Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

Jiarui Xu Sifei Liu* Arash Vahdat* Wonmin Byeon
Xiaolong Wang Shalini De Mello

STRUCT Group Seminar
Presenter: Haowei Kuang
2023.05.07

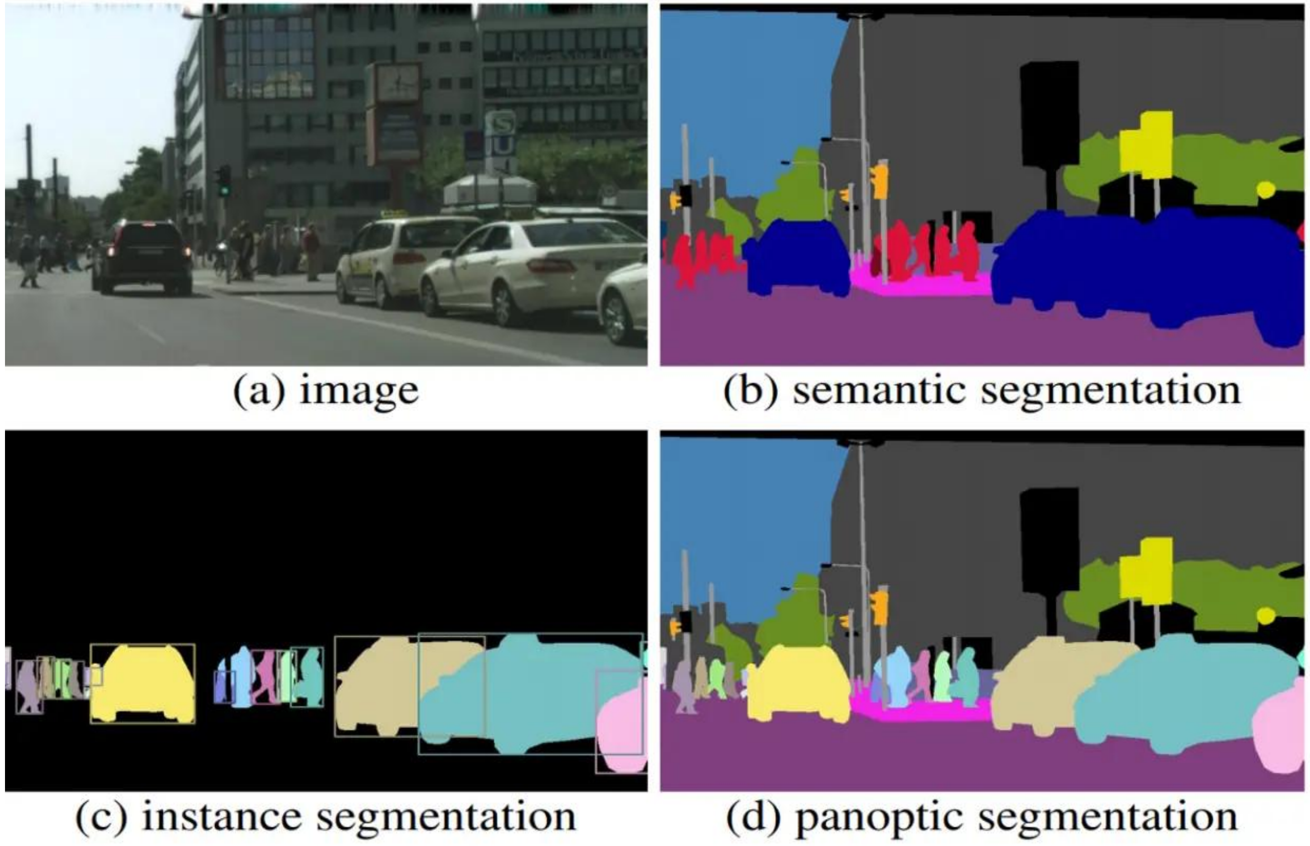
CVPR '2023 Highlight

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

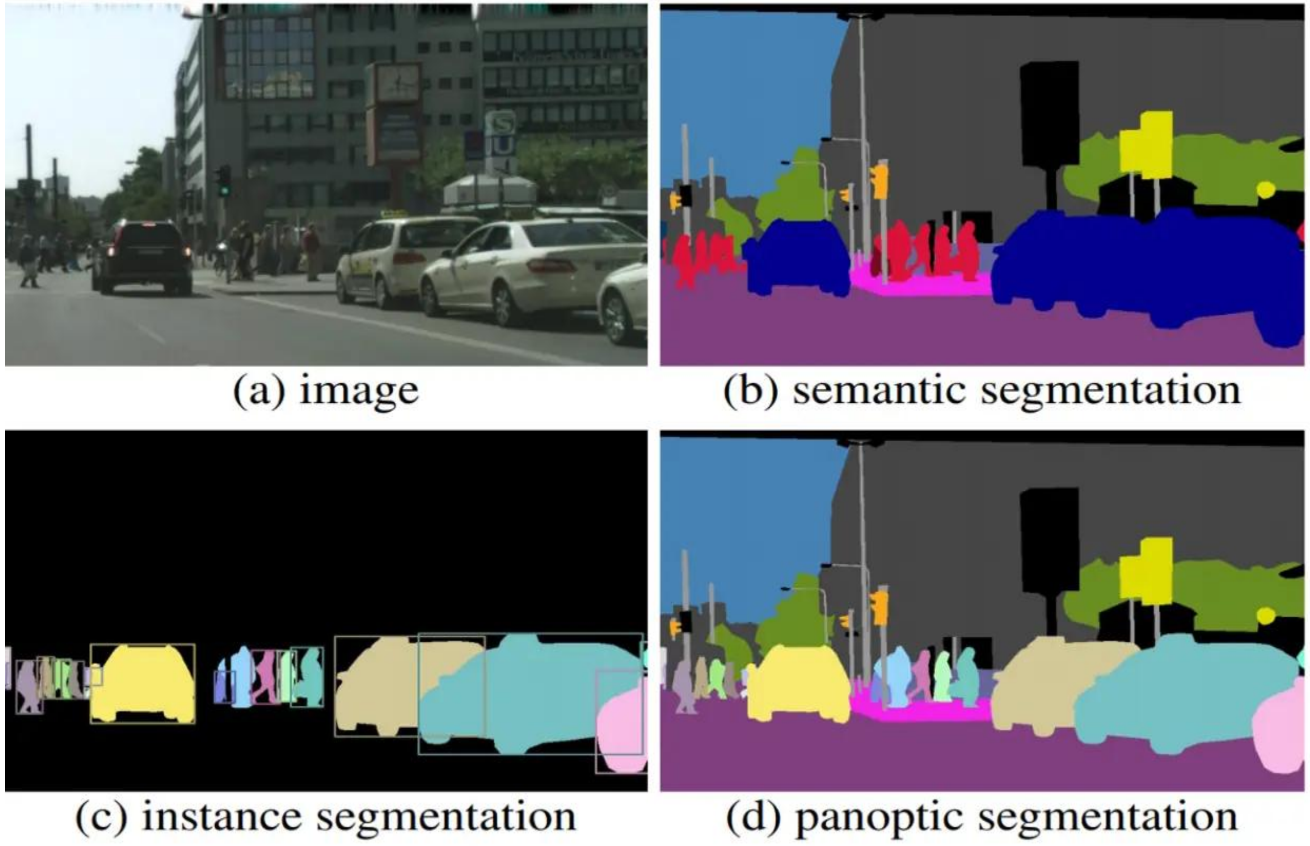
BACKGROUND: Open-Vocabulary Panoptic Segmentation

