Flaws can be Applause: Unleashing Potential of Segmenting Ambiguous Objects in SAM

Chenxin Li[1]*, **Yuzhi Huang**[2]*, **Wuyang Li**[1], **Hengyu Liu**[1],
Xinyu Liu[1], **Qing Xu**[3], **Zhen Chen**[4], **Yue Huang**[2], **Yixuan Yuan**[1]†
[1]The Chinese University of Hong Kong [2]Xiamen University
[3]University of Nottingham Ningbo China [4]Yale University

Abstract

As the vision foundation models like the Segment Anything Model (SAM) demonstrate potent universality, they also present challenges in giving ambiguous and uncertain predictions. Significant variations in the model output and granularity can occur with simply subtle changes in the prompt, contradicting the consensus requirement for the robustness of a model. While some established works have been dedicated to stabilizing and fortifying the prediction of SAM, this paper takes a unique path to explore how this flaw can be inverted into an advantage when modeling inherently ambiguous data distributions. We introduce an optimization framework based on a conditional variational autoencoder, which jointly models the prompt and the granularity of the object with a latent probability distribution. This approach enables the model to adaptively perceive and represent the real ambiguous label distribution, taming SAM to produce a series of diverse, convincing, and reasonable segmentation outputs controllably. Extensive experiments on several practical deployment scenarios involving ambiguity demonstrates the exceptional performance of our framework. Project page: `https://a-sa-m.github.io/`.

1 Introduction

The advent of Visual Foundational Models (VFMs) such as the Segment Anything Model (SAM) [22] has been unprecedented, largely due to the availability of vast datasets and computational resources. These models have exhibited remarkable generalization capabilities in zero-shot scenarios and the capacity to interact with human feedback. SAM, in particular, employs a specialized data engine to manage 11 million image masks, using a unique prompt-based segmentation framework to generate accurate masks for any object within a visual context, largely extending the capacity and generality of segmentation [27]. Such successes have been widely extended to various domains, such as medical imaging analysis [8, 35, 28], remote sensing [48], etc.

However, it has been observed that SAM suffers from severe *predictive ambiguity* to segment desired concepts [22] in practical scenarios. To uncover the factual basis, we delve into this ambiguity and detail it into two flaws according to experimental insight. Specifically, the first flaw lies in that SAM prediction is sensitive to slightly different prompt variants. As shown in Fig. 1 (a), we ground SAM into a realistic clinical scenario to segment the lesion in CT images. Even though three medical experts uniformly give rational box prompts covering the lesion, SAM, unfortunately, makes enormous differences among them, even including one wrong case. Further, we provide detailed statistics regarding IoU over the small perturbation (only 5 pixels) of box prompts, as shown in Fig. 1

*Equal Contribution: {chenxinli@link.cuhk.edu.hk, yzhuang13@stu.xmu.edu.cn}.
†Corresponding Author: {yxyuan@ee.cuhk.edu.hk}.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1449

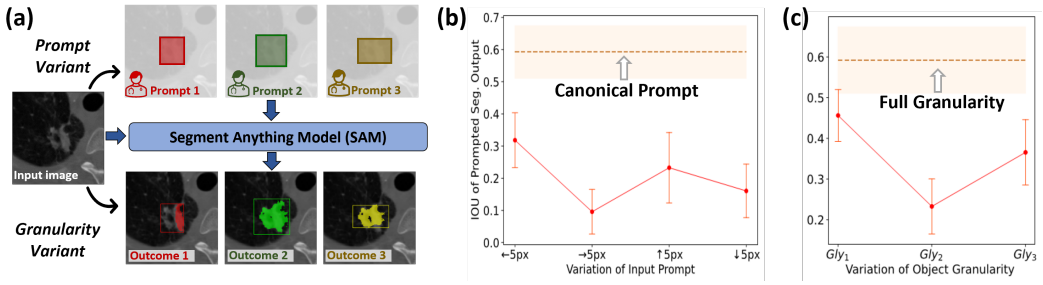

Figure 1: **Analysis of Inherent Ambiguity in SAM**. **(a)**: Feeding SAM with slightly different prompts from multiple experts for a single image can significantly alter the segmentation output. **(b)(c)**: We evaluate SAM using canonical box prompts and various perturbed versions, measuring the mean and variance of segmentation IoU on LIDC. Perturbations involve shifting the box five pixels in different directions and employing various granular outputs within SAM. Results highlight SAM's sensitivity to prompt variations and granularity.

(b). We can observe a significant IoU fluctuation over small prompt differences, which reveals that SAM prediction is highly sensitive to the prompt variants caused by such small perturbation[3].

The second flaw lies in the susceptibility of SAM output to the inherent structural granularity of the object. Despite the fact that VFMs like SAM gain generalizable knowledge and abilities from extensive datasets, they frequently forfeit the capacity to segment specific visual concepts as they are class agnostic, making the model unable to discern the difference between objects with different levels of semantic granularity. Consequently, for targets that are challenging to define, particularly those with rich internal hierarchical granularity, they are inclined to produce multiple candidate results at different granularities and amalgamate them, rather than directly outputting a definitive result. As depicted in Fig. 1 (c), we also discern that the multiple candidate results captured by SAM often exhibit significant differences, and the segmentation precision of outputs at different granularities diverges greatly when compared to the final integrated SAM output that incorporates multiple candidates.

Nonetheless, every coin has two sides. While these observations reveal SAM's faults regarding the output sensitivity, we explore a different perspective: *could this sensitivity flaw become an advantage in other cases, such as ambiguous segmentation, which requires the model to learn from a set of ambiguous object annotations caused by imaging noise, ambiguous tissue boundaries, and different annotator preferences?* Specifically, we are particularly interested in the following: ❶ Given the segmentation result's sensitivity to prompts, can we probabilistically model this prompt variation to tame prompt-sensitive SAM to controllably yield multiple likely results close to the actual fuzzy distribution? ❷ Considering the segmentation result's sensitivity to object structure, can we probabilistically model this granularity variation to tame granularity-sensitive SAM to controllably yield multiple likely results close to the actual ambiguous distribution?

Driven by these questions, we propose an innovative strategy, which flips the inherent category-agnostic ambiguity induced in SAM into a controlled ability to generate a range of feasible results for ambiguous segmentation tasks. Specifically, to simulate segmentation ambiguity under different prompts, we introduce context-aware prompt ambiguity modelling. This method probabilistically models the uncertainty of the prompts inputted into SAM using a latent learnable distribution, which adaptively perceives the specific ambiguity of different contexts. Furthermore, to simulate ambiguity caused by complex object structures at different granularities, we introduce granularity-aware object ambiguity modelling. This method introduces and enhances the visual ambiguity of objects at different granularities into SAM's original image embedding via a learnable embedding distribution. While establishing these two levels of ambiguity modelling, we introduce an efficient optimization strategy based on posterior constraints, allowing the model to mimic models that can perceive the actual ambiguous distribution. Our contributions can be summarized as follows:

[3]In practice, such perturbation commonly exists and cannot be controlled by users, even for medical experts.

https://doi.org/10.52202/079017-1449

- We explore the inherent ambiguity in SAM to flip this commonly seen disadvantage in deterministic segmentation tasks into an advantage for more practical ambiguous segmentation tasks that allow multiple possible outputs.
- We propose a $\mathcal{A}$-*SAM* framework that employs a learnable latent distribution to encapsulate the ambiguity at two strata, from prompts and object granularity.
- We introduce an optimization architecture based on variational autoencoders, which effectively represents ambiguity by constraining the sample embedding to align with those from a series of practically feasible annotations.
- Rigurous benchmarked experiments across a wide range of potential scenarios demonstrate that our method produces more accurate, diverse, and reasonable segmentation outputs.

