution (OOD) Generation for Rare Categories on FSS-1000

Besides introducing more diverse samples to few-shot learning of common categories, AlignDiff is most notable for helping with Out-Of-Distribution (OOD) generation of rare categories that plain text conditioning fails. In Tab. 1, we investigate the efficacy of AlignDiff to handle out-of-distribution generation on the FSS-1000 dataset [12] under the 1-shot setting. We use the HSNet [17] and VAT [10] as the FSS baseline and apply AlignDiff to provide additional samples for the support set during the few-shot testing. Note that DiffuMask is prohibitively expensive for this scenario since it requires training segmentation models for every category. Thus, we compare only with GD [14].

We perform **step-by-step ablations** from text-conditioned diffusion [14] to AlignDiff, which demonstrates the efficacy of our method. GD [14], which synthesizes samples using text conditioning, fails drastically on FSS-1000. The samples it synthesizes negatively impact the final performance. This is due to both inaccurate instances in the images caused by plain text conditioning and

Table 2: Comparison with state-of-the-arts on PASCAL-5^i and COCO-20^i in FSS settings. **Bold** and underlined indicate best and the second best methods.

Method	PASCAL-5^i 1-SHOT					PASCAL-5^i 5-SHOT				
	5^0	5^1	5^2	5^3	Mean	5^0	5^1	5^2	5^3	Mean
PFENet (TPAMI'20) [29]	60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
MLC (ICCV'21) [33]	60.8	71.3	61.5	56.9	62.6	65.8	74.9	**71.4**	63.1	68.8
HSNet (ICCV'21) [17]	65.7	70.3	63.2	61.9	65.3	72.0	73.6	68.7	68.4	70.7
NTRENet (CVPR'22) [15]	65.5	71.8	59.1	58.3	63.7	67.9	73.2	60.1	66.8	67.0
DCAMA (ECCV'22) [27]	62.5	70.8	64.5	56.4	63.5	70.0	73.8	66.8	65.0	68.9
VAT (ECCV'22) [10]	68.1	71.7	64.8	63.3	67.0	**72.6**	74.1	69.5	**69.5**	71.4
SCCAN (ICCV'23) [32]	69.1	74.0	66.3	61.6	67.7	71.6	75.2	69.5	66.5	70.7
SCCAN (ICCV'23) [32]+**AlignDiff**	**71.0**	**74.8**	**66.5**	**63.6**	**69.0**	**72.6**	**75.5**	70.3	68.1	**71.6**
	COCO-20^i 1-SHOT					COCO-20^i 5-SHOT				
PFENet (TPAMI'20) [29]	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	37.4
MLC (ICCV'21) [33]	**50.2**	37.8	27.1	30.4	36.4	**57.0**	46.2	37.3	37.2	44.4
HSNet (ICCV'21) [17]	37.6	44.5	44.4	40.7	41.8	45.1	52.3	48.5	47.9	48.5
NTRENet (CVPR'22) [15]	38.3	40.4	39.5	38.1	39.1	42.3	44.4	44.2	41.7	43.2
DCAMA (ECCV'22) [27]	41.5	46.2	45.2	41.3	43.5	48.0	58.0	54.3	47.1	51.9
SCCAN (ICCV'23) [32]	41.7	51.3	48.4	46.7	47.0	49.0	59.3	59.4	52.7	55.1
SCCAN (ICCV'23) [32]+**AlignDiff**	42.1	**52.5**	**49.0**	**47.8**	**47.9**	49.3	**59.9**	**59.5**	**53.0**	**55.4**

inaccurate mask generation. When we use our proposed semi-supervised mask generator, the overall IoU is improved since the masks provided are more accurate. However, the IoU of OOD categories is still worse than conditioning using a single real example even with accurate masks. Finally, we use full AlignDiff and apply the normalized masked textual inversion technique. In this case, both the overall IoU and the IoU of OOD categories surpass the original performance, highlighting the capability of AlignDiff to generate synthetic samples for rare out-of-distribution categories.