Panoptic Segmentation



BACKGROUND: Open-Vocabulary Panoptic Segmentation

Open-Vocabulary Segmentation

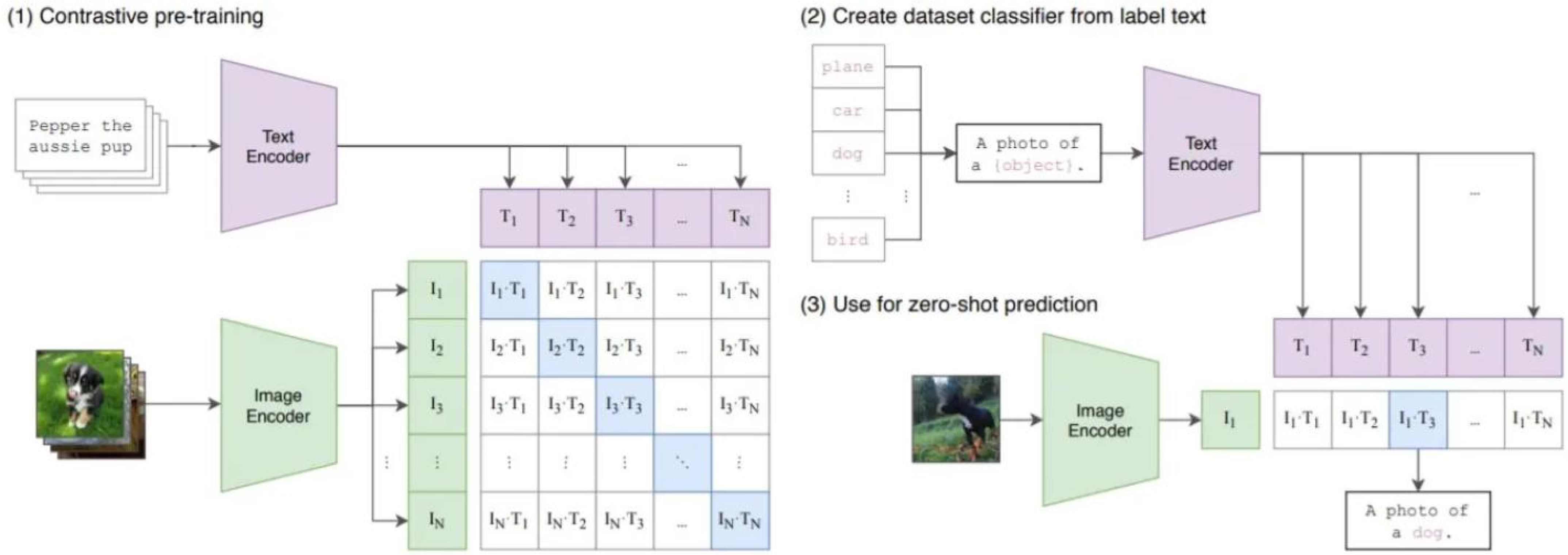


BACKGROUND: Open-Vocabulary Panoptic Segmentation

Open-Vocabulary Panoptic Segmentation

BACKGROUND: CLIP

CLIP

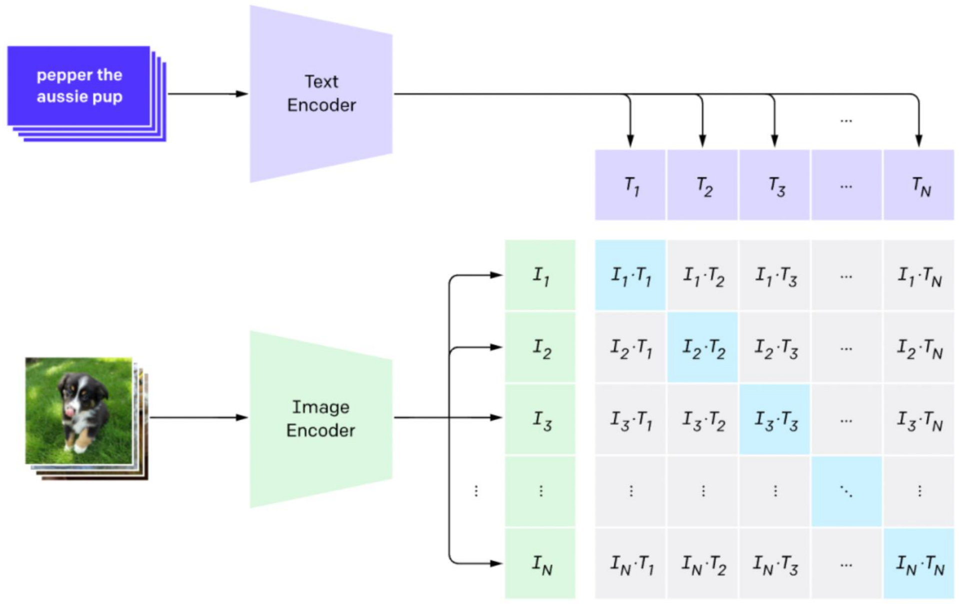


BACKGROUND: CLIP

CLIP

- Training Data: 400 Million Image-Text Pairs
- Deficiency:

Confuses the spatial relations between objects



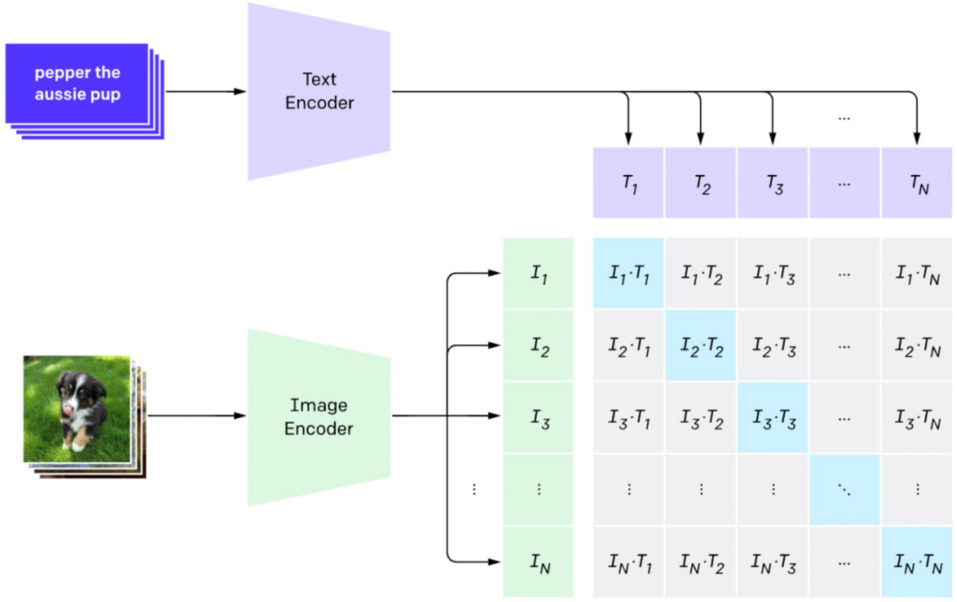
BACKGROUND: CLIP

CLIP

- Training Data: 400 Million Image-Text Pairs
- Deficiency:

