NeurIPS 2023

Dataset Diffusion: Diffusion-based Synthetic Dataset Generation for Pixel-Level Semantic Segmentation


Quang Nguyen


Truong Vu


Anh Tran


Khoi Nguyen

Motivations

Human annotations are expensive

The process of manually annotating can be time-consuming and expensive.

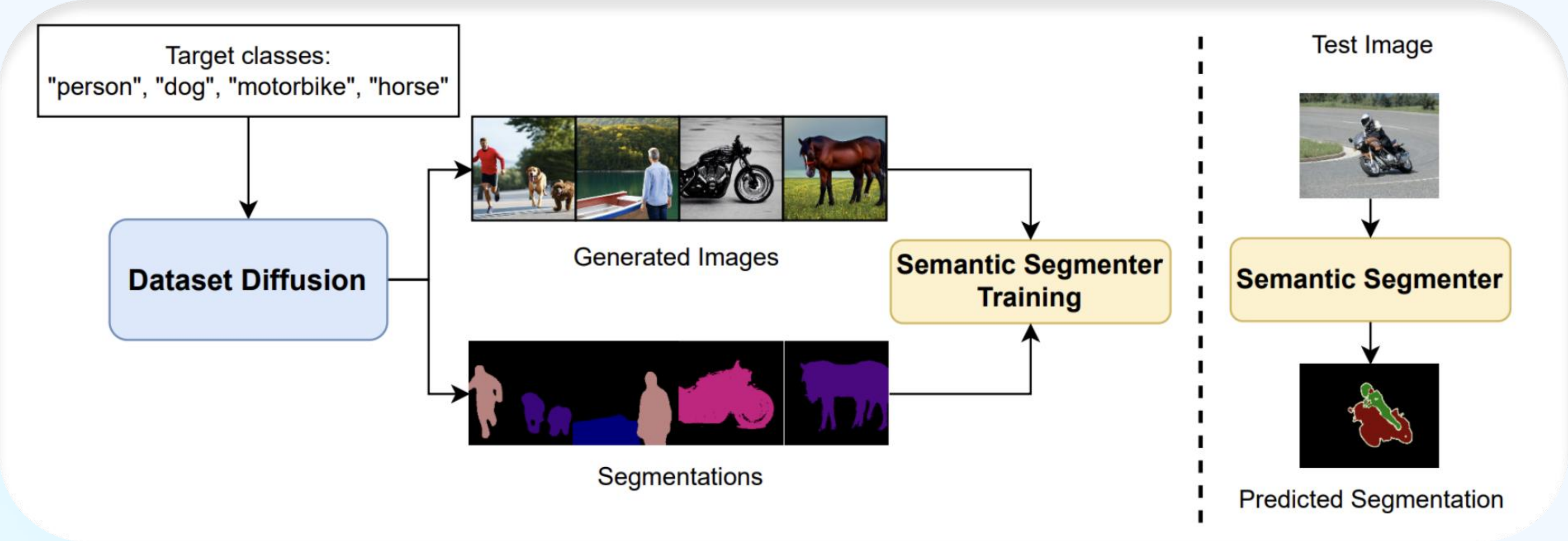
Leveraging pretrained diffusion models

Pretrained text-to-image Stable Diffusion model is pretrained on billions of images and ready to use for generating segmentation along with images

Explaining the results of diffusion models

Use segmentation of each selected words to verify whether the generated images are faithful to the given text prompts

