北京大学 PEKING UNIVERSITY

CleanDIFT: Diffusion Features without Noise

Nick Stracke*, Stefan Andreas Baumann*, Kolja Bauer*, Frank Fundel, Björn Ommer

CompVis @ LMU Munich

CVPR 2025 (Oral)

Presenter: Yixuan Zou

2026.3.15

Background

北京大学 PEKING UNIVERSITY

Can generative models provide useful visual representations?

“What I cannot create, I do not understand” —— Richard Feynman

Generative models can generate realistic images

→ Generative models learn image structure & semantics

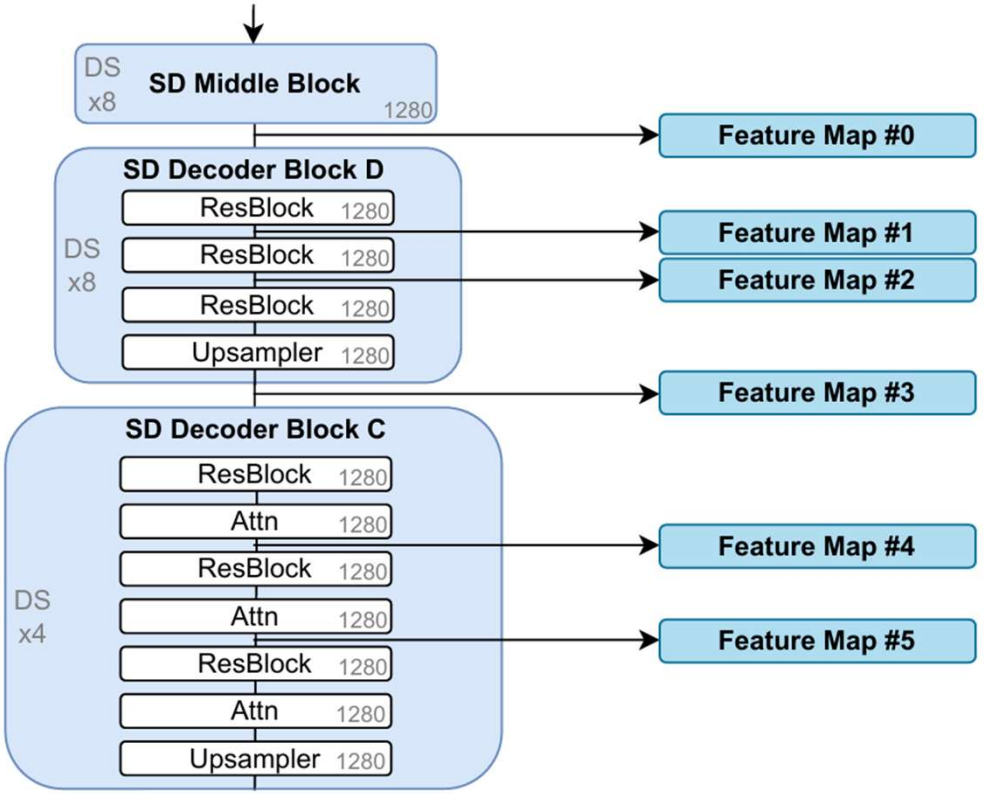
→ Generative models learn visual representations