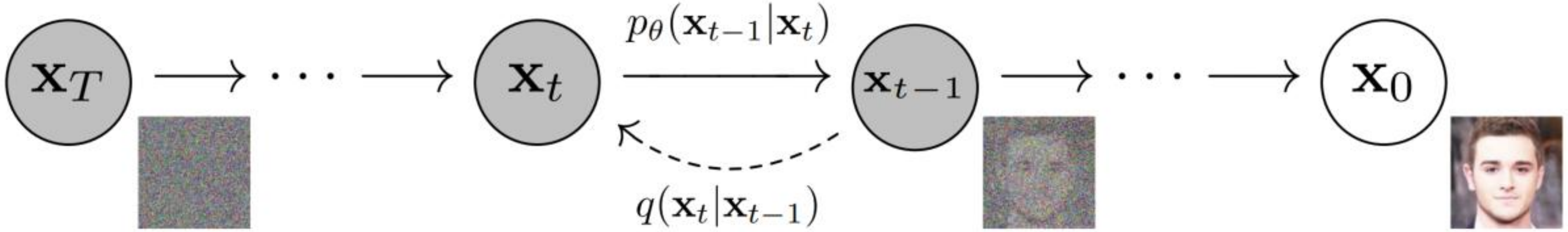
Confuses the spatial relations between objects