2 Related Work

Prompting Foundational Models for Segmentation. There has been a surge in the advancement of large-scale vision models for image segmentation, drawing inspiration from language foundation models [62, 4, 25, 39]. Segmentation Foundation Models (SFMs), such as the Segment Anything Model (SAM) [22] and SEEM [66], have delivered significant segmentation results across various downstream datasets [17, 38, 54]. SAM, leveraging a data engine incorporating model-in-the-loop annotation, learns a promotable segmentation framework that generalizes to downstream scenarios in a zero-shot manner. Meanwhile, other models like Painter [56] and SegGPT [57] introduce a robust in-context learning paradigm that enables segmenting any images given an image-mask prompt. On the contrary, SEEM [66] presents a general segmentation model prompted by multi-modal references, such as language and audio, thereby incorporating a wide range of semantic knowledge. These advancements in SFMs, driven by *promptable segmentation* design, involve two types of prompts: semantic prompts (e.g., free-form texts) and spatial prompts (e.g., points or bounding boxes) [22, 57, 40, 41, 43, 26].

Recently, the practice of adapting vision foundation models such as SAM [22] for application in medical image segmentation is garnering increasing interest [46, 10, 12, 44, 33]. A prevalent and cost-effective approach involves adapter techniques, which necessitate the inclusion of bottleneck modules with a finite number of parameters within the model. By fine-tuning these diminutive adapters, SAM can bridge the domain discrepancy between medical and natural images while preserving stellar performance. For instance, models like MSA [58], SAM-Med2D [9], and SAM-adapter [6] utilize adapter strategies to transfigure SAM for medical imaging, thereby achieving significant segmentation outcomes. Despite these advancements, acquiring suitable prompts for SFMs remains largely under-explored. This work aims to explore generating effective prompts for SAM, focusing on harnessing pre-training knowledge to complete ambiguous image segmentation.

Ambiguous Image Segmentation. Ambiguous image segmentation aims to model a range of, rather than single, segmentation labels [31, 65, 45]. A wealth of existing research has proposed various techniques to quantify uncertainty. Initial research focused on enhancing a traditional U-Net[51, 16, 29, 5] with a probabilistic component to generate multiple predictions for the same image. This was typically achieved by incorporating a conditional variational autoencoder (cVAE) [53], with the low-dimensional latent space encoding potential segmentation variations. Subsequent work further extended this setup to a hierarchical variant [3, 24, 64, 15]. Other research has utilized normalizing flows to allow for distribution in the cVAE [52, 55] to represent a discrete latent space [49] or incorporated variational dropout and directly used inter-grader variability as a training target. Several other methods [34] do not rely on the Probabilistic U-Net [47, 21, 32, 11, 60]. Monteiro *et al.* [47] proposed a network utilizing a low-rank multivariate normal distribution to model the logit distribution. Kassapis *et al.* [21] leveraged adversarial training to learn potential label maps based on the logits of a trained segmentation network. Zhang *et al.* [63] employed an autoregressive PixelCNN to model the conditional distribution between pixels [37, 36]. Finally, Gao *et al.* [13] used a mixture of stochastic experts, where each expert network estimates a mode of uncertainty, and a gating network predicts the probabilities of an input image being segmented by one of the experts. Unlike previous efforts [1, 42, 59, 7], our approach signifies the first exploration of leveraging the inherent properties in vision foundation models for ambiguous image segmentation.

https://doi.org/10.52202/079017-1449

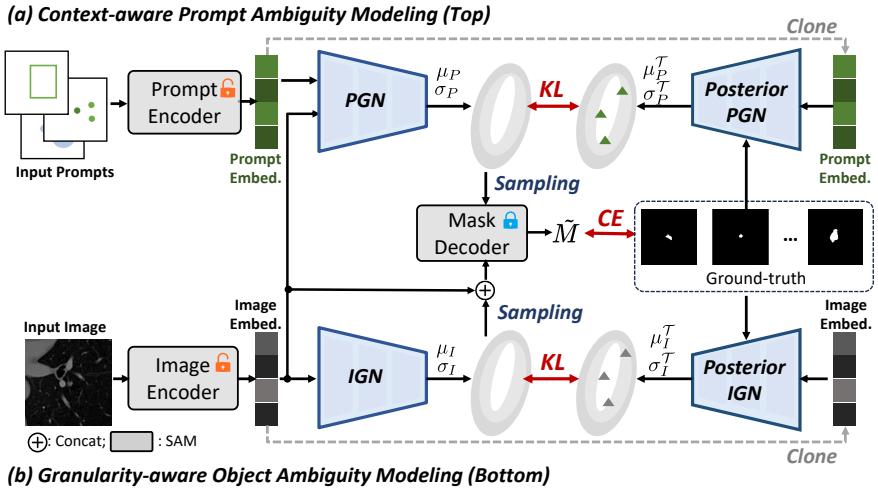

Figure 2: ***A-SAM* Training Pipeline.** We probabilistically model the prompt and object-level ambiguity by jointly probabilities the SAM embeddings with PGN and IGN, respectively.

3 Method

3.1 Revisiting SAM by Probabilistic Perspective

SAM: Segment Anything Model. Given an image I and a set of user-given prompts P, which could be a point, a box, or a rough mask, the Segment Anything Model (SAM) [22] employs a vision transformer-based image encoder Enc_I to extract salient image feature F_I and deploy a prompt encoder Enc_P with length k to encode prompt embeddings T_P, which are denoted as follows,

$$F_I = \mathrm{Enc}_I(I), \quad T_P = \mathrm{Enc}_P(P), \tag{1}$$

where $F_I \in \mathbb{R}^{h\times w\times c}$ and $T_P \in \mathbb{R}^{k\times c}$, where the resolution of the image feature map is represented by h, w, and the feature dimension is denoted by c. Subsequently, the encoded image and prompts are introduced into the decoder Dec_M for interaction based on attention mechanisms. SAM constructs the decoder's input tokens by concatenating several learnable mask tokens T_M as prefixes to the prompt tokens T_P. These mask tokens are accountable for generating the mask output M

$$M = \mathrm{Dec}_M(F_I,\, T_P,\, T_M), \tag{2}$$

***A-SAM*: Lifting SAM to Distributional Space.** Unlike the one-to-one deterministic mapping in SAM, we formulate a probabilistic latent distribution to enable the one-to-many ambiguous mapping named *A-SAM*, with each observation being a sample from this hidden distribution. To this end, the prompt and image embedding can be probabilistically formulated as a distribution:

$$\tilde{T_P} \sim \mathcal{P}_P(\Theta), \quad \tilde{F_I} \sim \mathcal{P}_I(\Phi), \tag{3}$$

where $\mathcal{P}_P$ and $\mathcal{P}_I$ denote a latent distribution for prompt embedding and image embedding, respectively. $\tilde{T_P}$ and $\tilde{F_I}$ denotes a prompt at one sampling from the defined latent distribution at one time. Formally, by implementing multiple rounds of sampling, we can construct a distributional mapping of segmentation outputs with respect to their prompts, formulated as the format of expectation,

$$\tilde{M} = \mathrm{Dec}_M\left(\tilde{F_I},\, \tilde{T_P},\, T_M\right), \quad st.\ \tilde{T_P} \sim \mathcal{P}_P(\Theta), \tilde{F_I} \sim \mathcal{P}_I(\Phi), \tag{4}$$

where $\tilde{M}$ denotes the SAM output corresponding to one prompt sampling, which can also be interpreted as the sampling from a virtual distribution $\mathcal{P}_M(\Omega)$ for the segmentation results obeying parameters Ω. As a result, we can construct an optimized probability distribution $\tilde{T_P} \sim \mathcal{P}_P(\Theta)$ and $\tilde{F_I} \sim \mathcal{P}_I(\Phi)$ by narrowing the gap between $\tilde{M} \sim \mathcal{P}_M(\Omega)$ and ground-truth distribution.