Problem Statement

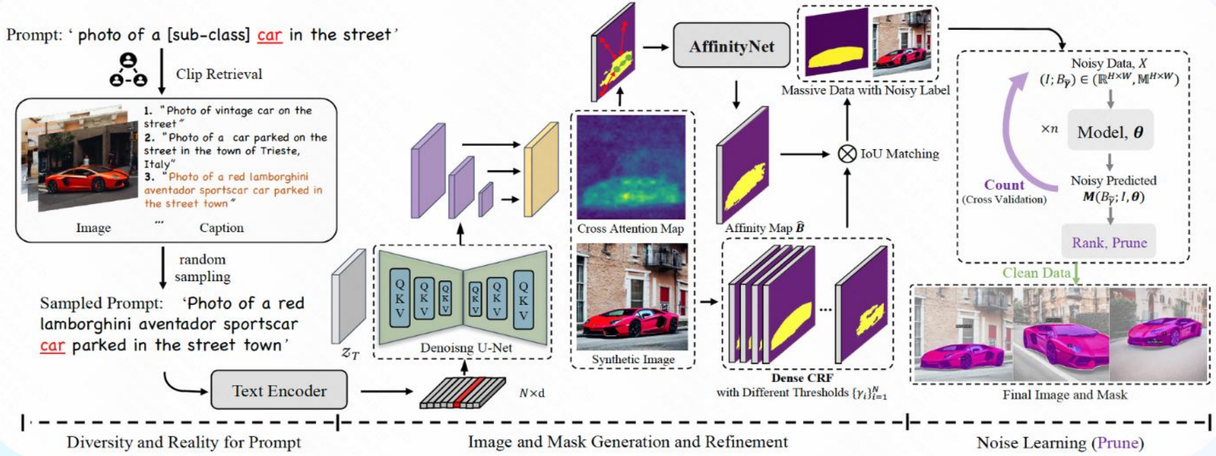


Challenges:

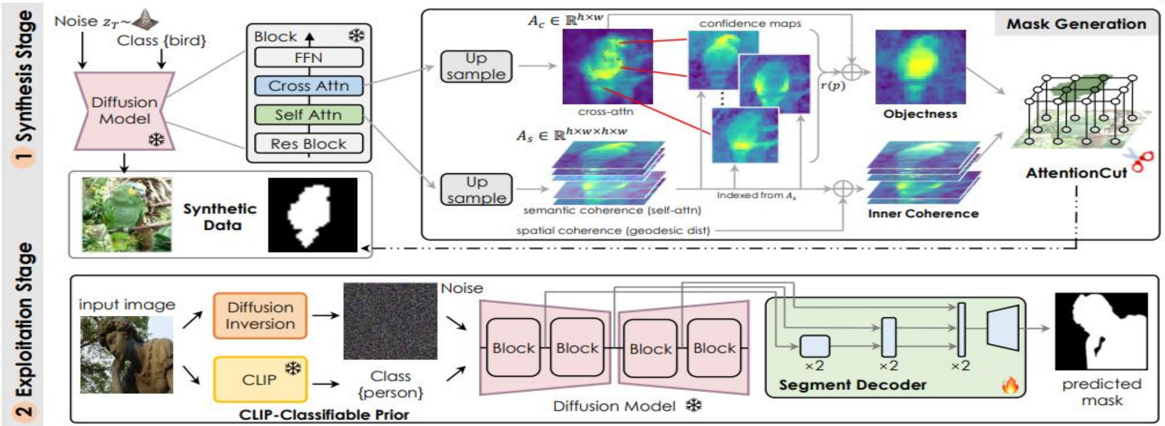
- Stable Diffusion is not trained on pixel-level annotation
- Stable Diffusion is only designed for generating images!