4.3 Few-Shot Segmentation (FSS)

In Tab. 2, we present results of combining AlignDiff with a recent method, SCCAN [32], to tackle the few-shot segmentation tasks on the Pascal-5^i and the COCO-20^i datasets. As expected, since Pascal and COCO contain only common classes, AlignDiff successfully augments the performance on both datasets. This marks that AlignDiff can augment state-of-the-art networks in all commonly used datasets including Pascal-5^i, COCO-20^i, as well as the FSS-1000 [12] that is challenging to simple diffusion models.

4.4 Generalized Few-Shot Segmentation (GFSS)

In Tab. 3, we report results on the challenging GFSS setting on the Pascal-5i^i dataset and the COCO-20^i datasets, where the model is required to segment both the base and novel categories. We augment several baselines with AlignDiff to illustrate how much the AlignDiff can help improve the performance of few-shot segmentation in a model-agnostic manner.

Table 3: AlignDiff can be applied in a model-agnostic fashion to GFSS, where models are required to segment both base and novel classes. AlignDiff improves underlying models across different few-shot settings on COCO-20^i and PASCAL-5^i. The best results are **bolded**. HM stands for harmonic mean. *: Simple fine-tuning yields bad performance due to catastrophic forgetting [1].

Method	Base	Novel	HM	Base	Novel	HM
	PASCAL-5^i 1-SHOT			PASCAL-5^i 5-SHOT		
PIFS [1]	64.1	16.9	26.7	64.5	27.5	38.6
GFS [28]	65.7	15.1	24.6	66.1	22.4	33.5
FINETUNE*	47.2	3.9	7.2	58.7	7.7	13.6
FINETUNE + GD [14]	28.1	20.3	23.6	32.0	22.5	26.4
FINETUNE+**ALIGNDIFF**	66.2(+19.0)	**44.9(+41.0)**	**53.5(+46.3)**	65.9(+7.2)	45.1(+37.4)	53.6(+40.0)
GAPS [20]	66.8	23.6	34.9	68.2	43.9	53.4
GAPS + GD† [14]	66.8	41.2	51.0	**68.5**	44.0	53.6
GAPS+**ALIGNDIFF** [20]	**67.3(+0.5)**	43.3(+19.7)	52.7(+17.8)	68.4(+0.2)	**47.4(+3.5)**	**56.0(+2.6)**
	COCO-20^i 1-SHOT			COCO-20^i 5-SHOT		
PIFS [1]	40.4	10.4	16.5	41.1	18.3	25.3
GFS [28]	44.6	7.1	12.2	45.2	11.1	17.8
FINETUNE*	38.5	4.8	8.5	39.5	11.5	17.8
FINETUNE + GD [14]	25.8	17.2	20.6	31.9	23.6	27.2
FINETUNE+**ALIGNDIFF**	41.7(+3.2)	22.4(+17.6)	29.1(+20.6)	41.8(+2.3)	27.9(+16.4)	33.5(+15.7)
GAPS [20]	**46.8**	12.7	20.0	**49.1**	25.8	33.8
GAPS + GD† [14]	47.1	21.8	29.9	46.5	29.0	35.7
GAPS+**ALIGNDIFF** [20]	46.7(-0.1)	**23.1(+10.4)**	**30.9(+10.9)**	47.9(-1.2)	**30.3(+4.5)**	**37.1(+3.3)**

GFSS Baselines. We use three recent works [1, 20, 28] from GFSS as baselines. GFS [28] proposed the setting of GFSS, but it focuses only on test-time optimization and does not fine-tune on $\mathcal{D}^N$. PIFS [1] formulates GFSS as a continual few-shot learning task and proposes to fine-tune the model using a few novel samples. Finally, GAPS [20] views GFSS as continual learning and combines memory-replay with copy-paste to further increase the GFSS performance. In this work, we choose GAPS [20] as the main baseline due to its state-of-the-art performance in GFSS and we treat images from AlignDiff as instances that can be copied and pasted in GAPS.

AlignDiff consistently improves novel IoU under few-shot settings. Across all task settings, AlignDiff is able to synthesize more diverse samples to aid few-shot learning of novel categories. Most notably, on the impoverished 1-shot setting, GAPS+AlignDiff improves the novel IoU by approximately 80% on both Pascal-5i^i and COCO-20^i. AlignDiff also consistently outperforms GD [14]. Note that the performance gap between AlignDiff and GD is more drastic on simple fine-tuning because GAPS [20] has built-in copy-paste for scene layout, which **highlights the potential of AlignDiff for standard segmentation via copy-pasting generated samples**.