Background

Diffusion model learns visual representations

Diffusion features are intermediate U-Net representations

We can use **diffusion features** for downstream tasks



北京大学 PEKING UNIVERSITY

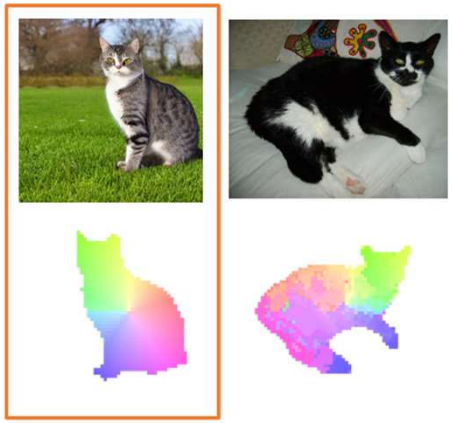
Diffusion features for downstream tasks

DIFT

Vanilla feature utilization

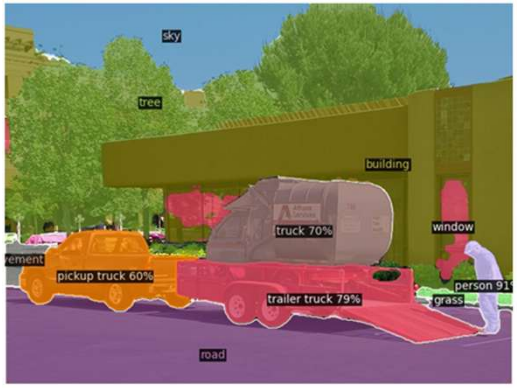
A Tale of Two Features

Diffusion-DINOv2 fusion



ODISE

Open-vocabulary segmentation



EmerDiff

Semantic segmentaion

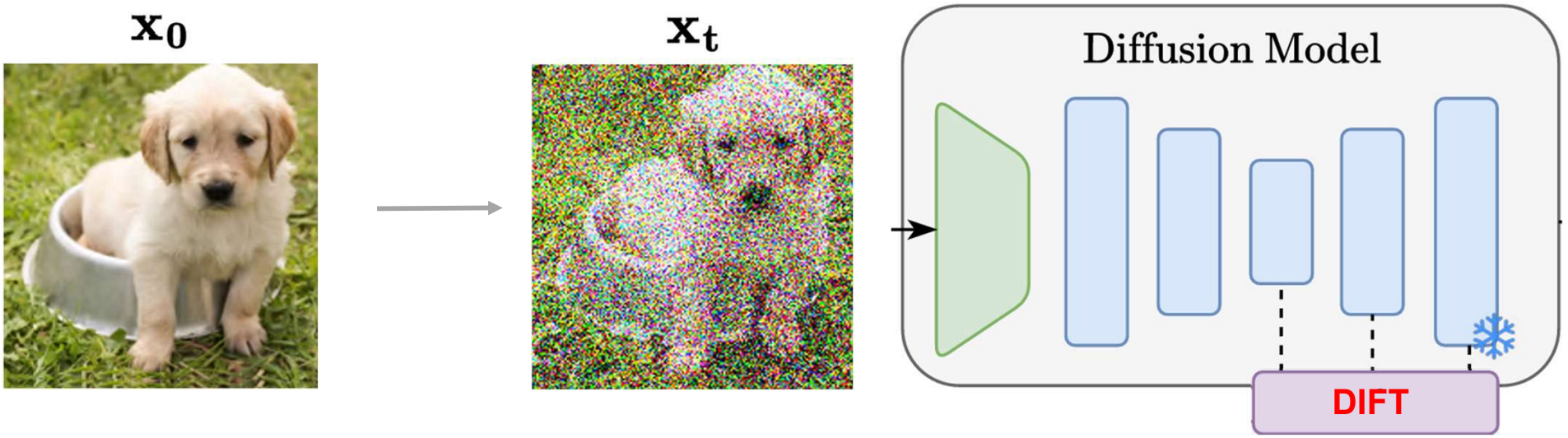
Semantic Correspondence Detection

Find semantically matching points across different images

DIffusion FeaTures (DIFT)

We can use **DIFT** to extract correspondences

How to extract DIFT on real images? → Simply add noise



DIffusion FeaTures (DIFT)