Bottleneck for Open-Vocabulary Panoptic Segmentation



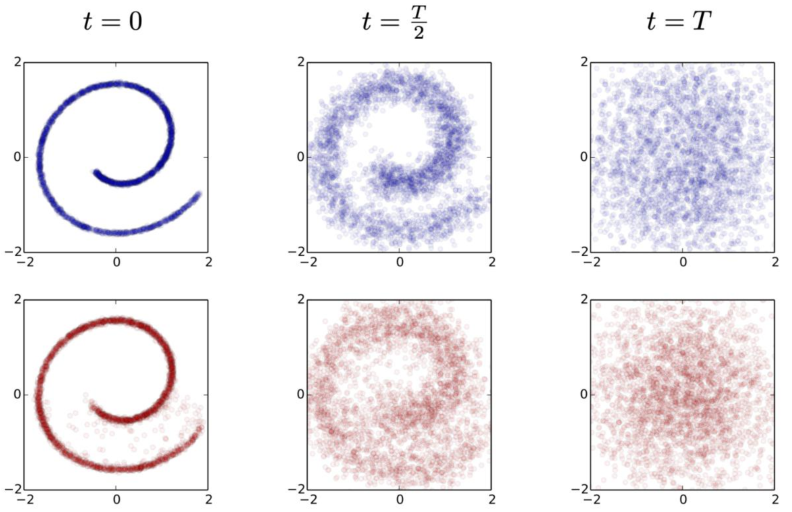
BACKGROUND: Diffusion

Diffusion



The forward trajectory
$q(\mathbf{x}_{0:T})$

The reverse trajectory
$p_\theta(\mathbf{x}_{0:T})$

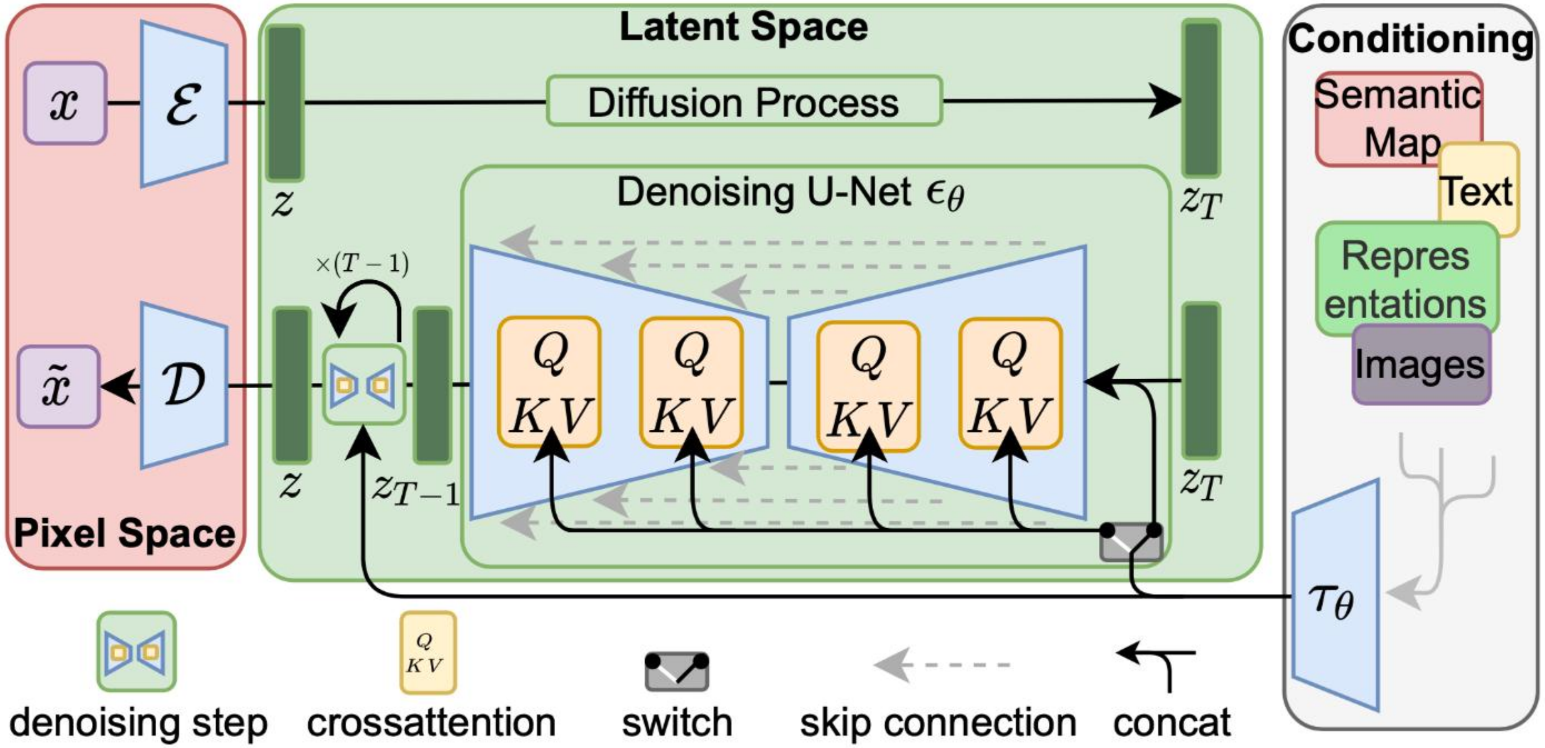


$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

BACKGROUND: Diffusion

Text-to-Image Diffusion Model

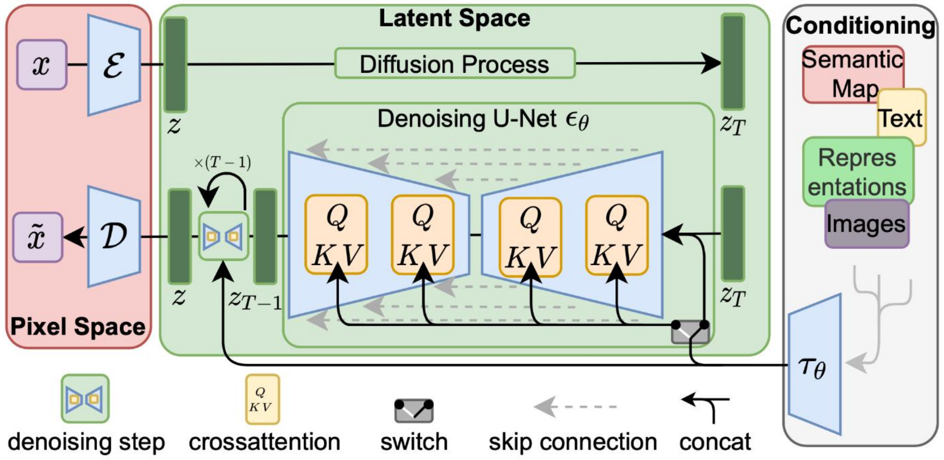
- Stable Diffusion



BACKGROUND: Diffusion

Text-to-Image Diffusion Model

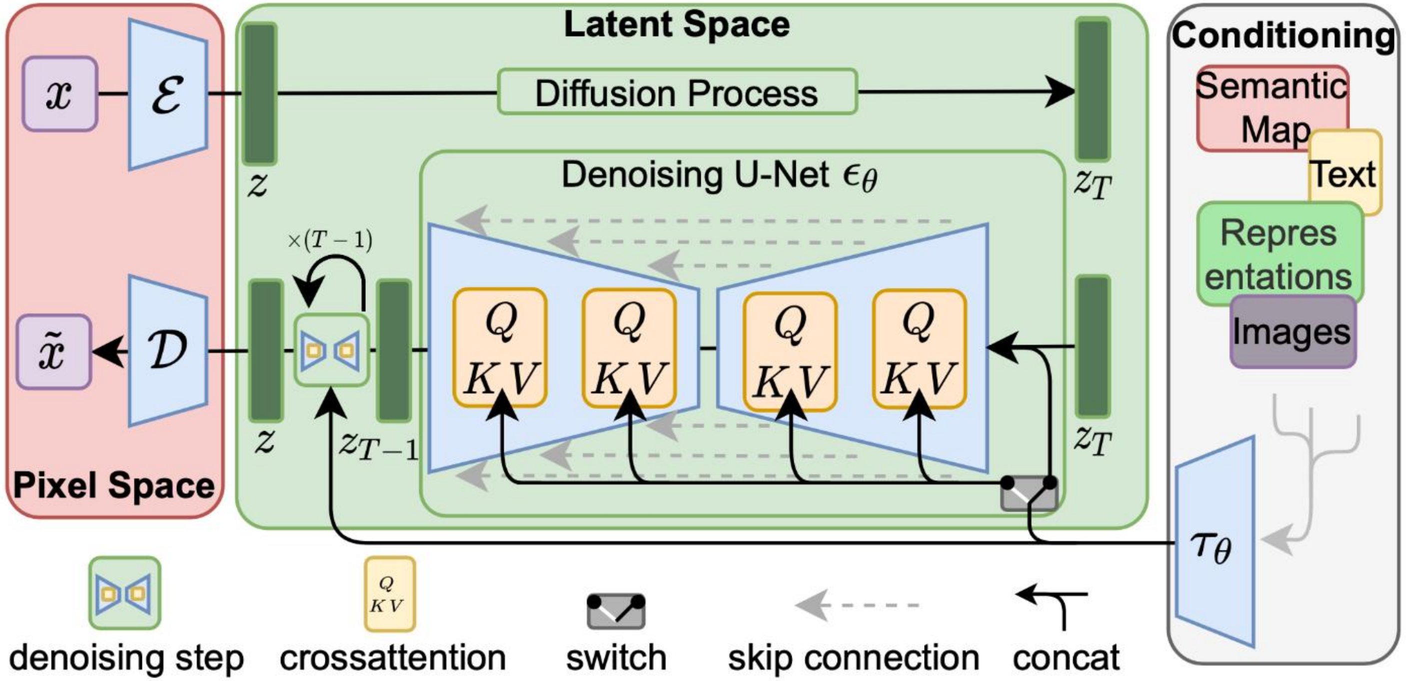
- Stable Diffusion
  - Training Data：5.8 Billion Image-Text Pairs
  - Parameters: More than 1 Billion



BACKGROUND: Diffusion

Text-to-Image Diffusion Model

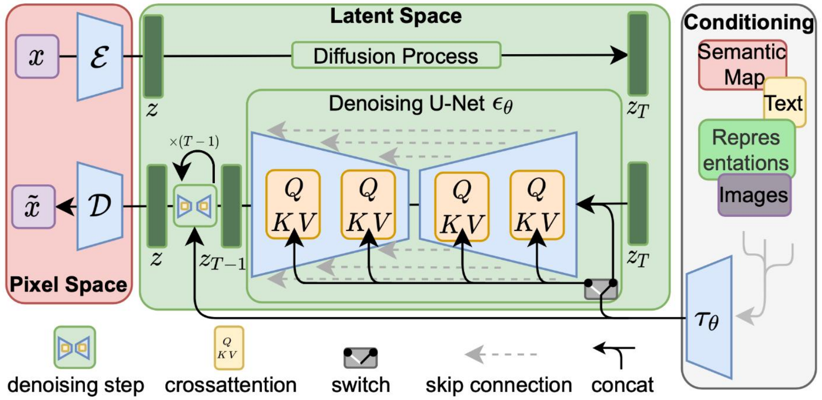
- Stable Diffusion
  - Only be used for generation?