***A-SAM*: Inference Stage.** After the training, *A-SAM* can model two types of latent distribution, representing the ambiguity of the prompt variation and the varied object granularity, respectively.

Based on this formulation, each latent sample drawn from the distribution represents a segmentation candidate. Concretely, to predict a set of m segmentations, we apply the network m times to the same input image. In each iteration $i \in \{1, \ldots, m\}$, we draw a random sample regarding prompt embedding $\tilde{T_P} \in \mathbb{R}^{N_P}$ and image embedding $\tilde{F_I} \in \mathbb{R}^{N_I}$ respectively from $\mathcal{P}_P(\Theta)$ and $\mathcal{P}_I(\Phi)$. The final prediction maps $\tilde{M} \sim \mathcal{P}_M(\Omega)$ can be obtained by Eq. 4. In what follows, we will primarily focus on the methodology of training our overall ambiguous segmentation framework.

3.2 Context-aware Prompt Ambiguity Modeling

Distributional Prompt Representation. To model the distribution of prompt embedding, it is imperative to estimate the parameters Θ of this distribution. We adopt an axisymmetric Gaussian distribution to characterize the prompt embedding, which is dictated by two crucial parameters, including mean μ and standard deviation σ. Then, we can sample a prompt embedding from the given Gaussian distribution, which is shown as

$$\tilde{T_P} \sim \mathcal{P}_P(\Theta) = \mathcal{N}(\mu_P, \mathrm{diag}(\sigma_P)), \tag{5}$$

where μ_P and σ_P denotes the parameters characterized for prompt P. μ_P and σ_P respectively denote the mean and standard deviation of the Axis Gaussian Distribution generated by the network, where $\mu_P, \sigma_P \in \mathbb{R}^{N_P}$. This simple yet effective formulation enables the discrete prompt to be continuously represented in the probabilistic latent space, making the uncertainty estimation available.

Context-aware Prompt Embedding Generation. To parameterize the latent prompt distribution, we propose a prompt generation network (PGN) to effectively model the aforementioned Gaussian related to prompt embedding. This network is simply designed to include several convolution blocks. Considering the variation in salient regions within the image context, the required prompt positions and sizes should also be varied. Therefore, we incorporate image context as prior knowledge into PGN during the forward inference process. By integrating this prior knowledge, the network can customize a unique prompt-associated axial Gaussian distribution for each image I, thereby achieving adaptive and infinite sampling in the latent prompt distribution:

$$[\mu_P, \sigma_P] = F_{PGN}(T_P, F_I; \Theta), \tag{6}$$

where the parameters of the prompt generation network F_{PGN} are modeled by the parameters Θ of our desired distribution, as each set of generated mean and variance uniquely specifies a distribution. Previous research indicates that allowing the model to conditionally perceive the ground truth label distribution during the training phase enhances training stability for such tasks that exhibit significant uncertainty. Thus, a posterior version for prompt generation network F^{post}_{PGN}, parameterized by $\Theta^{\mathcal{T}}$, is further introduced during training, learning to generate the effective distribution for prompt embedding when accessing the ground-truth label distribution:

$$[\mu_P^{\mathcal{T}}, \sigma_P^{\mathcal{T}}] = F^{post}_{PGN}(T_P, F_I, GT; \Theta^{\mathcal{T}}) \tag{7}$$

We only employ this posterior network during training and guide the standard network, which cannot perceive the true labels during testing, to achieve viable performance via a KL loss.

$$\mathcal{L}_{PKL} = D_{\mathrm{KL}}\Big(\mathcal{N}(\mu_P^{\mathcal{T}}, \mathrm{diag}(\sigma_P^{\mathcal{T}})) \parallel \mathcal{N}(\mu_P, \mathrm{diag}(\sigma_P))\Big) \tag{8}$$

3.3 Granularity-aware Object Ambiguity Modeling

Distributional Object Representation. The model integrates object ambiguity from a probabilistic perspective, enhancing its ability to solve problems in innovative ways. We instantiate various segmentation labels to represent ambiguity levels and incorporate them as priors in visual feature extraction, influencing object-related feature modeling for the input image I,

$$\tilde{F_I} = \mathrm{Concat}(F_I, \tilde{F_I\prime}), \quad \tilde{F_I\prime} \sim \mathcal{P}_I(\Phi) = \mathcal{N}(\mu_I, \mathrm{diag}(\sigma_I)), \tag{9}$$

where μ_I and σ_I respectively denote the mean and standard deviation of the Axis Gaussian Distribution generated by the network, where $\mu_I, \sigma_I \in \mathbb{R}^{N_I}$. This approach enhances our understanding and depiction of object diversity and uncertainty in a more profound and lucid manner, enabling better adaptability and handling of variation and uncertainty.

https://doi.org/10.52202/079017-1449

Granularity-aware Image Embedding Generation. SAM generates multiple candidate segmentation masks for the same object at different granularities and levels, demonstrating the inherent ambiguity of SAM related to object granularity. This inspires us to leverage this aspect to enhance the model's perception of ambiguous objects. We subsequently introduce an image generation network (IGN) for object embedding to model this distribution:

$$[\mu_I, \sigma_I] = F_{IGN}(F_I; \Phi), \tag{10}$$

where the parameters of the image generation network F_{IGN} are modeled by the parameters Φ of our desired distribution, as each set of generated mean and variance uniquely specifies a distribution. Similar to the previous introduction of a posterior network to enhance learning in modeling prompt ambiguity, we introduce a posterior version of IGN, denoted as F_{IGN}^{post}, for perceptible labels.

$$[\mu_I^{\mathcal{T}}, \sigma_I^{\mathcal{T}}] = F_{IGN}^{post}(F_I, GT; \Phi^{\mathcal{T}}) \tag{11}$$

We only employ this posterior network during training and guide the standard network, which cannot perceive the true labels during testing, to achieve viable performance via a KL loss.

$$\mathcal{L}_{IKL} = D_{\mathrm{KL}}\Big(\mathcal{N}(\mu_I^{\mathcal{T}}, \mathrm{diag}(\sigma_I^{\mathcal{T}})) \parallel \mathcal{N}(\mu_I, \mathrm{diag}(\sigma_I))\Big) \tag{12}$$

3.4 Overall Optimization

In line with current SAM techniques that generate segmentation masks for the same object at different granularity levels, the present approach utilizes this feature to enhance the model's capture of multi-level object concepts through a common ensemble strategy. Specifically, given the multiple candidate outputs from SAM, represented as $\{\tilde{M}^1, \tilde{M}^2..., \tilde{M}^n\}$, where n is the number of scales. By introducing a set of learnable mask weights $\mathcal{W} = \{w_1, w_2..., w_n\} \in \mathbb{R}^n$, the final mask output can be fine-tuned and obtained through a weighted sum calculation:

$$\tilde{M} = \Sigma_{i=1}^{n} w_i \odot \tilde{M}^i, \tag{13}$$

where $w_1, w_2, ..., w_n$ are initialized to $\frac{1}{n}$ and subsequently fine-tuned to enable the model effectively being aware of the object scales. By adaptively integrating multiple scale masks, the model's perception and modeling capabilities for complex target diversity are further enhanced.