How to use DIFT to find correspondence? → Feature matching

Source Point p_1

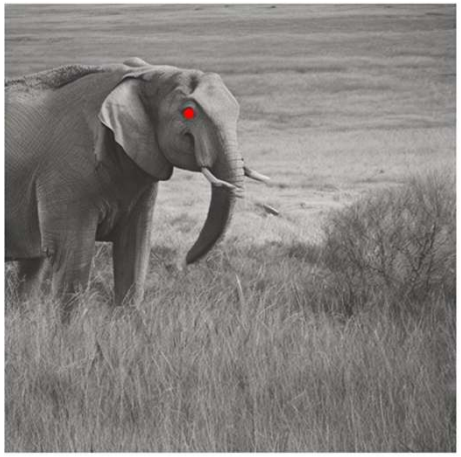


Image I_1

$$p_2 = \arg\min_p d(F_1(p_1), F_2(p))$$

F_1 is DIFT of I_1

F_2 is DIFT of I_2

d is cosine distance

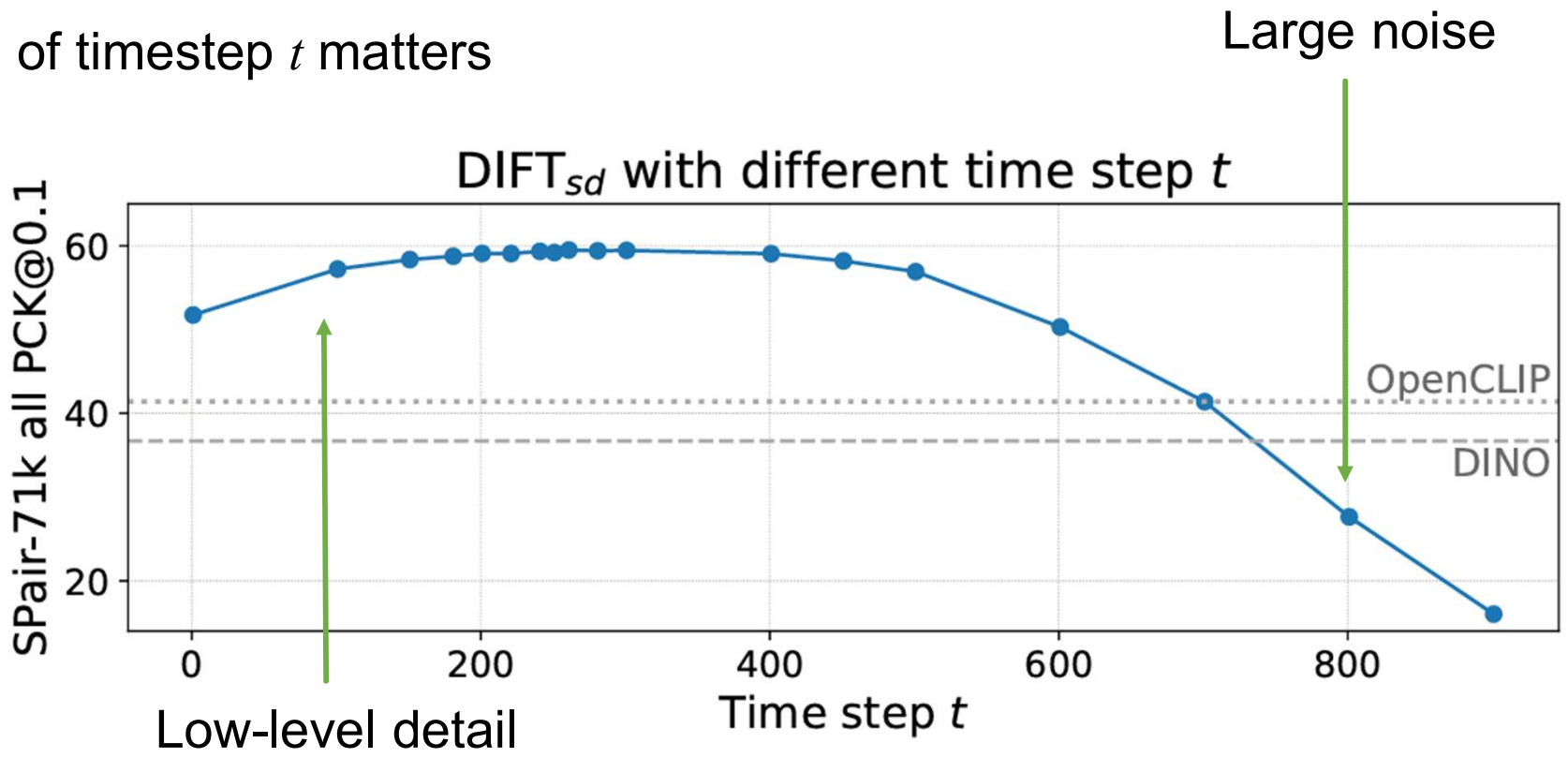
Predicted Target Point p_2

Image I_2

Emergent Correspondence from Image Diffusion, NeurIPS 2023

DIffusion FeaTures (DIFT)

Choice of timestep t matters

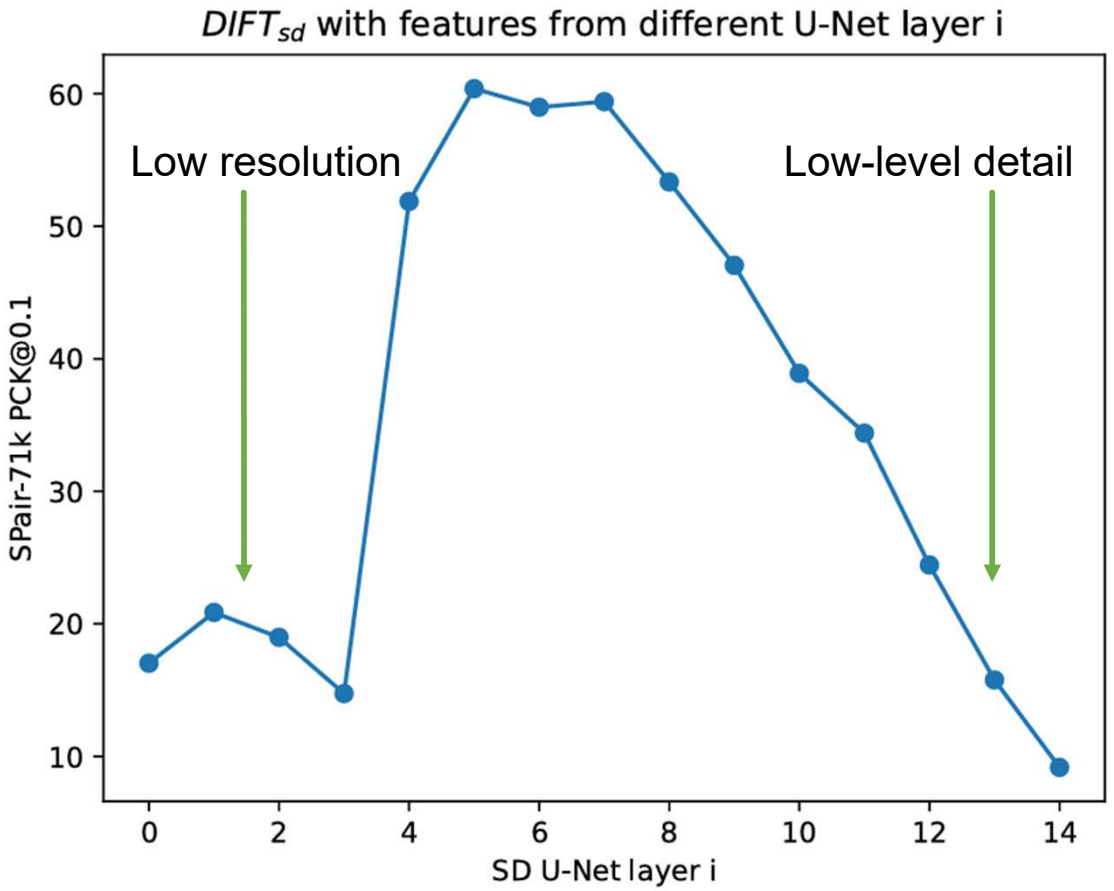


Emergent Correspondence from Image Diffusion, NeurIPS 2023

DIffusion FeaTures (DIFT)