BACKGROUND: Diffusion

Text-to-Image Diffusion Model

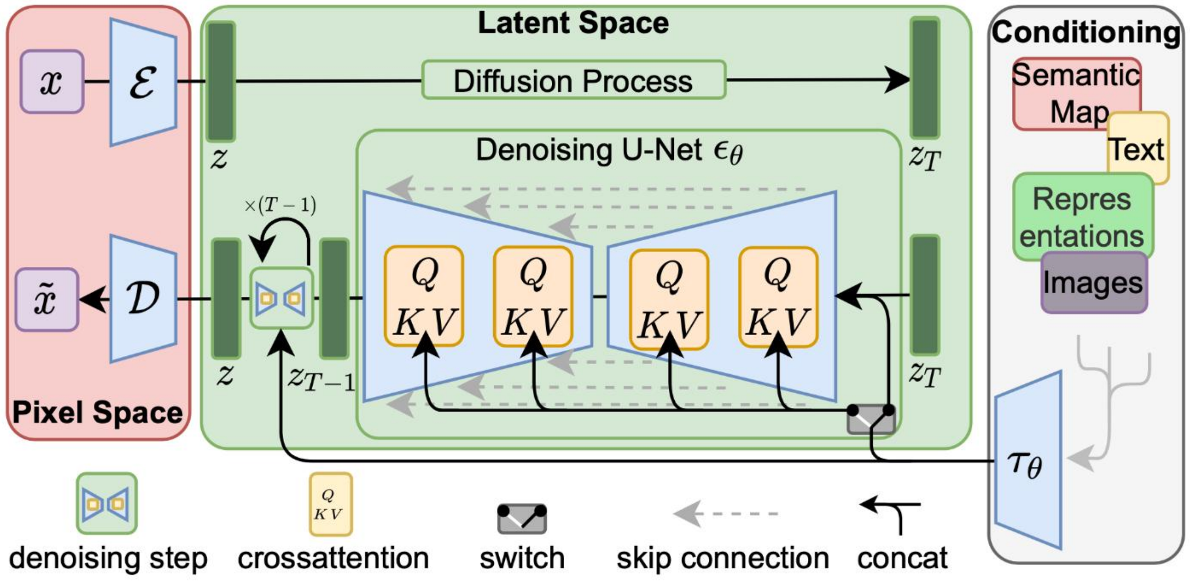
- Stable Diffusion
  - Only be used for generation?



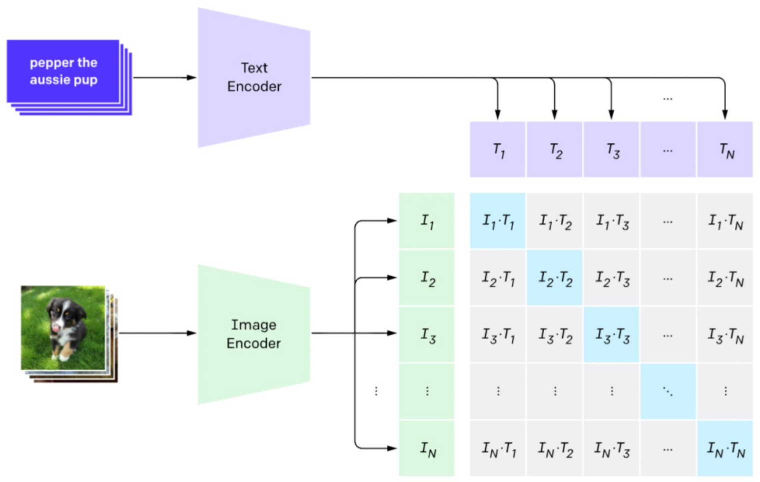
Input Image

K-Means Clustering of Frozen Diffusion Features

BACKGROUND: Motivation



+



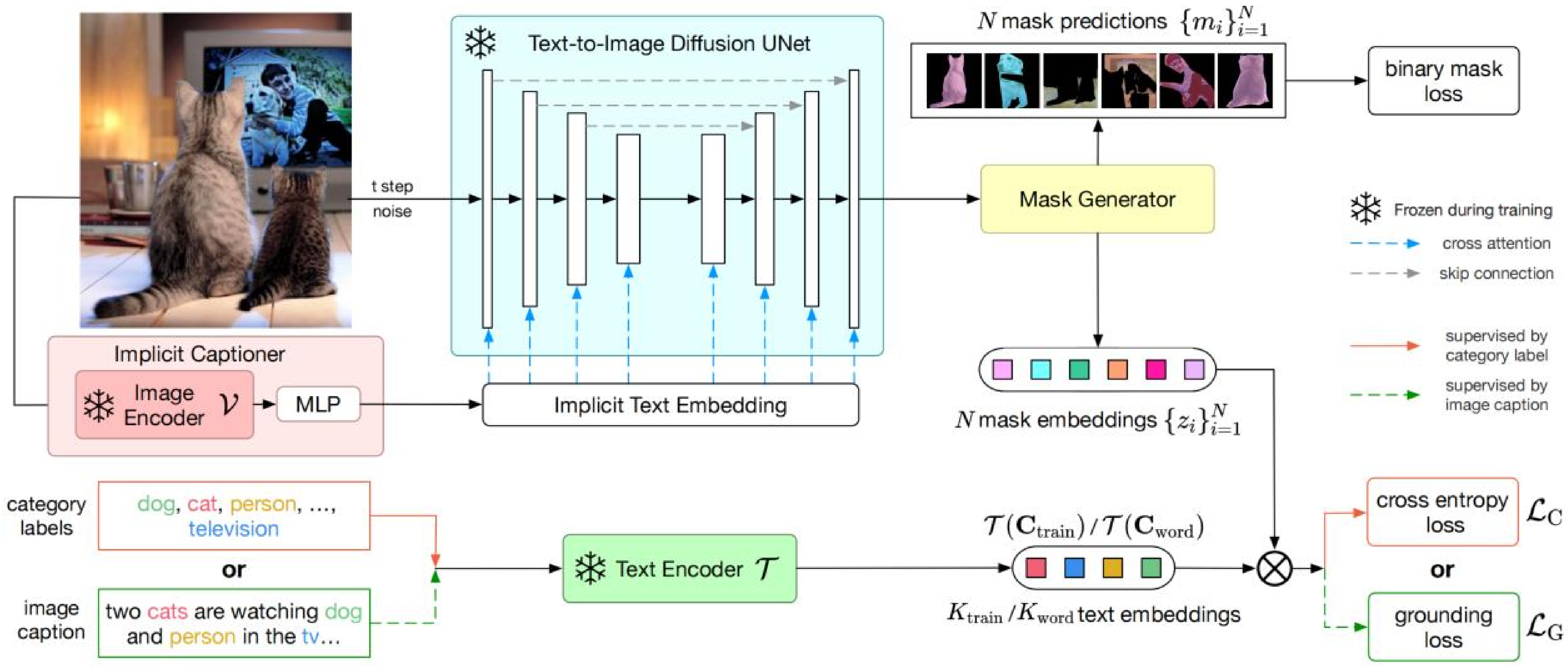
?