When the representation from SAM is combined with the ground-truth segmentation GT from the training samples, a guide-providing teacher prediction segmentation $\tilde{M}^{\mathcal{T}}$ is created. A cross-entropy loss $CE(\cdot, \cdot)$ is employed to penalize the discrepancies between the distribution of $\tilde{M}^{\mathcal{T}}$ and GT, i.e., $\mathcal{P}_M(\Omega)$ and $\mathcal{P}_{GT}$, where the distribution of GT is a constant value that does not need to be parameterised, as: $\mathcal{L}_{Seg} = CE(GT, \tilde{M})$. Additionally, the KL losses introduced to regularize training in prompt ambiguity and image ambiguity, respectively, in Eq. 8 and Eq. 12, are amalgamated into a weighted sum with weight coefficient of α_P and α_I:

$$\mathcal{L}_{All} = \mathcal{L}_{Seg} + \alpha_P \cdot \mathcal{L}_{PKL} + \alpha_I \cdot \mathcal{L}_{IKL}. \tag{14}$$

The model is trained from scratch using randomly initialized weights. Parameters requiring update include the prompt encoder and image encoder within SAM as well as PGN, posterior PGN, IGN, and posterior IGN. The KL loss during training aligns the distribution of perceptually true segmentation labels (encoded segmentation variants) with the distribution that is imperceptible at inference time. Adhering to this training objective, the eventual distribution is adjusted to encompass all segmentation variants for a specific input image.

4 Experiment

4.1 Experimental Setup

Dataset. Four datasets are utilized for comparison. The LIDC-IDRI dataset [2] is used for lung lesion segmentation, consisting of lung computed tomography scans from 1010 subjects with annotations from four domain experts. This dataset accurately captures the typical ambiguity found in CT imaging. The BraTS 2017 dataset [18] is used for 3D brain tumor segmentation, comprising 285 cases of 3D MRI images. Each image includes 155 slices and four modes (T1, T1ce, T2, and Flair). These slices

Figure 3: Qualitative comparison with prompted segmentation models adapted for ambiguous segmentation. Examples include three ground-truth expert labels and sampled segmentation masks.

Table 1: Comparison with prompted segmentation models adapted for ambiguous segmentation.

Metric	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑
Method	LIDC				BRATS			
SegGPT w/ Point shift	0.462	0.280	0.573	0.153	0.451	0.032	0.144	0.046
SegGPT w/ Box shift	0.392	0.354	0.638	0.325	0.348	0.082	0.224	0.146
SEEM w/ Mask shift	0.381	0.401	0.692	0.272	0.210	0.228	0.281	0.194
SAM w/ Point shift	0.377	0.365	0.650	0.337	0.252	0.169	0.334	0.238
SAM w/ Box shift	0.361	0.380	0.673	0.253	0.239	0.242	0.344	0.246
$\mathcal{A}$-*SAM* (Ours)	0.228	0.717	0.948	0.356	0.193	0.610	0.864	0.423
Method	ISBI				Sim10K			
SegGPT w/ Point shift	0.649	0.662	0.659	0.323	0.259	0.128	0.151	0.127
SegGPT w/ Box shift	0.527	0.772	0.874	0.536	0.272	0.152	0.220	0.183
SEEM w/ Mask shift	0.522	0.821	0.908	0.760	0.238	0.271	0.344	0.246
SAM w/ Point shift	0.513	0.782	0.886	0.681	0.265	0.155	0.229	0.189
SAM w/ Box shift	0.491	0.792	0.896	0.685	0.255	0.160	0.239	0.199
$\mathcal{A}$-*SAM* (Ours)	0.276	0.835	0.926	0.904	0.233	0.637	0.851	0.327

are annotated into four classes: Background, Non-enhancing/Necrotic Tumor Core, Edema, and Enhancing Tumor Core. The ISBI 2016 dataset [14] contains 900 dermoscopic images for training and 379 images for testing, all annotated by an expert with the lesion area. The images are resized and padded to maintain a uniform scale. The SIM 10k dataset [19] consists of 10,000 images rendered by the gaming engine Grand Theft Auto, providing bounding boxes of 58,701 cars in training images.

Implementation Details. For the LIDC dataset, we use the included four expert annotations to represent different ambiguous segmentation labels. In the case of the BraTS dataset, we amalgamate annotations from different categories into a binary mask, creating multiple segmentation masks to mimic real-world ambiguous segmentation scenarios. For the ISBI dataset, we use the single label provided. All three datasets are optimized using the Adam optimizer, with a learning rate of 1e-4, over 100 epochs. For the SIM 10k dataset, we select images where pixels from two instances overlap, creating three potential masks. Optimization for this dataset is carried out with Adam optimizer over 500 epochs, with a learning rate of 1e-4. The trade-off coefficients are set as $\alpha_P = \alpha_I = 1$.

Evaluation Metrics. Four metrics are used for evaluation: Generalized Energy Distance (GED), Hungarian-Matched Intersection over Union (HM-IoU), Maximum Dice Matching (D_{max}), and Average Dice Matching (D_{mean}). GED is a metric used in ambiguous image segmentation tasks that compares the distribution of segmentations. It leverages the distance between observations, where lower energy signifies a better match between prediction and the ground truth. HM-IoU calculates the optimal match of Intersection over Union (IoU) between annotations and predictions using Hungarian algorithm, providing an accurate representation of sample fidelity. D_{max} and D_{mean} represent the best and average Dice scores between each prediction result and each ground truth, respectively.

4.2 Comparison to Prompted Segmentation Models

Tab. 1 presents the quantitative results on four datasets, offering a comparison with the current state-of-the-art prompting-based segmentation models adapted for ambiguous segmentation tasks. Specifically, we have adapted SegGPT [57], a SAM-like prompt-based segmentation approach that

https://doi.org/10.52202/079017-1449

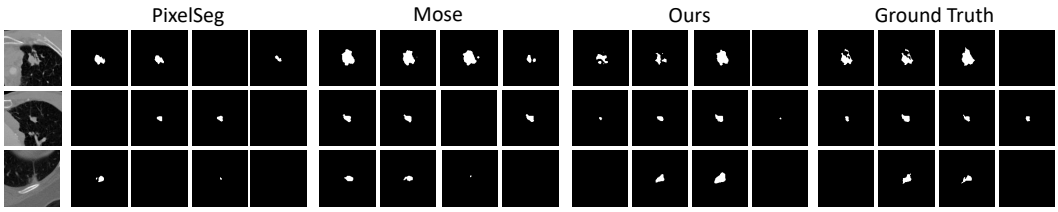

Figure 4: Qualitative comparison with efforts specially designed for ambiguous segmentation. Examples include four ground-truth expert labels and sampled segmentation masks.