Choice of U-Net layer matters

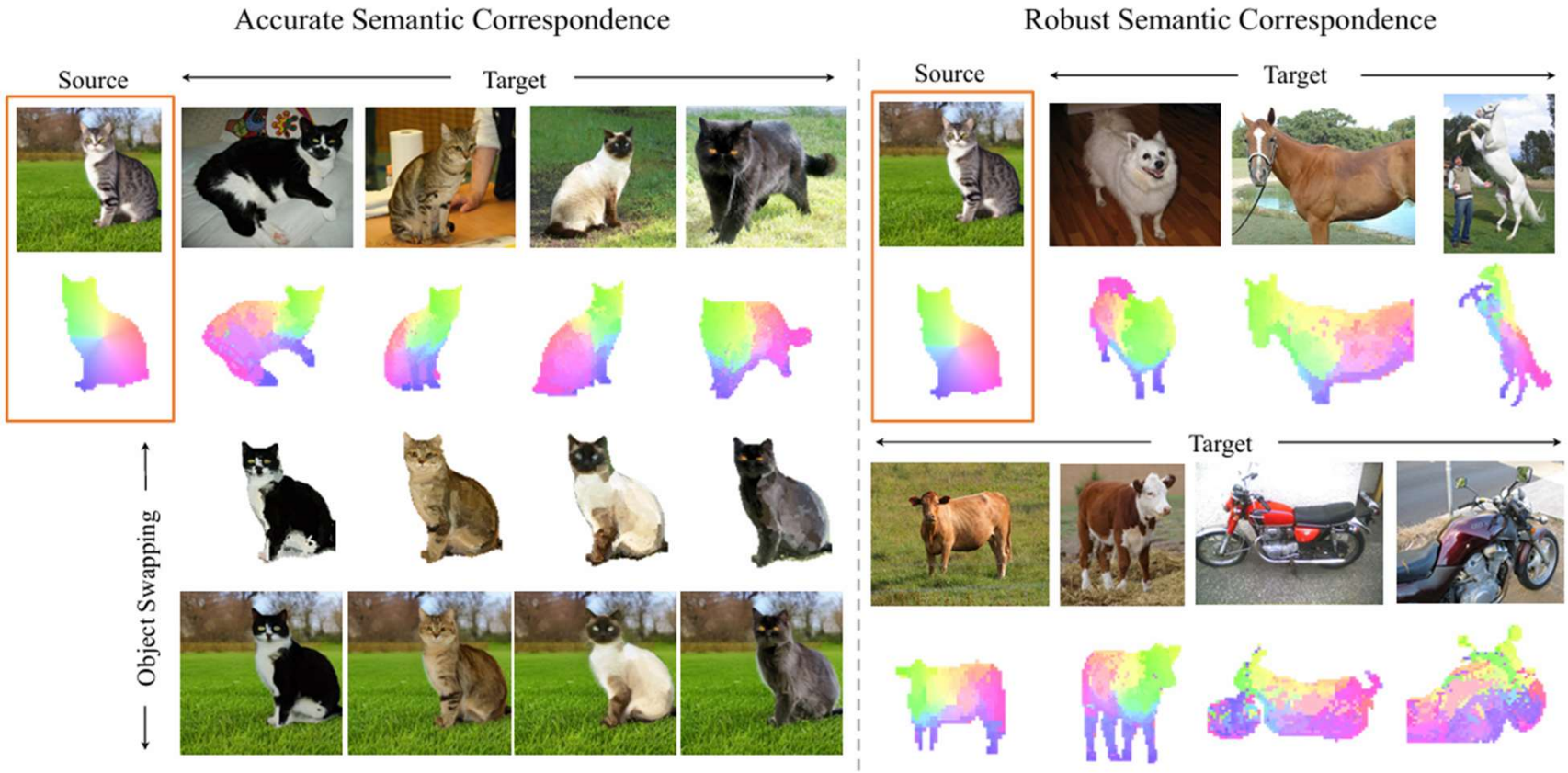
- Layers 0-3 (UpBlock1)
- Layers 4-7 (UpBlock2)
- Layers 8-11 (UpBlock3)
- Layers 12-14 (UpBlock4)