Prior Work

DiffuMask



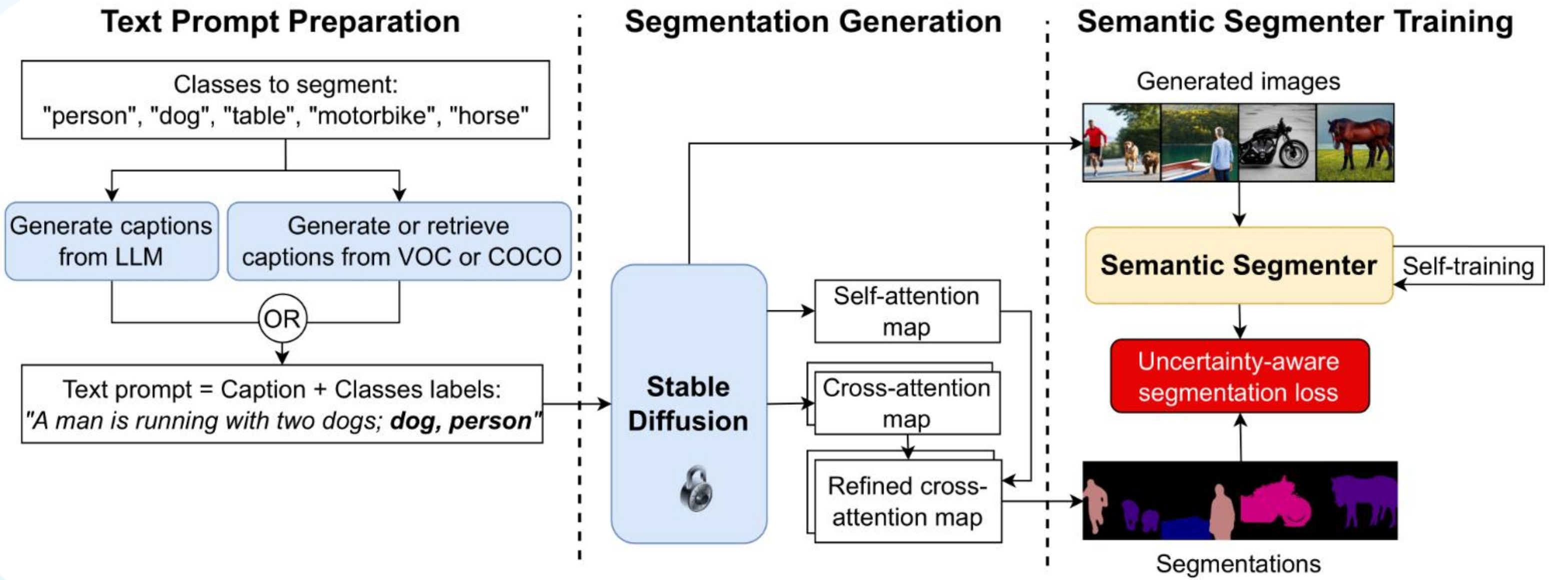
DiffusionSeg



Limitations:

- They can only generate single object segmentation mask per image.
- Requires complex post-processing step to obtain final segmentation mask: DenseCRF and GraphCut

Overview of our Approach



Our contributions:

- Introduce simple but effective text prompts design for generating more objects
- Employ self and cross-attention maps to produce segmentation maps

Class-prompt Appending

Several problems with the generated (provided) captions: missing classes or mismatched classes

Caption: A photograph of a kitchen inside a house.
Provided classes: **bottle**, **microwave**, **sink**, **refrigerator**

Caption: A **bike** leaning against a sign in Scotland.
Provided classes: **bicycle**, **backpack**, **bottle**

Caption: A **man** riding a dirt **bike** in a forest.
Provided classes: **person**, **motorcycle**

"A photo graph of a kitchen inside a house; bottle microwave sink refrigerator"

"A bike leaning against a sign in Scotland; a bicycle backpack bottle"

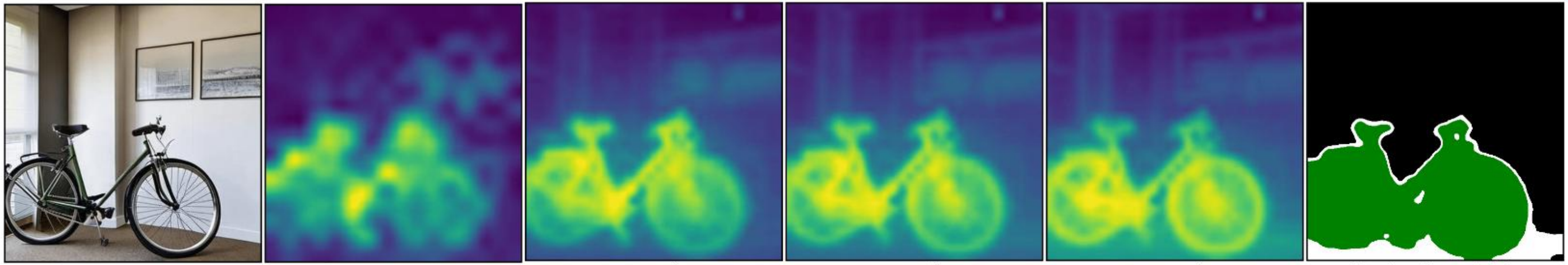
"A man riding a dirt bike in a forest; person motorcycle"