Table 2: Quantitative comparison on ambiguous segmentation on the LIDC dataset.

Method	GED ↓	HM-IOU ↑	D_{max} ↑
Prob UNet	0.324	0.423	-
HProb UNet	0.270	0.530	-
PHiseg	0.262	0.595	-
SSN	0.259	0.555	-
CAR	0.252	0.549	0.732
PixelSeg	0.243	0.614	0.814
CIMD	0.234	0.587	-
Mose	0.234	0.623	0.702
$\mathcal{A}$-*SAM* (Ours)	0.230	0.763	0.959

Table 3: Quantitative comparison on ambiguous segmentation on the BraTS dataset.

Method	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑
Prob UNet	0.225	0.521	0.645	0.364
PixelSeg	0.419	0.528	0.785	0.361
$\mathcal{A}$-*SAM* (Ours)	0.192	0.603	0.886	0.438

Table 4: Quantitative comparison on ambiguous segmentation on the ISBI dataset.

Method	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑
UNet	-	0.815	0.902	0.902
Prob UNet	0.329	0.824	0.914	0.894
PHiseg	0.289	0.788	0.912	0.871
cFlow	0.306	0.822	0.918	0.892
$\mathcal{A}$-*SAM* (Ours)	0.267	0.834	0.918	0.905

supports multimodal prompt input [30]. To simulate the actual scenario of prompt-based segmentation, we use the smallest box containing all segmentation labels in each image's mask as the standard prompt. The standard prompt is then randomly perturbed multiple times by scaling [0.8,1.2] and shifting up, down, left, or right by [-8,8] pixels to obtain different ambiguous segmentation results. We have also adapted SEEM [66], a segmentation model that follows the in-context learning paradigm. Given a reference image and a reference segmentation mask, it segments the object in the query image. We select an image with multiple mask labels and apply random perturbations on the reference segmentation mask by shifting [-8,8] pixels left, right, up, or down, resulting in multiple different results on the query image. Comparing with SAM, we employ the same paradigm for prompt acquisition and perturbation as SegGPT. We find that $\mathcal{A}$-*SAM* outperforms the state-of-the-art segmentation foundational models based on point or box in terms of diversity and accuracy. In addition, $\mathcal{A}$-*SAM* surpasses SEEM, a segmentation paradigm directly based on masks, in all aspects. This indicates that our designed strategy accurately captures the ambiguous attributes present in different images and objects, effectively achieving a balance between diversity and accuracy in ambiguous segmentation tasks. Fig. 3 further illustrates the qualitative results of our method in comparison with existing techniques. Our proposed $\mathcal{A}$-*SAM* yields segmentations that preserve a greater degree of accurate object detail, particularly boundary specifics, and offers an exceptional visual representation of potential diversity, as compared to other technologies.

4.3 Comparison to Conventional Ambiguous Segmentation Models

The numerical outcomes across the four datasets are delineated in Tab. 2, 3, 4, and 5, where we draw comparisons with contemporary leading-edge classical ambiguous segmentation methodologies. Precisely, the comparative method results on the LIDC and ISBI datasets are direct quotations from their respective papers, while outcomes on other datasets are predicated on our reimple-

Table 5: Quantitative comparison on ambiguous segmentation on the SIM 10k dataset.

Method	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑
Prob UNet	0.292	0.391	0.462	0.229
PixelSeg	0.398	0.525	0.644	0.358
$\mathcal{A}$-*SAM* (Ours)	0.241	0.596	0.833	0.421

mentation of their official code. The methods compared encompass recent ambiguous segmentation methodologies that amalgamate a conditional variational autoencoder and UNet: Probabilistic U-Net [23], Hierarchical Probabilistic U-Net (HProb UNet) [24], PhiSeg [3], Stochastic Segmentation Networks (SSN) [47], PixelSeg [63], and the ambiguous segmentation endeavor of Calibrated Adversarial Refinement (CAR) [20] and Collective Intelligence Medical Diffusion (CIMD) [50] that integrate generative models. It also includes ensemble-based techniques utilizing a Mixture-of-expert like the Mix of Stochastic Experts (Mose) [13]. We observe that the $\mathcal{A}$-*SAM* transcends state-of-the-art methodologies that amalgamate a conditional variational autoencoder and UNet in terms of diversity and precision. Additionally, the $\mathcal{A}$-*SAM* outstrips segmentation paradigms based on generative models or ensembles in all dimensions. This suggests that our strategy accurately apprehends the ambiguous characteristics inherent in varied images and objects, effectively attaining equilibrium between diversity and precision in the ambiguous segmentation task. Fig. 4 further elucidates the qualitative outcomes of our methodology juxtaposed with extant technologies. In contrast to other technologies, the segmentations engendered by our proposed $\mathcal{A}$-*SAM* preserve a higher degree of exact object detail, particularly boundary details, and provide a distinctive visual representation of potential diversity.

4.4 Further Empirical Results

Table 6: Ablation study on the proposed key components.

No Key Components	GED ↓	HM-IOU ↑	D_{max} ↑	D_{mean} ↑
No Ambiguity Modeling	0.361	0.380	0.673	0.253
No Object Ambiguity	0.370	0.389	0.691	0.230
No Prompt Ambiguity	0.308	0.674	0.930	0.336
No Posterior Distillation	0.266	0.385	0.805	0.341
$\mathcal{A}$-*SAM* (Ours)	0.228	0.717	0.948	0.356

Ablation Study. Tab. 6 delineates the consequences of eliminating various principal strategies of the $\mathcal{A}$-*SAM*. *No Ambiguity Modeling* precludes all ambiguous modeling, including both prompt embedding and image embedding levels. At this stage, we employ the smallest box encompassing all segmentation labels within each image mask as the standard prompt. This standard prompt is then randomly perturbed numerous times by scaling [0.8, 1.2] and shifting up, down, left, or right by [-8, 8] pixels, yielding different ambiguous segmentation outcomes. *No Object Ambiguity* and *No Prompt Ambiguity* either eradicates the ambiguity related to the object or the prompt, that is, it alters its object embedding or prompt embedding to a regular deterministic rather than an ambiguous characteristic. *No Posterior Distillation* eliminates a process of network training guided by a teacher network that can perceive the actual labels. We discover that when any component is excised, the performance correspondingly deteriorates, which underscores the effectiveness of our proposed several strategies.

Robustness Analysis. Fig. 5 reports the IoU performance of $\mathcal{A}$-*SAM* under a variety of instantaneous perturbations. The blue and red solid lines respectively illustrate the performance changes of the SAM model and our method when prompted by different disturbances, while the dashed lines depict the performance of both models under standard prompts, serving as an upper bound. We selected both light and severe degrees of perturbation. Specifically, 'Shift' indicates a random offset of the box by [0,5] pixels, 'Scale' represents a random scaling of the box by [0.85,1.15], 'Shift+' denotes a random offset of the box by [0,8] pixels, and 'Scale+' implies a random scaling of the box by [0.7,1.3]. Compared to the vanilla SAM baseline, $\mathcal{A}$-*SAM* demonstrates robustness against various instantaneous perturbations.