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

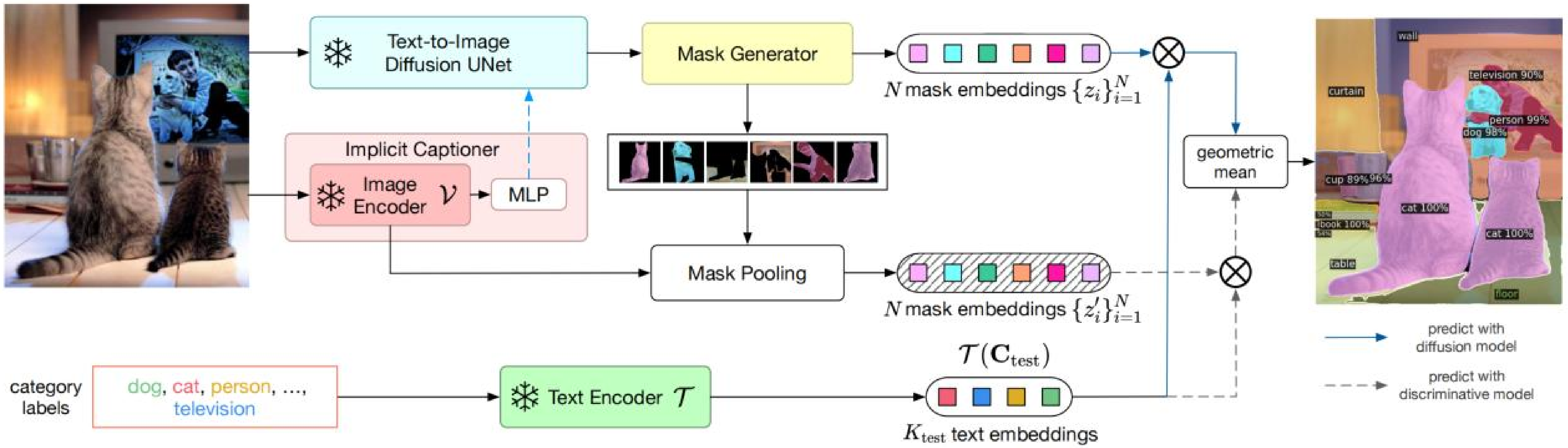
METHOD

Training Pipeline



METHOD

Testing Pipeline



OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

EXPERIMENTS

Text-to-Image Diffusion Model: Stable Diffusion

Diffusion Time Step t = 0

Mask generator: Mask2Former

Details

- Training for 90k iterations on COCO(BSZ=64)
- 28.1M trainable parameters(1.8%)

Evaluation

- Panoptic Quality(PQ), mAP, mIoU

$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}.$$

EXPERIMENTS: Comparisons

Comparisons on Open-Vocabulary Panoptic Segmentation

- Training on COCO, testing on ADE20K

	Supervision			ADE20K			COCO		
Method	label	mask	caption	PQ	mAP	mIoU	PQ	mAP	mIoU
MaskCLIP [16]	✓	✓		15.1	6.0	23.7	-	-	-
ODISE (Ours)	✓	✓		**22.6**	**14.4**	**29.9**	**55.4**	**46.0**	**65.2**
ODISE (Ours)		✓	✓	**23.4**	**13.9**	**28.7**	**45.6**	**38.4**	**52.4**

EXPERIMENTS: Comparisons

Comparisons on Open-Vocabulary Semantic Segmentation

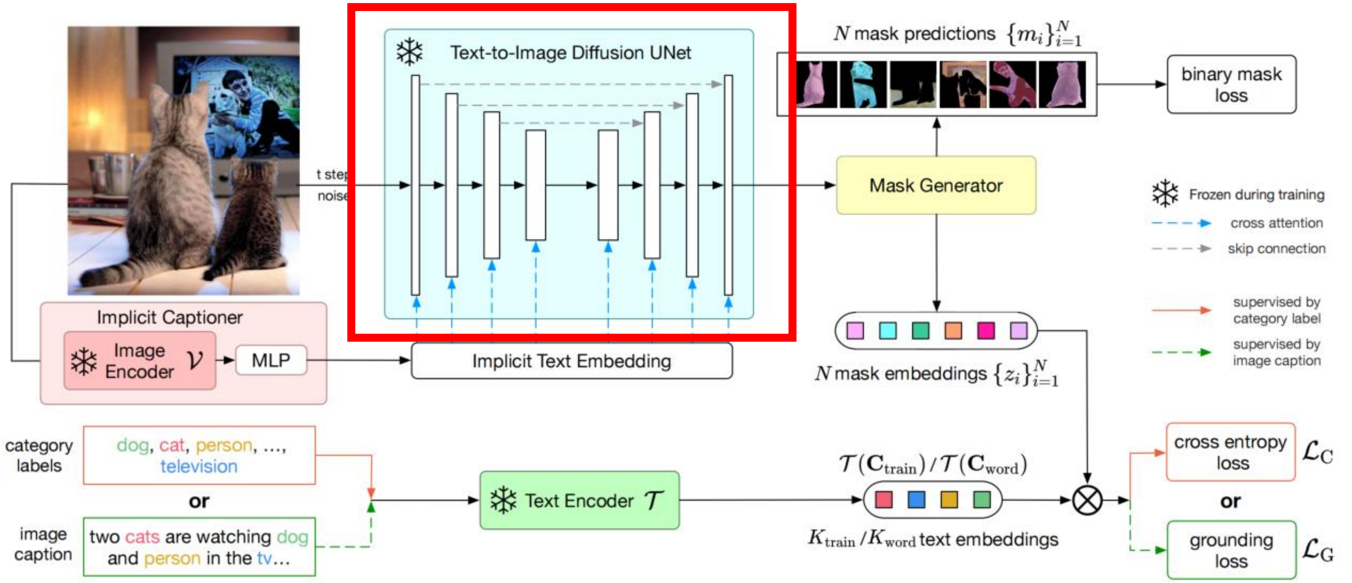
- ADE20K: A-150, A-847
- Pascal Context: PC-59, PC-459
- Pascal VOC dataset

Method	Training Dataset	Supervision label	Supervision mask	Supervision caption	mIoU A-847	mIoU PC-459	mIoU A-150	mIoU PC-59	mIoU PAS-21	mIoU COCO
SPNet [82]	Pascal VOC	✓	✓		-	-	-	24.3	18.3	-
ZS3Net [4]	Pascal VOC	✓	✓		-	-	-	19.4	38.3	-
LSeg [40]	Pascal VOC	✓	✓		-	-	-	-	47.4	-
SimBaseline [84]	COCO	✓	✓		-	-	15.3	-	74.5	-
ZegFormer [15]	COCO	✓	✓		-	-	16.4	-	73.3	-
LSeg+ [23]	COCO	✓	✓		3.8	7.8	18.0	46.5	-	55.1
MaskCLIP [16]	COCO	✓	✓		8.2	10.0	23.7	45.9	-	-
ODISE (Ours)	COCO	✓	✓		**11.1**	**14.5**	**29.9**	**57.3**	**84.6**	**65.2**
GroupViT [83]	GCC+YFCC			✓	4.3	4.9	10.6	25.9	50.7	21.1
OpenSeg [23]	COCO		✓	✓	6.3	9.0	21.1	42.1	-	36.1
ODISE (Ours)	COCO		✓	✓	**11.0**	**13.8**	**28.7**	**55.3**	**82.7**	**52.4**