Stable Diffusion

Generating Segmentation Map from Self and Cross-attention Maps

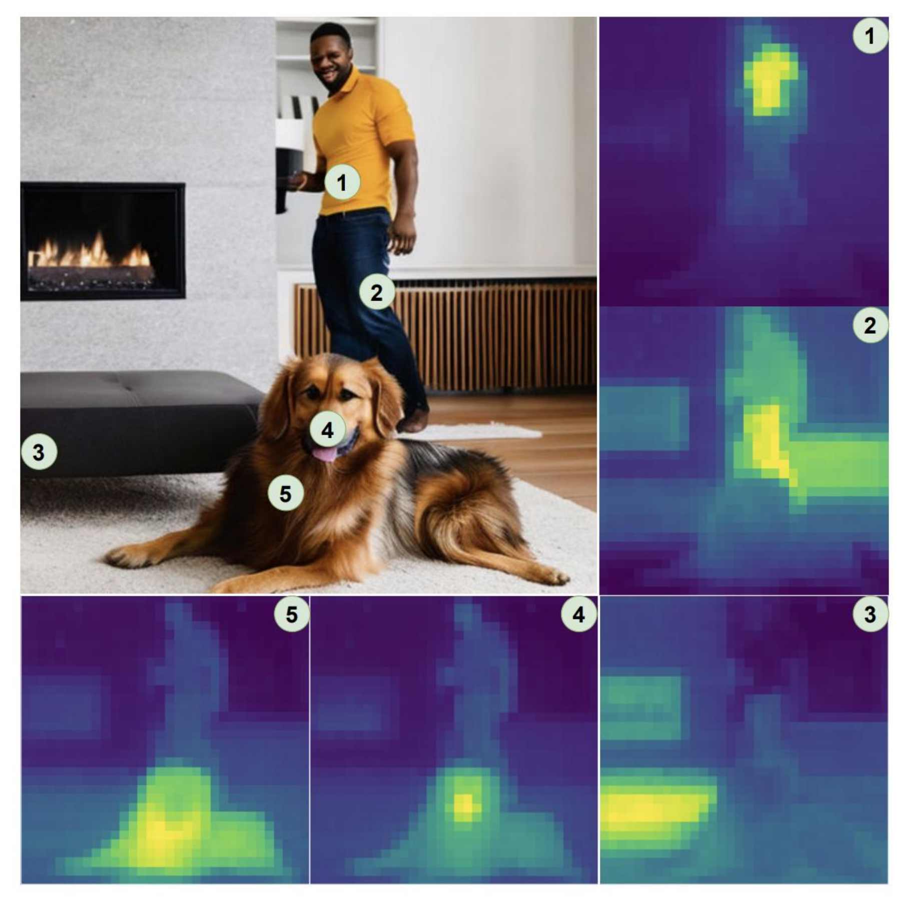
Refine the cross-attention maps A_c by using the self-attention A_S exponentiation

$$\mathcal{A}_C^* = (\mathcal{A}_S)^\tau \cdot \mathcal{A}_C$$



Generated image | Cross-attention | $\tau = 1$ | $\tau = 2$ | $\tau = 4$ | Mask + Uncertainty

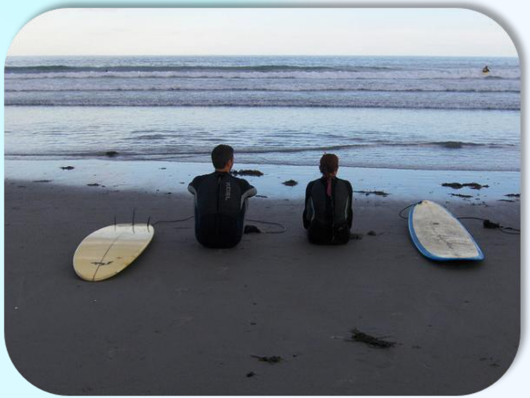
Why Self-attention help improve cross-attention?