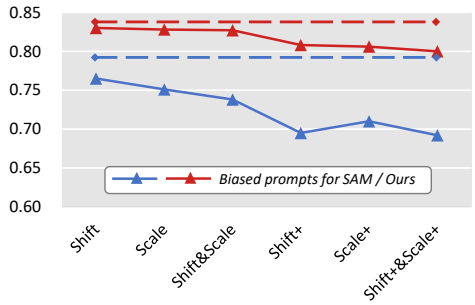

Figure 5: Robustness analysis of our $\mathcal{A}$-*SAM* framework over the SAM baseline against prompt perturbation.

5 Conclusion

The continued evolution of vision foundation models like the Segment Anything Model (SAM) demonstrates impactful universality, while also posing challenges in producing ambiguous and uncertain predictions. Minor changes in the prompt can cause significant variations in the model's output, challenging its required robustness. While many works aim to stabilize SAM's prediction capabilities,

https://doi.org/10.52202/079017-1449

this paper uniquely explores leveraging this perceived flaw to advantageously model inherently ambiguous data distributions. We introduce an innovative optimization framework grounded in a conditional variational autoencoder, which cohesively models the prompt and the object granularity with a latent probability distribution. This approach endows the model with the capacity to adaptively perceive and represent the genuine ambiguous label distribution, thereby enabling SAM to generate a controlled series of diverse, persuasive, and reasonable segmentation outputs. Our comprehensive experiments across multiple practical deployment scenarios involving ambiguity underscore the exceptional performance of our framework, thereby illuminating the need for increased focus on addressing related challenges and opportunities.

Acknowledgement. This work was supported by the Hong Kong Research Grants Council (RGC) General Research Fund under Grant 11211221, 14204321. This work was also supported in part by the National Natural Science Foundation of China under Grant 82172033, Grant 82272071 and in part by the Dreams Foundation of Jianghuai Advance Technology Center, and in part by the Open Fund of the National Key Laboratory of Infrared Detection Technologies.

References

[1] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Sci. Rep.*, 2024.

[2] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

[3] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 119–127. Springer, 2019.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Amanda, S. Agarwal, A. Herbert-Voss, G. Krueger, H. Tom, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, C. Benjamin, J. Clark, C. Berner, M. Sam, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020.

[5] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[6] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3367–3375, 2023.

[7] Z. Chen, W. Li, X. Xing, and Y. Yuan. Medical federated learning with joint graph purification for noisy label learning. *MedIA*, 2023.

[8] Z. Chen, Q. Xu, X. Liu, and Y. Yuan. Un-sam: Universal prompt-free segmentation for generalized nuclei images. *ArXiv*, abs/2402.16663, 2024.

[9] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[10] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023.

[11] Z. Ding, Q. Dong, H. Xu, C. Li, X. Ding, and Y. Huang. Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In *International Conference on Neural Information Processing*, pages 133–144. Springer, 2022.

[12] Y. Gao, W. Xia, D. Hu, and X. Gao. Desam: Decoupling segment anything model for generalizable medical image segmentation. *arXiv preprint arXiv:2306.00499*, 2023.

[13] Z. Gao, Y. Chen, C. Zhang, and X. He. Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic expert. *arXiv preprint arXiv:2212.07328*, 2022.

[14] D. Gutman, N. C. F. Codella, E. M. Celebi, B. Helba, M. Marchetti, N. K. Mishra, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *ArXiv*, abs/1605.01397, 2016.

[15] Z. He, W. Li, Y. Jiang, Z. Peng, P. Wang, X. Li, T. Liu, J. Han, T. Zhang, and Y. Yuan. F2tnet: Fmri to t1w mri knowledge transfer network for brain multi-phenotype prediction. In *MICCAI*, pages 265–275, 2024.

[16] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. IEEE, 2020.

[17] Y. Huang, C. Li, Z. Lin, H. Liu, H. Xu, Y. Liu, Y. Huang, X. Ding, X. Tu, and Y. Yuan. P2sam: Probabilistically prompted sams are efficient segmentator for ambiguous medical images. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9779–9788, 2024.

 https://doi.org/10.52202/079017-1449

[18] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pages 287–297. Springer, 2018.

[19] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016.

[20] E. Kassapis, G. Dikov, D. K. Gupta, and C. Nugteren. Calibrated adversarial refinement for stochastic semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7037–7047, 2020.

[21] E. Kassapis, G. Dikov, D. K. Gupta, and C. Nugteren. Calibrated adversarial refinement for stochastic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7057–7067, 2021.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[23] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

[24] S. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. J. Rezende, S. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.

[25] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[26] C. Li, B. Y. Feng, Y. Liu, H. Liu, C. Wang, W. Yu, and Y. Yuan. Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting. *arXiv preprint arXiv:2407.01029*, 2024.

[27] C. Li, X. Lin, Y. Mao, W. Lin, Q. Qi, X. Ding, Y. Huang, D. Liang, and Y. Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141:105144, 2022.

[28] C. Li, H. Liu, Y. Liu, B. Y. Feng, W. Li, X. Liu, Z. Chen, J. Shao, and Y. Yuan. Endora: Video generation models as endoscopy simulators. *arXiv preprint arXiv:2403.11050*, 2024.

[29] C. Li, X. Liu, W. Li, C. Wang, H. Liu, and Y. Yuan. U-kan makes strong backbone for medical image segmentation and generation. *arXiv preprint arXiv:2406.02918*, 2024.

[30] C. Li, X. Liu, C. Wang, Y. Liu, W. Yu, J. Shao, and Y. Yuan. Gtp-4o: Modality-prompted heterogeneous graph learning for omni-modal biomedical representation. *arXiv preprint arXiv:2407.05540*, 2024.

[31] C. Li, W. Ma, L. Sun, X. Ding, Y. Huang, G. Wang, and Y. Yu. Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. *Neural Computing and Applications*, pages 1–14.

[32] C. Li, Y. Zhang, J. Li, Y. Huang, and X. Ding. Unsupervised anomaly segmentation using image-semantic cycle translation. *arXiv preprint arXiv:2103.09094*, 2021.

[33] C. Li, Y. Zhang, Z. Liang, W. Ma, Y. Huang, and X. Ding. Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 61–65. IEEE, 2021.

[34] W. Li, X. Guo, and Y. Yuan. Novel scenes & classes: Towards adaptive open-set object detection. In *ICCV*, pages 15780–15790, 2023.

[35] W. Li, X. Liu, Q. Yang, and Y. Yuan. From static to dynamic diagnostics: Boosting medical image analysis via motion-informed generative videos. In *MICCAI*, 2024.

[36] W. Li, X. Liu, X. Yao, and Y. Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *AAAI*, pages 1421–1428, 2022.

[37] W. Li, X. Liu, and Y. Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, 2022.

[38] H. Liu, Y. Liu, C. Li, W. Li, and Y. Yuan. Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. *arXiv preprint arXiv:2406.16073*, 2024.

[39] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[40] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[41] X. Liu, W. Li, and Y. Yuan. Intervention & interaction federated abnormality detection with noisy clients. In *MICCAI*, pages 309–319, 2022.

[42] X. Liu, W. Li, and Y. Yuan. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–66. Springer, 2024.

[43] Y. Liu, C. Li, C. Yang, and Y. Yuan. Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. *arXiv preprint arXiv:2401.12561*, 2024.