4.5 Ablation Studies

Qualitative samples. We perform extensive visualization to demonstrate the quality of synthetic samples that AlignDiff generates. We include some qualita-

Fig. 4: Qualitative samples for AlignDiff on FSS-1000 [12]. Note how text-to-image synthesis fails for rare classes. AlignDiff generates instances within class and with varying lighting conditions. (From top to bottom: Samarra Mosque, phonograph, Pidan, American Chameleon, and chess queen).

tive examples in Fig. 1 and Fig. 4. These qualitative examples demonstrate the capability of AlignDiff on OOD categories from FSS-1000, where plain text conditioning fails. Notice that AlignDiff is capable of picking up fine-grained details of different types of chess in the synthesized image and textures.

Comparison with DiffuMask [31]. In Table. 5, we compare AlignDiff with DiffuMask on the 5 classes of the Pascal-5^3 dataset on the GFSS setting with GAPS [20]+AlignDiff. We empirically measure the GPU hours for synthesis and use the final novel IoU on the GFSS task setting as a quantitative measurement of the mask quality. We observe that the mask quality of AlignDiff and DiffuMask is on-par with similar accuracy, whereas AlignDiff is much more efficient.

Dissection of Mask-level Alignment. Our mask-level alignment is based on a key observation: both cross-attention maps from Diffusion and masks predicted by FSS models (conditioned on a few real images) are not perfect, but **they have different failure patterns**. Cross-attention maps have strong generalizability but with unrefined boundary. Conversely, recent FSS models have refined boundary, but lack generalizability due to the limited real samples. Our scoring stage captures masks where both models agree on *with good boundary and intra-class variance.*

Table 4: Comparisons of different mask generation methods using Pascal-5^2 categories. An off-the-shelf model trained on abundant data [4] provides reference labels.

Method	Images	IoU
Coarse Attention [8, 31]	All	78.6
FSS-only [5]	All	88.9
AlignDiff	All	**92.9**
Coarse Attention [8, 31]	$\mathcal{D}_{\text{good}}$	91.4
FSS-only [5]	$\mathcal{D}_{\text{good}}$	93.7
AlignDiff	$\mathcal{D}_{\text{good}}$	**94.2**

Table 5: Comparison of AlignDiff to DiffuMask [31] on mask generation on Pascal-5^3 GFSS. GPU hours are reported for the generation of masks for a total of 5 classes with 2,000 images per class on a single RTX3090. AlignDiff is much more efficient with on-par efficiency.

Method	GPU hrs↓	Novel IoU↑
AlignDiff	**15**	**41.5**
DiffuMask [31]	130	40.7

To validate the above claim, we perform additional experiments to evaluate the quality of masks throughout the mask alignment process. Specifically, we use a 2D segmentation model trained on abundant data as an oracle to provide reference masks for data synthesized for Pascal-5^2. The results are presented in Tab. 4: coarse attention maps and FSS-only are unsatisfactory due to failure modes mentioned above, but AlignDiff yields significantly improved masks. Notably, both models generate near-perfect initial estimation for images in the post-scoring $\mathcal{D}_{\text{good}}$ set ($\mathcal{D}_{\text{good}}$ here is discussed in Algo. 1).

Timing. We compare the timing of AlignDiff and DiffuMask [31] in Tab. 5. For both methods, generating 10,000 images require ≈14 GPU hours using Stable Diffusion [23]. For mask generation, AlignDiff's semi-supervised pipeline only requires single-step feedforward conditioning and inference with less than 0.3 seconds per image. DiffuMask [31] requires training of segmentation models per category with amortized cost of ≈ 40 seconds per image.

Supplementary Material For more results, such as more iamges of generated samples and masks generated by coarse attention/AlignDiff to show mask improvement, please refer to the supplementary material.