Emergent Correspondence from Image Diffusion, NeurIPS 2023

北京大学 PEKING UNIVERSITY

A Tale of Two Features



A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence, NeurIPS 2023

A Tale of Two Features

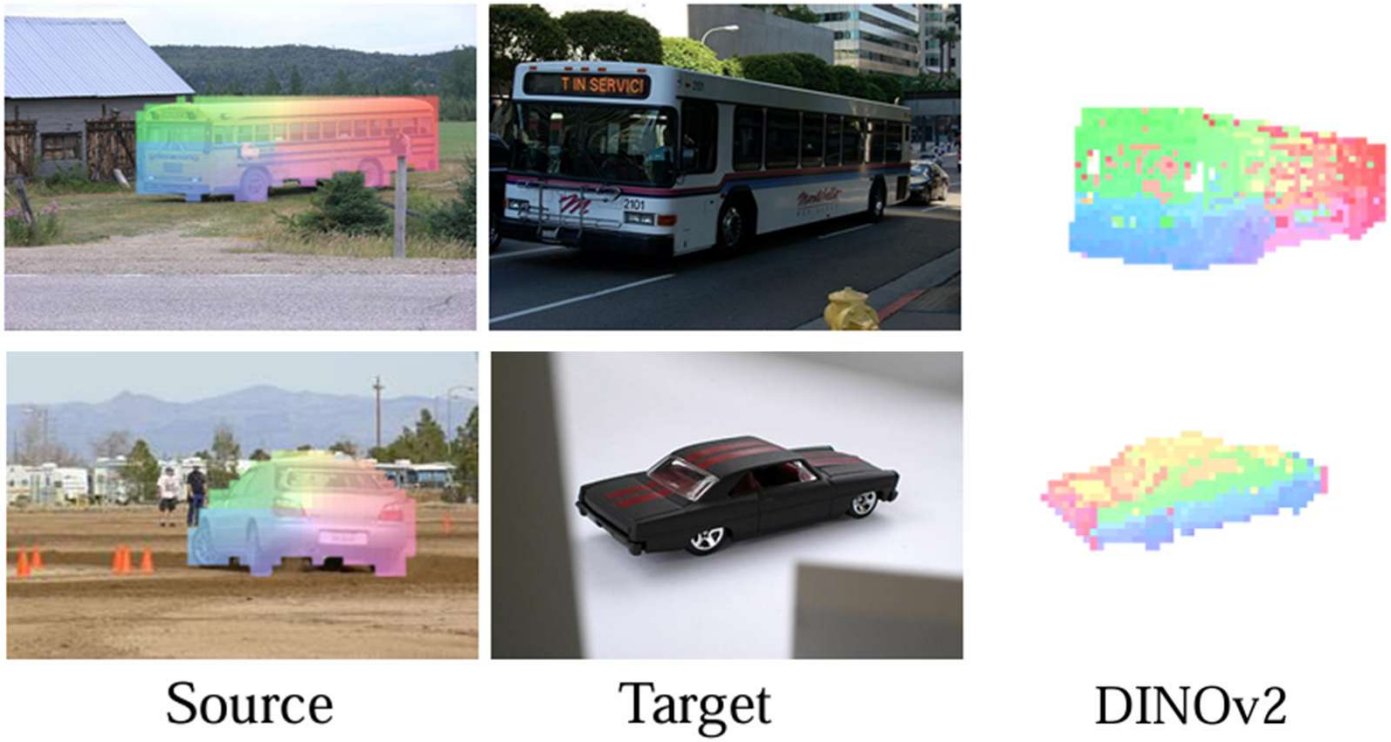
Properties of diffusion features

- Spatially Coherent
- Semantically Inaccurate

A Tale of Two Features

Properties of DINOv2 features

- Spatially InCoherent
- Semantically Accuarate



北京大学 PEKING UNIVERSITY

A Tale of Two Features

Fuse diffusion and DINO features is effective

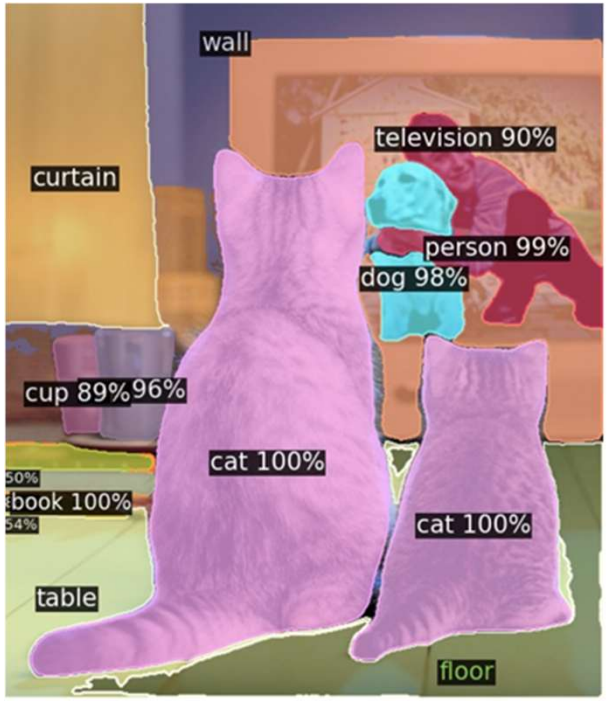
$$\mathcal{F}_{\text{FUSE}} = (\alpha||\mathcal{F}_{\text{SD}}||_2,\ (1-\alpha)||\mathcal{F}_{\text{DINO}}||_2)$$

	Method	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Dog	Horse	Motor	Person	Plant	Sheep	Train	TV	**All**
U^N	DINOv1-ViT-S/8 [2]	57.2	24.1	67.4	24.5	26.8	29.0	27.1	52.1	15.7	42.4	43.3	30.1	23.2	40.7	16.6	24.1	31.0	24.9	33.3
	DINOv2-ViT-B/14	72.7	62.0	85.2	**41.3**	40.4	52.3	51.5	71.1	36.2	67.1	64.6	67.6	61.0	68.2	30.7	62.0	54.3	24.2	55.6
	Stable Diffusion (**Ours**)	63.1	55.6	80.2	33.8	44.9	49.3	47.8	74.4	38.4	70.8	53.7	61.1	54.4	55.0	54.8	53.5	65.0	53.3	57.2
	Fuse-ViT-B/14 (**Ours**)	**73.0**	**64.1**	**86.4**	40.7	**52.9**	**55.0**	**53.8**	**78.6**	**45.5**	**77.3**	**64.7**	**69.7**	**63.3**	**69.2**	**58.4**	**67.6**	**66.2**	**53.5**	**64.0**