Experiments

- **Datasets:**
  - Training: introduce new **synth-VOC** and **synth-COCO** benchmarks which only contain text prompts taken from the provided/generated captions of VOC and COCO
  - Testing: the test set of
    - **PASCAL-VOC 2012:** 20 object classes and 1,456 test images.
    - **COCO 2017:** 80 object classes and 5K validataion images.
- **Metric:** mIoU

"a photo of person"



"two people sit in the sand with their surf boards."

"a man riding a horse"

Quantitative results

Segmenter	Backbone	VOC dataset: Training set	VOC dataset: Val	VOC dataset: Test	COCO dataset: Training set	COCO dataset: Val
DeepLabV3	ResNet50	VOC's training (11.5k images)	77.4	75.2	COCO's training (2017: 118k images)	48.9
DeepLabV3	ResNet101		79.9 ↓	79.8 ↓		54.9
Mask2Former	ResNet50		77.3	77.2		57.8
Mask2Former	ResNet50	DiffuMask [8] (60k images)	57.4 ↑	-	-	-
DeepLabV3	ResNet50	Dataset Diffusion (40k images)	61.6	59.0	Dataset Diffusion (80k images)	32.4
DeepLabV3	ResNet101		64.8	64.6		34.2
Mask2Former	ResNet50		60.2	60.5		31.0

On VOC, our approach yields satisfactory results of 64.8 mIoU when compared to the real training set of 79.9 mIoU.

Ours outperforms DiffuMask by a large margin of 4.2 mIoU using the same Resnet50 backbone.

Quantitative results

Segmenter	Backbone	VOC dataset			COCO dataset	
		Training set	Val	Test	Training set	Val
DeepLabV3	ResNet50	VOC's training (11.5k images)	77.4	75.2	COCO's training (2017: 118k images)	48.9
DeepLabV3	ResNet101		79.9	79.8		54.9
Mask2Former	ResNet50		77.3	77.2		57.8
Mask2Former	ResNet50	DiffuMask [8] (60k images)	57.4	-	-	-
DeepLabV3	ResNet50	Dataset Diffusion (40k images)	61.6	59.0	Dataset Diffusion (80k images)	32.4
DeepLabV3	ResNet101		64.8	64.6		34.2
Mask2Former	ResNet50		60.2	60.5		31.0

Dataset Diffusion achieves a promising result of 34.2 mIoU compared to 54.9 mIoU of real COCO training set.

Ablation Study

Method	Example	mIoU (%)
1: Simple text prompts	a photo of an aeroplane	54.7
2: Captions only	a large white airplane sitting on top of a boat	50.8
3: Class labels only	aeroplane boat	57.4
4: Simple text prompts + class labels	a photo of an aeroplane; aeroplane boat	57.6
5: Caption + class labels	a large white plane sitting on top of a boat; aeroplane boat	**62.0**

Study on different text prompt selection

Cross-attention	Self-attention	Uncertainty	Self-training	TTA	mIoU (%)
✓					44.8
✓	✓				61.0
✓	✓	✓			62.0
✓	✓	✓	✓		62.7
✓	✓	✓	✓	✓	**64.3**