5 Conclusion

In this paper, we present AlignDiff, which aligns standard Diffusion models for out-of-distribution generation and accurate pixel-level annotations. AlignDiff can be combined with copy-paste [7, 20] and recent conditioning methods [13] to improve general segmentations, which we leave for future research.

Limitation. Though AlignDiff can adapt to synthesize instances of rare categories that plain text conditioning fails with as few as a single image, AlignDiff may fail on tasks with drastic domain gaps, such as medical image segmentation.

Societal Impact. AlignDiff was developed using Stable Diffusion [23], which is trained on datasets known to contain societal biases such as gender bias [26].

Acknowledgement. This work was supported in part by NIFA Award 2020-67021-32799 and NSF Grant 2106825. This work used NVIDIA GPU at NCSA Delta via allocations CIS220014 and CIS230012 from the ACCESS program.

References

1. Cermelli, F., Mancini, M., Xian, Y., Akata, Z., Caputo, B.: Prototype-based incremental few-shot semantic segmentation. In: BMVC (2021) 4, 9, 12
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 3
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 3
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. NeurIPS (2021) 3, 14
5. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: ECCV (2022) 1, 3, 4, 10, 14
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 3, 6, 7
7. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR (2021) 14
8. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 8, 14
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020) 5
10. Hong, S., Cho, S., Nam, J., Lin, S., Kim, S.: Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In: ECCV (2022) 10, 11
11. Hu, T., Zhang, J., Yi, R., Du, Y., Chen, X., Liu, L., Wang, Y., Wang, C.: Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024) 4
12. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: FSS-1000: A 1000-class dataset for few-shot segmentation. In: CVPR (2020) 3, 9, 10, 11, 13
13. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023) 14
14. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Guiding text-to-image diffusion model towards grounded generation. arXiv preprint arXiv:2301.05221 (2023) 4, 6, 9, 10, 12
15. Liu, Y., Liu, N., Cao, Q., Yao, X., Han, J., Shao, L.: Learning non-target knowledge for few-shot semantic segmentation. In: CVPR (2022) 11
16. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015) 3
17. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV (2021) 1, 3, 4, 5, 9, 10, 11
18. Myers-Dean, J., Zhao, Y., Price, B., Cohen, S., Gurari, D.: Generalized few-shot semantic segmentation: All you need is fine-tuning. arXiv preprint arXiv:2112.10982 (2021) 4
19. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. NeurIPS (2023) 4
20. Qiu, R.Z., Chen, P., Sun, W., Wang, Y.X., Hauser, K.: GAPS: Few-shot incremental semantic segmentation via guided copy-paste synthesis. In: CVPRW (2023) 1, 4, 5, 9, 12, 13, 14

21. Qiu, R.Z., Sun, Y., Marques, J.M.C., Hauser, K.: Real-time semantic 3d reconstruction for high-touch surface recognition for robotic disinfection. In: IROS (2022) 3
22. Qiu, R.: Towards real-time robotics perception with continual adaptation. Ph.D. thesis, University of Illinois at Urbana-Champaign (2023) 4
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 3, 4, 14
24. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 6, 7
25. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022) 2
26. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022) 14
27. Shi, X., Wei, D., Zhang, Y., Lu, D., Ning, M., Chen, J., Ma, K., Zheng, Y.: Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation. In: ECCV (2022) 11
28. Tian, Z., Lai, X., Jiang, L., Liu, S., Shu, M., Zhao, H., Jia, J.: Generalized few-shot semantic segmentation. In: CVPR (2022) 1, 4, 5, 9, 12
29. Tian, Z., Zhao, H., Shu, M., Yang, Z., Li, R., Jia, J.: Prior guided feature enrichment network for few-shot segmentation. TPAMI (2020) 1, 3, 4, 9, 11
30. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV (2019) 3, 4, 9
31. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: ICCV (2023) 3, 4, 6, 8, 9, 13, 14
32. Xu, Q., Zhao, W., Lin, G., Long, C.: Self-calibrated cross attention network for few-shot segmentation. In: ICCV (2023) 3, 9, 11
33. Yang, L., Zhuo, W., Qi, L., Shi, Y., Gao, Y.: Mining latent classes for few-shot segmentation. In: ICCV (2021) 11
34. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017) 3
35. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: ICCV (2021) 3