Per-class and average PCK@0.10 on test split

ODISE

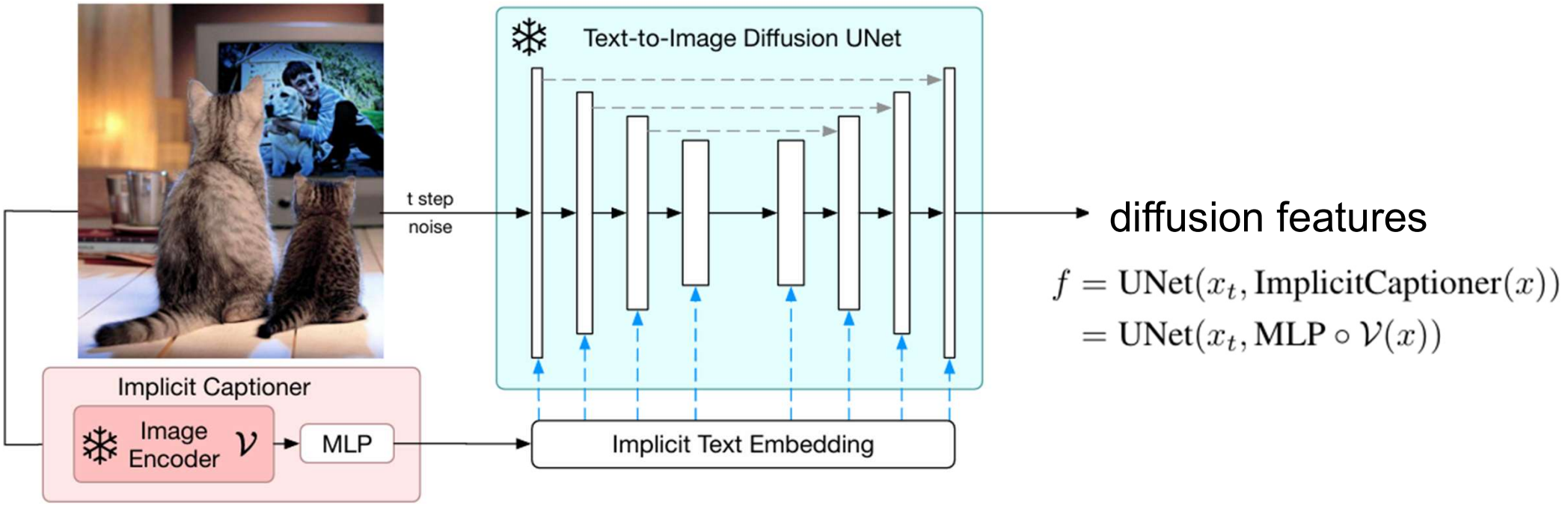
We can use diffusion features for panoptic segmentation



ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023

ODISE

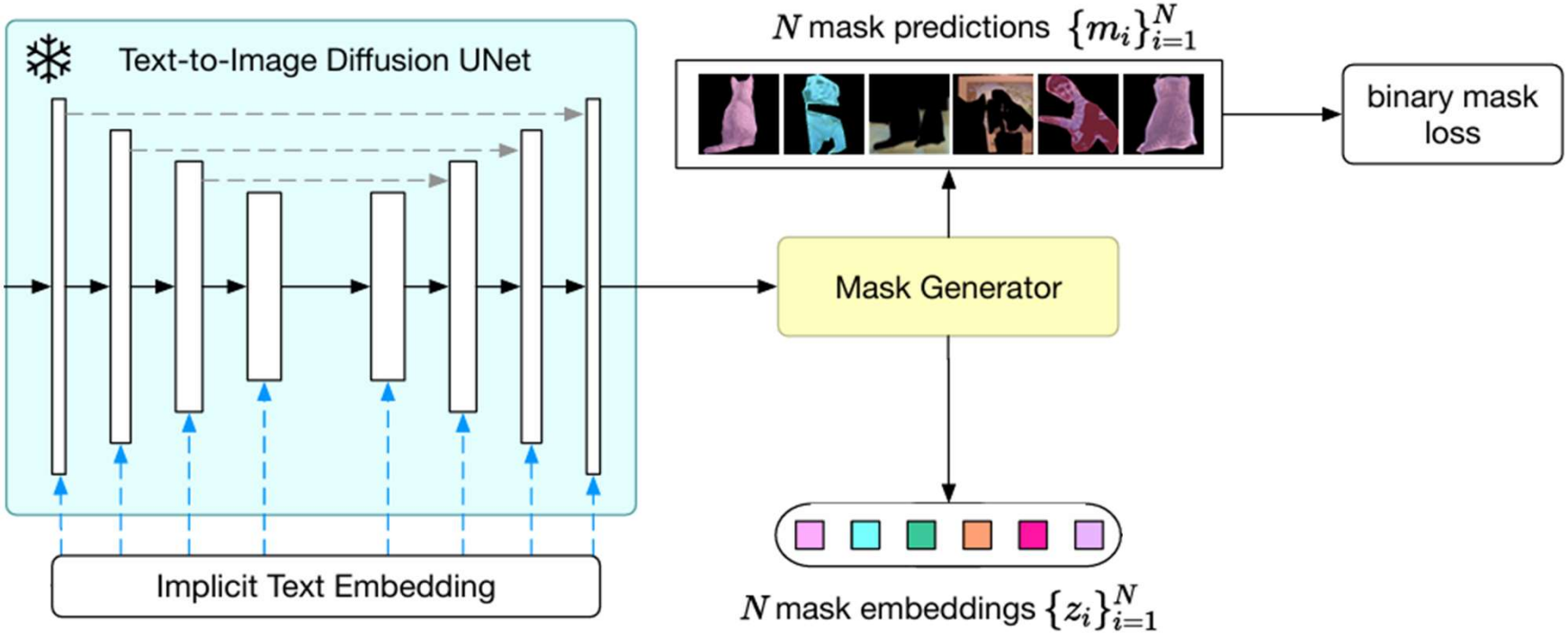
Extract diffusion features from a frozen Stable Diffusion UNet



ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023

ODISE

Obtain **binary masks** and **mask embedding features** from diffusion features



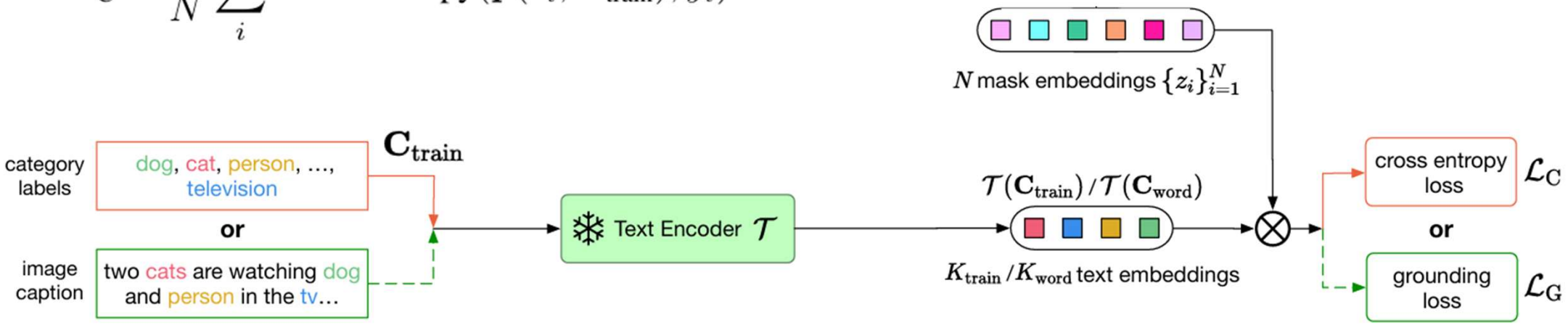
ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023

ODISE

Category label supervision and image caption supervision

$$\mathbf{p}(z_i, \mathbf{C}_{\text{train}}) = \text{Softmax}(z_i \cdot \mathcal{T}(\mathbf{C}_{\text{train}})/\tau)$$

$$\mathcal{L}_{\text{C}} = \frac{1}{N}\sum_{i}^{N} \text{CrossEntropy}(\mathbf{p}(z_i, \mathbf{C}_{\text{train}}), y_i)$$



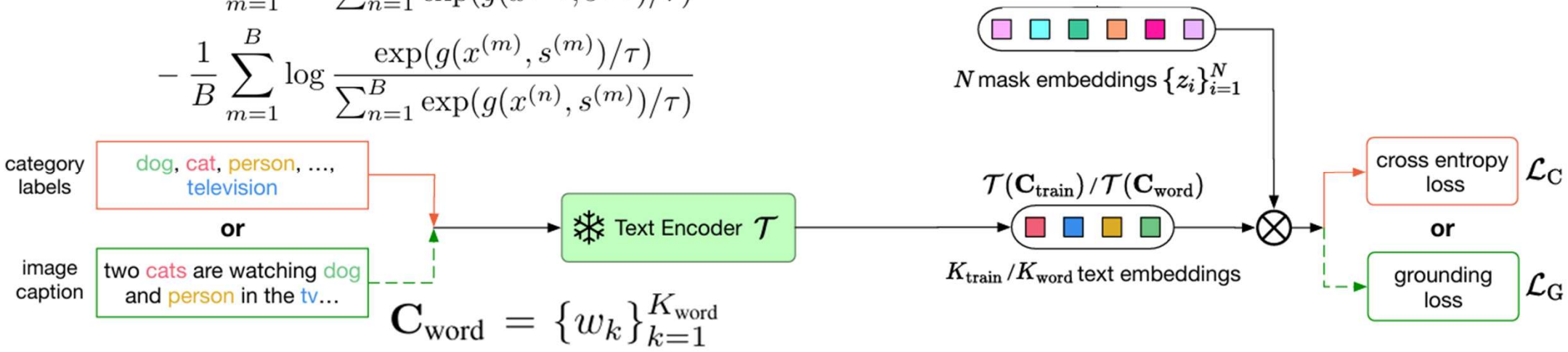
ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023

ODISE

Category label supervision and image caption supervision

$$g(x^{(m)}, s^{(m)}) = \frac{1}{K}\sum^{K}\sum^{N}\mathbf{p}(z_i, \mathbf{C}_{\text{word}})_k \cdot \langle z_i, \mathcal{T}(w_k)\rangle$$