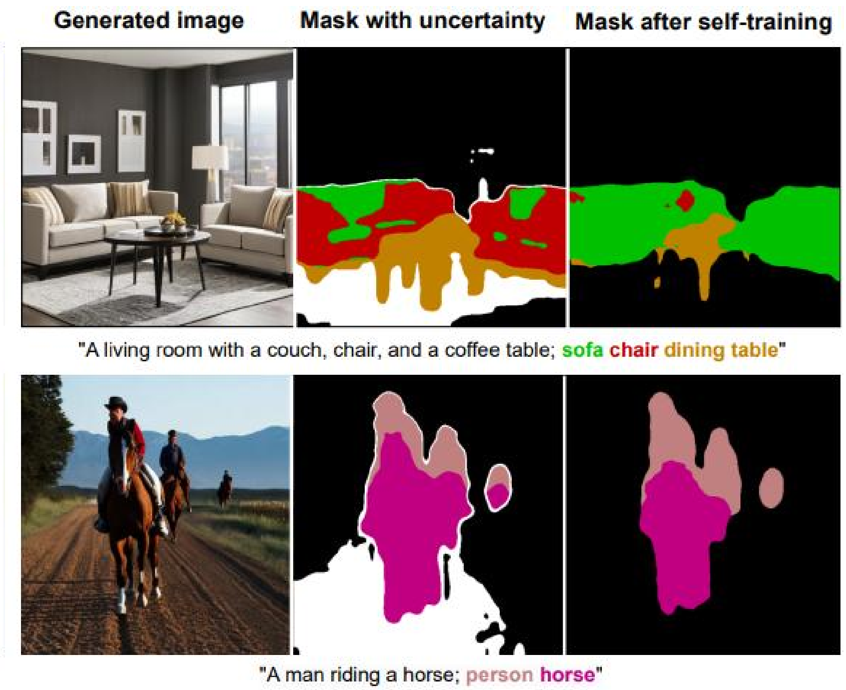
Study on different components

	Self-attention	
Cross-attention	32	64
8	39.7	38.1
16	**62.0**	59.6
32	52.8	50.9
64	35.4	31.5
16, 32	59.7	57.3
16, 32, 64	59.1	57.2

Study on different feature scales

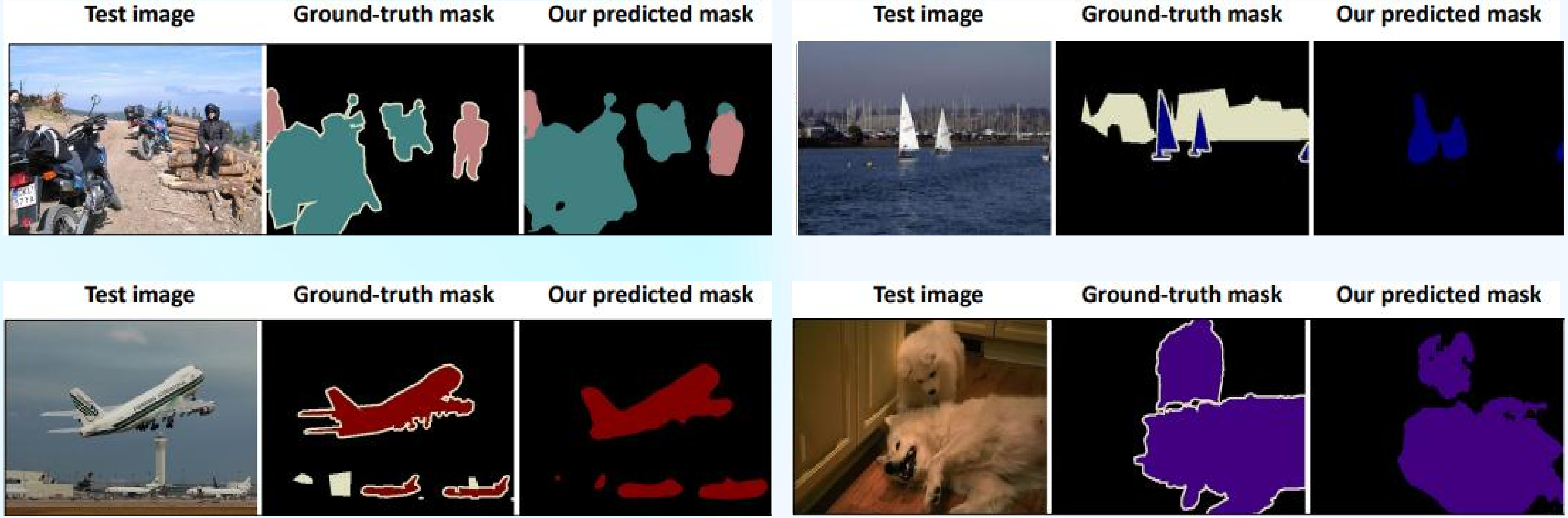
Our Generated Images and Segmentations

Dataset Diffusion can generate the high quality semantic masks.



Selftraining helps correct mis-segmented objects in some cases but can harm the original mask for small objects.

Segmentation results on VOC val set



Thank you!

@VinAI

https://www.vinai.io/