EXPERIMENTS: Ablation Study

Visual Representations

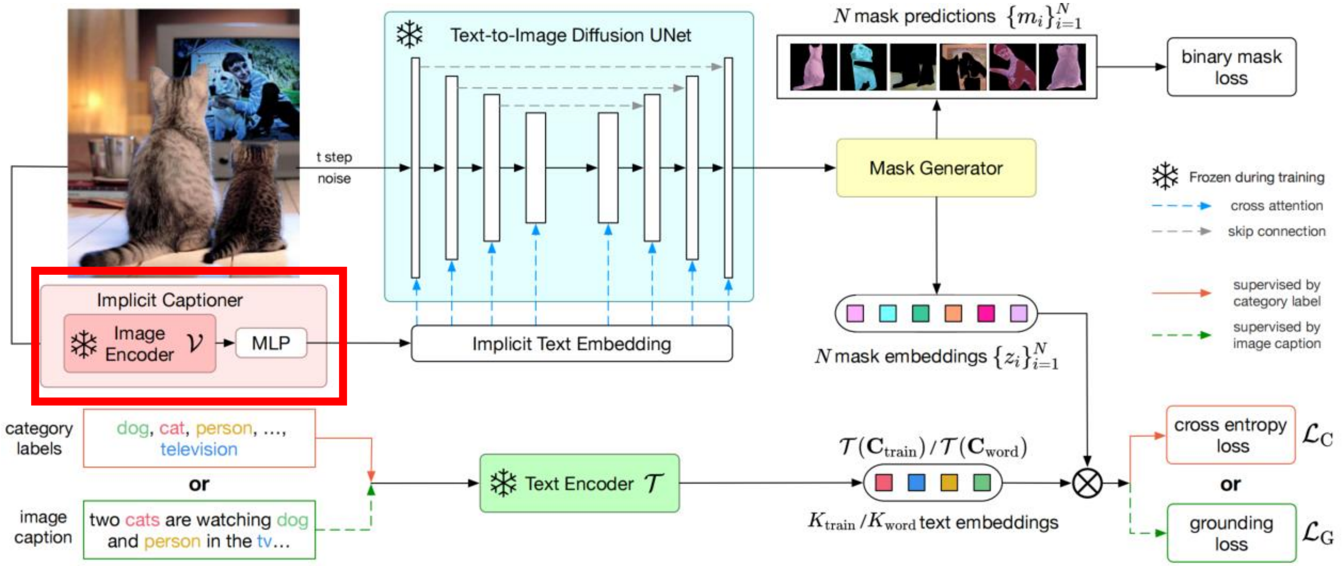
Model	Training Data	ADE20K PQ	ADE20K mAP	ADE20K mIoU	COCO PQ	COCO mAP	COCO mIoU
Pre-trained with class labels							
DeiT-v3(H) [75]	IN-21k	21.4	11.4	28.0	41.4	29.2	52.3
Swin(H) [51]	IN-21k	20.9	10.7	27.7	42.4	31.6	54.0
ConvNeXt(H) [52]	IN-21k	21.0	11.0	27.8	43.1	33.1	54.3
MViT(H) [46]	IN-21k	21.1	11.6	28.1	44.0	36.3	**54.5**
LDM [66]	IN-1k	20.7	10.9	26.5	41.7	35.3	50.6
Pre-trained with self-supervision							
MoCo-v3(H) [8]	IN-1k	19.3	9.6	25.8	37.1	26.8	47.1
DINO(B) [6]	IN-1k	20.6	10.5	26.3	39.5	29.8	49.5
MAE(H) [28]	IN-1k	21.5	10.9	27.6	37.9	31.6	46.3
BEiT-v2(H) [61]	IN-21k	21.4	11.4	28.0	41.4	29.2	52.3
Pre-trained with text							
CLIP(L) [62]	WIT	20.4	9.6	27.0	40.6	26.7	52.1
CLIP(H) [62]	LAION	21.2	10.8	28.1	41.0	27.9	52.1
ODISE	LAION	**23.3**	**13.0**	**29.2**	**44.2**	**38.3**	53.8



EXPERIMENTS: Ablation Study

Captioning Generators

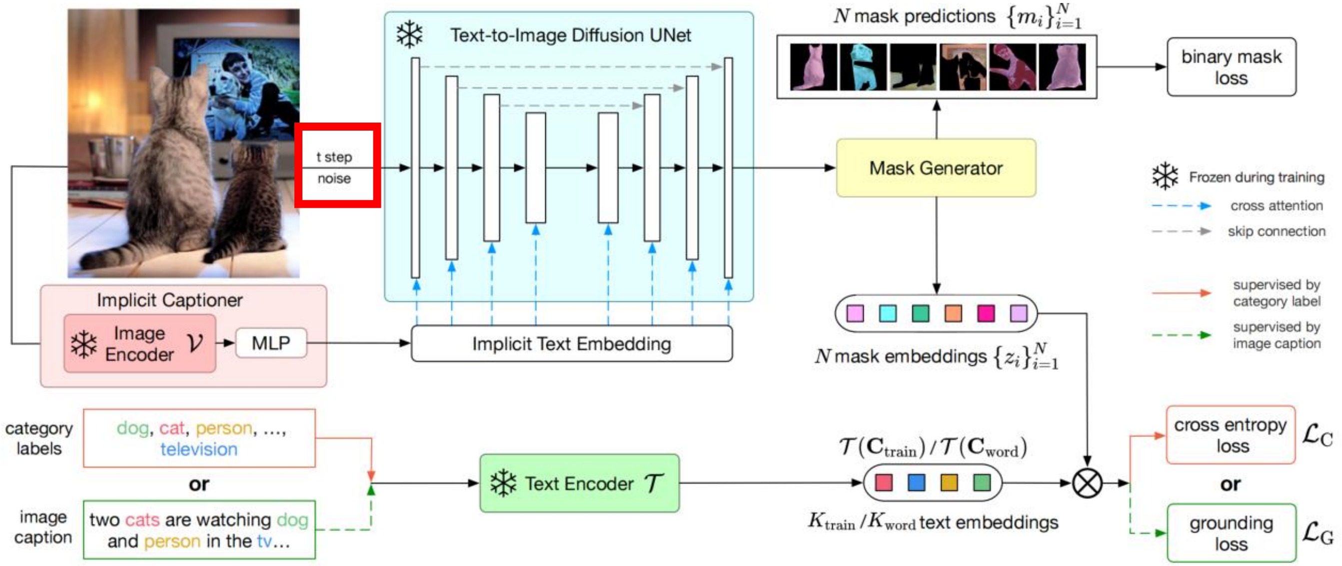
	Captioner	ADE20K PQ	ADE20K mAP	ADE20K mIoU	COCO PQ	COCO mAP	COCO mIoU
(a)	Empty	21.8	11.8	27.3	43.5	37.0	52.3
(b)	Heuristic [90]	22.2	12.1	28.1	44.0	36.3	53.3
(c)	BLIP [43]	22.3	12.4	28.2	44.1	37.1	53.6
(d)	Implict	**23.3**	**13.0**	**29.2**	**44.2**	**38.3**	**53.8**