[44] Y. Liu, W. Li, C. Wang, H. Chen, and Y. Yuan. When 3d partial points meets sam: Tooth point cloud segmentation with sparse labels. In *MICCAI*, pages 778–788, 2024.

[45] Z. Lou, Q. Xu, Z. Jiang, X. He, Z. Chen, Y. Wang, C. Li, M. M. He, and W. Duan. Nusegdg: Integration of heterogeneous space and gaussian kernel for domain-generalized nuclei segmentation. *arXiv preprint arXiv:2408.11787*, 2024.

[46] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[47] M. Monteiro, L. Le Folgoc, D. Coelho de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020.

[48] L. P. Osco, Q. Wu, E. L. de Lemos, W. N. Gonçalves, A. P. M. Ramos, J. Li, and J. M. Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *ArXiv*, abs/2306.16623, 2023.

[49] D. Qiu and L. M. Lui. Modal uncertainty estimation via discrete latent representation. *arXiv preprint arXiv:2007.12858*, 2020.

[50] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11546, 2023.

[51] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[52] R. Selvan, F. Faye, J. Middleton, and A. Pai. Uncertainty quantification in medical image segmentation with normalizing flows. In *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*, pages 80–90. Springer, 2020.

[53] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[54] L. Sun, C. Li, X. Ding, Y. Huang, Z. Chen, G. Wang, Y. Yu, and J. Paisley. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140:105067, 2022.

[55] M. A. Valiuddin, C. G. Viviers, R. J. van Sloun, P. H. de With, and F. van der Sommen. Improving aleatoric uncertainty quantification in multi-annotated medical image segmentation with normalizing flows. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis: 3rd International Workshop, UNSURE 2021, and 6th International Workshop, PIPPI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 3*, pages 75–88. Springer, 2021.

 https://doi.org/10.52202/079017-1449

[56] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang. Images speak in images: A generalist painter for in-context visual learning. Dec 2022.

[57] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang. Seggpt: Segmenting everything in context. *ArXiv*, abs/2304.03284, 2023.

[58] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

[59] J. M. Y. Y. Wuyang Li, Xinyu Liu. Cliff: Continual latent diffusion for open-vocabulary object detection. In *ECCV*, 2024.

[60] H. Xu, Y. Zhang, L. Sun, C. Li, Y. Huang, and X. Ding. Afsc: Adaptive fourier space compression for anomaly detection. *arXiv preprint arXiv:2204.07963*, 2022.

[61] Y. Yuan, M. Chao, and Y.-C. Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36:1876–1886, 2017.

[62] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention.

[63] W. Zhang, X. Zhang, S. Huang, Y. Lu, and K. Wang. Pixelseg: Pixel-by-pixel stochastic semantic segmentation for ambiguous medical images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4742–4750, 2022.

[64] W. Zhang, X. Zhang, S. Huang, Y. Lu, and K. Wang. A probabilistic model for controlling diversity and accuracy of ambiguous medical image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4751–4759, 2022.

[65] Y. Zhang, C. Li, X. Lin, L. Sun, Y. Zhuang, Y. Huang, X. Ding, X. Liu, and Y. Yu. Generator versus segmentor: Pseudo-healthy synthesis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 150–160. Springer, 2021.

[66] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix

The following contents are provided in the supplements:

- Limitation and Border Impact.
- More experimental details (Sec. 4.1 in main paper).
- Details of network architecture of *$\mathcal{A}$-SAM*.
- More visualization about our experiments. (Sec. 4.2 and Sec. 4.3 in the main paper).

A.1 Limitation and Border Impact

Limitation. The current approach is constrained by the uncontrolled and unquantifiable nature of uncertainty. This limitation means that the accuracy of handling uncertainty varies across different scenarios. Further systematic analysis is required to comprehend the underlying factors that result in some scenarios being more manageable than others in terms of uncertainty.

Broader Impacts. Ambiguous segmentation and uncertainty handling are essential in fields such as image processing and medical diagnostics. Fuzzy segmentation improves our understanding of complex or unclear image content. Proper uncertainty management can enhance prediction accuracy and decision-making, especially in medical diagnostics, aiding in precise disease diagnosis and treatment planning. Ensuring image processing safety is also critical for user privacy and data security, thereby building trust and satisfaction.

A.2 Detailed Experimental Setup

Dataset. Four datasets are used for comparison. **LIDC-IDRI** [2] is a dataset for lung lesion segmentation, which encompasses a voluminous collection of lung computed tomography scans from 1010 distinct subjects, with manual annotations provided by a panel composed of four domain experts. A diversified panel of 12 radiologists leveraged their expertise to provide annotation masks for the dataset, a characteristic that allows the dataset to reflect the typical ambiguity frequently encountered in CT imaging, thereby ensuring comprehensive, accurate annotations that represent a broad range of expert opinions. The resolution of all images is 128×128. **BraTS 2017** [18] is a dataset for 3D brain tumor segmentation, which consists of 285 cases of 3D MRI images, each image comprising 155 slices. Each slice exhibits four modes (T1, T1ce, T2, and Flair) and is meticulously annotated by professional radiologists into four classes: Background (BG), Non-enhancing/Necrotic Tumor Core (NET), Edema (OD), and Enhancing Tumor Core (ET). The resolution of all images is 240×240. **ISBI 2016** [14] is a dermoscopy dataset containing 900 dermoscopic images for training and 379 images for testing. Each image is 8-bit RGB and is annotated by an expert with the lesion area. To keep each image at the same scale, we follow [61] to resize the images to 256×192 and pad the top and bottom with 32 pixels, respectively, to get 256×256 images. **SIM 10k** [19] consists of 10,000 images that are rendered by the gaming engine Grand Theft Auto. In SIM 10k, bounding boxes of 58,701 cars are provided in the 10,000 training images. All images are used in the training.

Implementation Details. For LIDC, we directly employ the four expert annotations included in the dataset to represent four different ambiguous segmentation labels [2]. For BraTS, we overlay and amalgamate annotations from disparate categories, subsequently transforming the outcome into a binary mask that comprises solely the foreground and background [18]. This process is geared towards generating multiple segmentation masks to mimic real-world ambiguous segmentation scenarios, thereby augmenting the rigor and reliability of the experiment. For ISBI, we directly use the single label included in the dataset [14]. For the aforementioned three datasets, we carry out optimization using the Adam optimizer, with a learning rate of 1e-4, over 100 epochs. For the SIM 10k dataset, we contemplate a practical overlap setting. Specifically, we select images in which pixels from two instances within a frame overlap, thereafter creating three potential masks from the two overlapping instances. These masks could represent the first object, the second object, or the union of both. For SIM 10k [19], we carry out optimization using the Adam optimizer with a learning rate of 1e-4, over 500 epochs. The trade-off coefficients in Eq. 14 are set as $\alpha_P = \alpha_I = 1$.