EXPERIMENTS: Ablation Study

Diffusion Time Steps

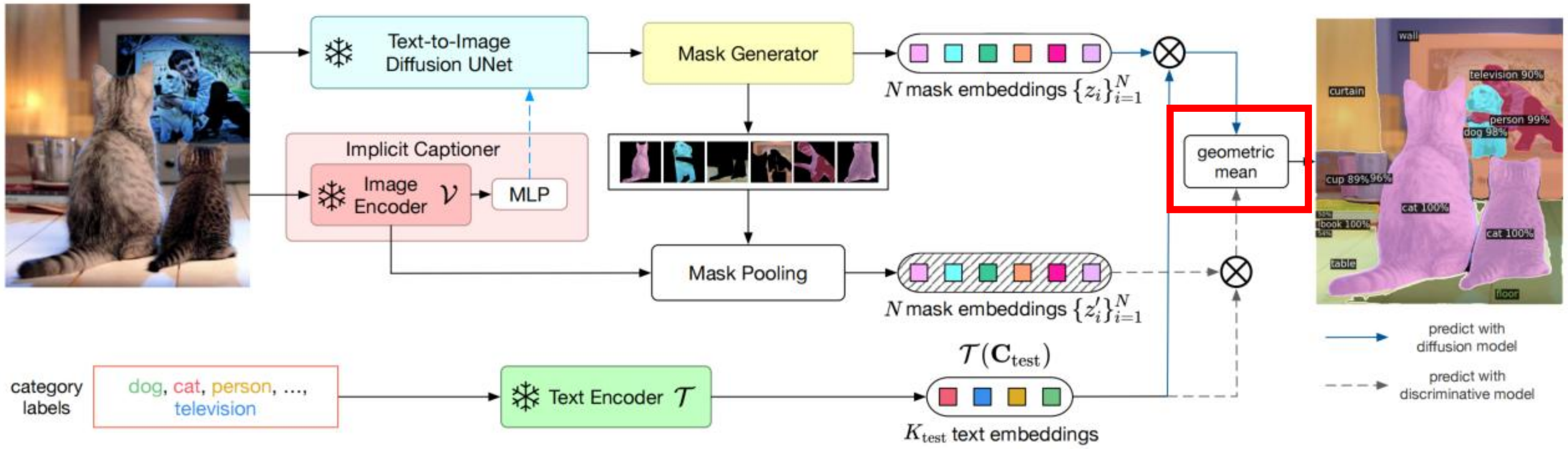
time step	ADE20K PQ	ADE20K mAP	ADE20K mIoU	COCO PQ	COCO mAP	COCO mIoU
0	**23.3**	**13.0**	29.2	**44.2**	**38.3**	**53.8**
100	22.8	12.5	29.3	43.2	36.4	52.3
200	21.5	11.9	28.0	42.4	35.1	51.7
500	20.9	11.1	27.0	38.2	31.1	47.6
0+100+200	23.1	12.9	**29.7**	43.7	37.4	53.0
learnable	22.8	12.9	29.2	44.0	37.5	53.4



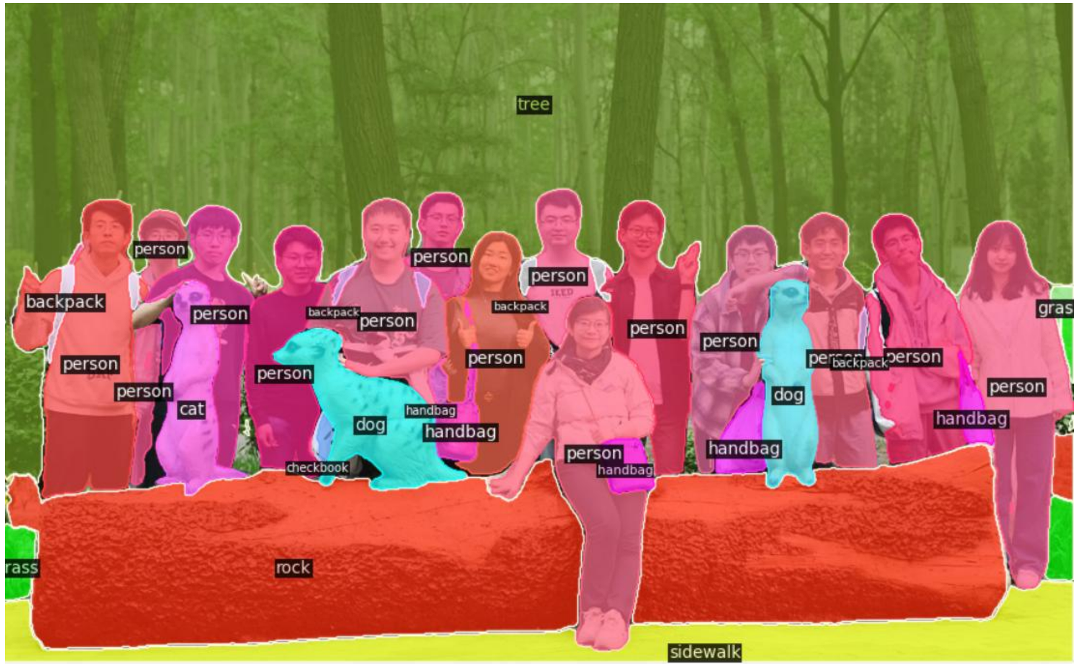
EXPERIMENTS: Ablation Study

Mask Classifiers

model		ADE20K			COCO		
diffusion	discriminative	PQ	mAP	mIoU	PQ	mAP	mIoU
	✓	15.0	9.6	17.5	26.5	23.5	23.6
✓		20.1	10.3	24.4	42.3	37.8	52.0
✓	✓	**23.3**	**13.0**	**29.2**	**44.2**	**38.3**	**53.8**



EXPERIMENTS: Other attempts



sidewalk, mongoose

EXPERIMENTS: Other attempts

sidewalk, mongoose

OUTLINE

- Authorship
- Background
- Method
- Experiments
- Conclusion

CONCLUSION

- Taking the first step in leveraging the frozen internal representation of large-scale text-to-image diffusion models for downstream recognition tasks
- This work demonstrates that text-to-image diffusion models are not only capable of generating plausible image but also of learning rich semantic representations

Thanks for listening!