Evaluation Metrics. Four metrics are used for strict evaluation. **Generalized Energy Distance (GED)** is a commonly used metric in ambiguous image segmentation tasks that leverages distance between observations by comparing the distribution of segmentations [23], as $D^2_{GED}(P_{gt}, P_{out}) = 2\mathbb{E}[d(S,Y)] - \mathbb{E}[d(S,S')] - \mathbb{E}[d(Y,Y')]$, where d corresponds to the distance measure $d(x,y) = 1 - IoU(x,y)$, Y and Y' are independent samples of P_{gt} and S and S' are sampled from P_{out}. Lower energy indicates better agreement between prediction and the ground truth distribution of segmentations. **Hungarian-Matched Intersection over Union (HM-IoU)** is used by calculating the optimal match of Intersection over Union (IoU) between annotations and predictions, which is searched by Hungarian algorithm. This metric offers a more accurate representation of

https://doi.org/10.52202/079017-1449

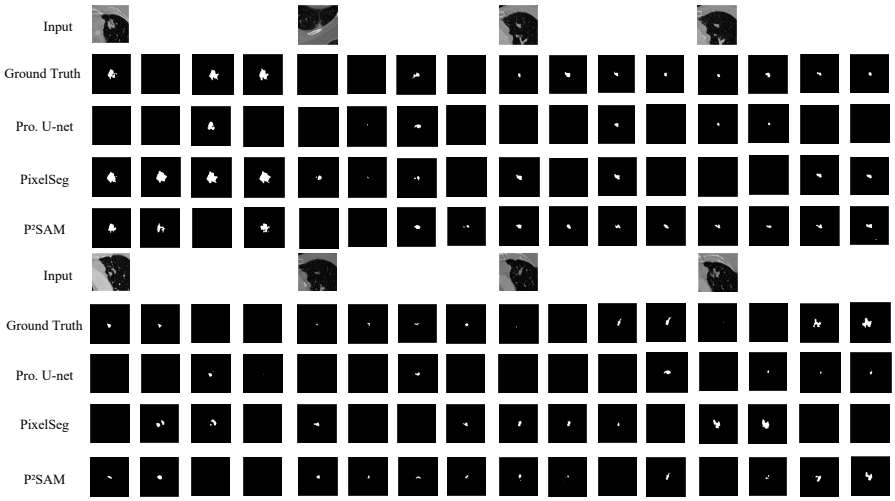

Figure 6: More visualization on the LIDC dataset, displaying only the first 4 samples.

sample fidelity, contrasting the Generalized Energy Distance (GED) which tends to over-reward sample diversity. The Hungarian algorithm identifies the best one-to-one correspondence between objects in two sets. In this context, we utilize $IoU(Y, Y')$ to determine the similarity between two samples. **Maximum & Average Dice Matching (**D_{max} **&** D_{mean}**)** is respectively the best and average results over the Dice scores between each prediction result and each ground truth.

A.3 Network Architecture

A.3.1 Prompted SAM

In the experiment, we adopted the Vit-b version of the SAM model and accommodated Enc_I by reducing the size of the output feature map F_I by $\frac{1}{8}$ compared with the original. This change is expected to reduce the required memory usage during the training process and accelerate the inference speed of the model. In addition, we adjusted the SAM model to multi output mode with 8 outputs, and set the pixel mean and pixel std parameters to 0 and 1, respectively.

A.3.2 Prompt Generation Network (PGN)

The network mainly consists of two parts. (1) Encoder: This part contains 4 convolutional blocks, each with 3 convolutional layers inside. These 4 convolutional blocks have channel numbers of 32, 64, 128 and 192, respectively, to gradually extract and deepen features. (2) Axis Gaussian Generation Network: This network consists of a 1x1 convolutional layer with 256 channels and an axial Gaussian distribution generator. This design first increases the dimensionality of the feature map output by the Encoder through a 1x1 convolutional layer to obtain 256 dimensional μ and σ, and then these two parameters are fed into a Gaussian generator to generate the distribution of $\tilde{T_P}$.

A.3.3 Diversity-aware Assembling Module

In our experimental design, we set the number of mask weights $\mathcal{W}$ to 8 and initialize each weight to $\frac{1}{8}$. This setting aims to correspond to 8 outputs of SAM model. In the first stage of the experiment, these weight $\mathcal{W}$ will be trained to meet the model requirements. In the second stage, we will freeze these weights. This is to enable the prompt generation network to generate more diverse and representative segmentation results, thereby effectively guiding the modeling of $\tilde{T_P}$.

A.4 More Visualization about Experiments

As demonstrated in Fig. 6 and Fig. 7, these illustrations provide an extensive visualization of our research outcomes. These figures meticulously depict various aspects of our data, aiding readers in gaining a profound understanding of our research findings.

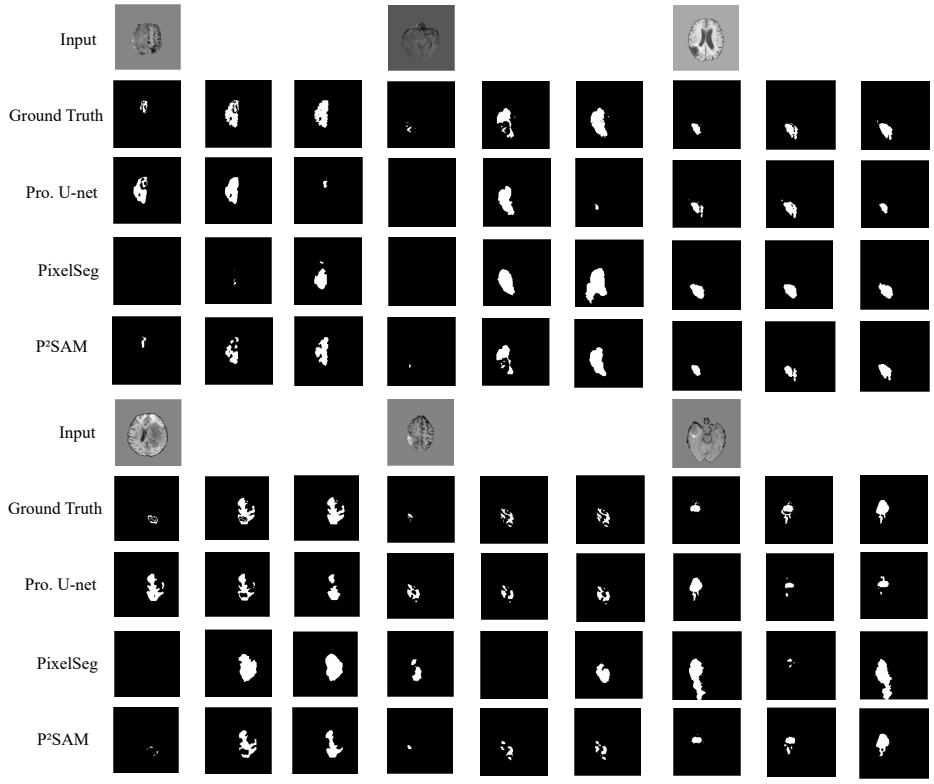

Figure 7: More visualization on the BraTS2017 dataset, displaying only the first 4 samples.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction sections offer a comprehensive discussion of the manuscript's context, intuition, and ambitions, as well as its contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed by authors at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

https://doi.org/10.52202/079017-1449

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The pipeline of the methods and the details of experiments are presented with corresponding reproducible credentials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All utilized data are sourced from open-access platforms. The code, which will be made publicly available, is uploaded as a zip file.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The pipeline of the methods and the details of experiments are presented with corresponding reproducible credentials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results contain the standard deviation of the results over several random runs.

https://doi.org/10.52202/079017-1449

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details of experiments are presented with corresponding reproducible credentials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original owners of assets, including data and models, used in the paper, are properly credited and are the license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

https://doi.org/10.52202/079017-1449

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced in the paper are well documented and provided alongside the assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.