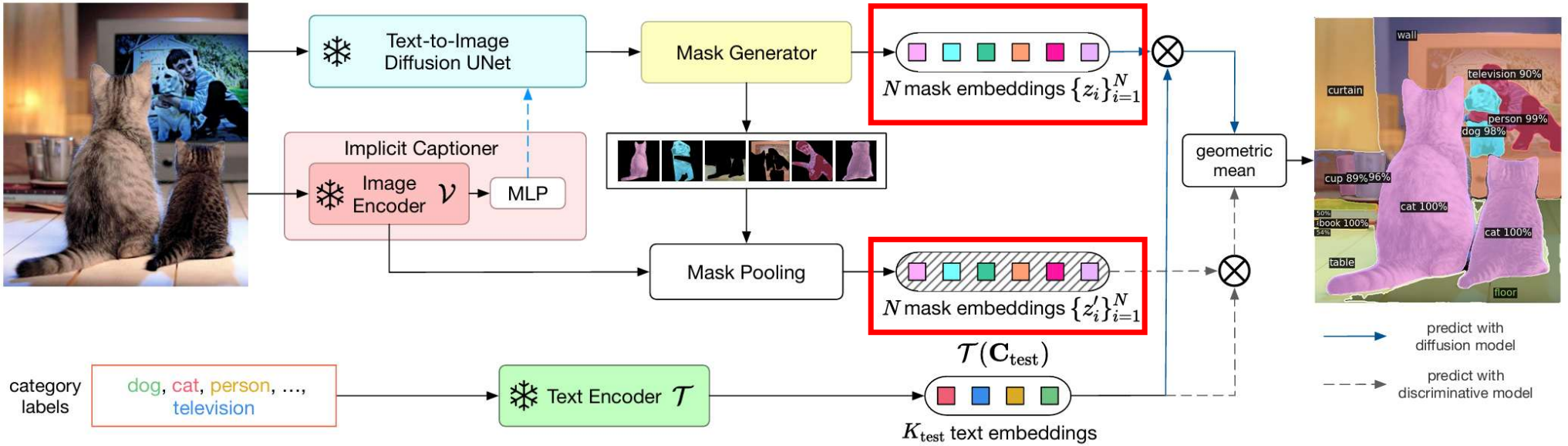
$$\mathcal{L}_{\text{G}} = -\frac{1}{B}\sum_{m=1}^{B}\log\frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^{B}\exp(g(x^{(m)}, s^{(n)})/\tau)} - \frac{1}{B}\sum_{m=1}^{B}\log\frac{\exp(g(x^{(m)}, s^{(m)})/\tau)}{\sum_{n=1}^{B}\exp(g(x^{(n)}, s^{(m)})/\tau)}$$



ODISE: Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models, CVPR 2023

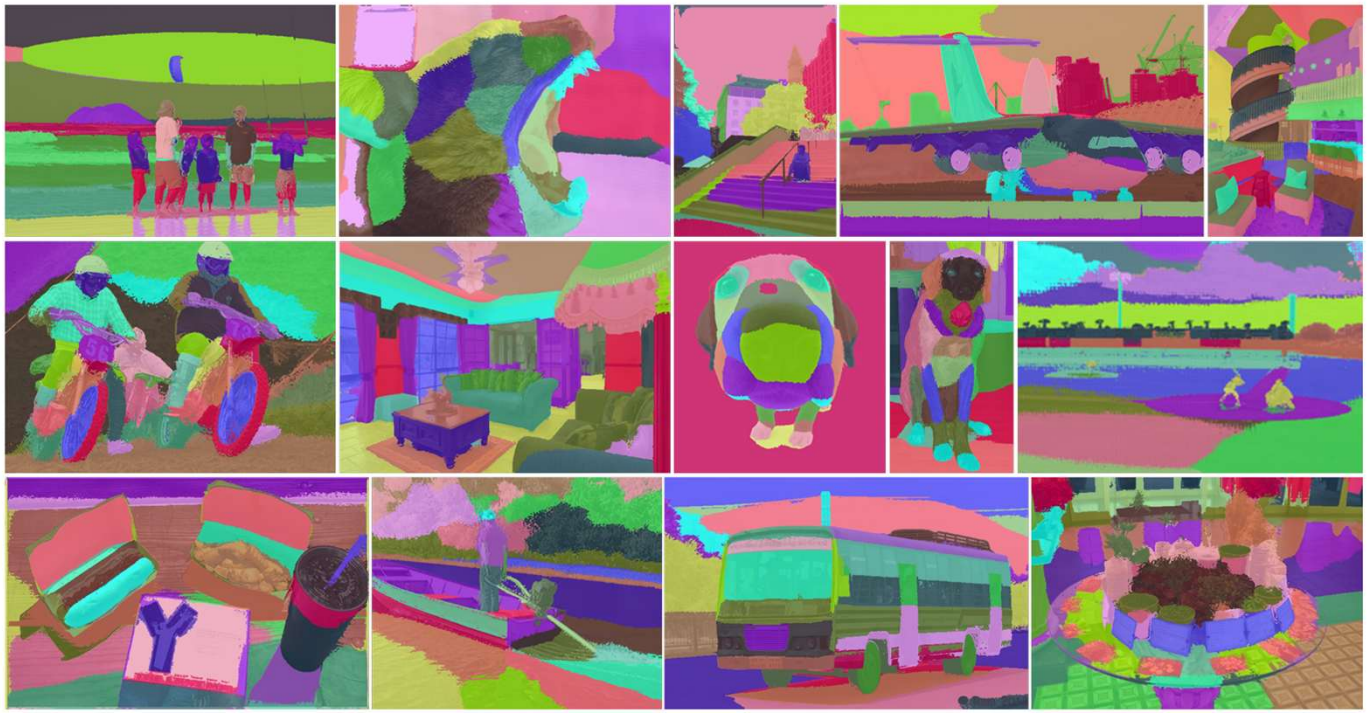
ODISE

For inference, fuse predictions from the **diffusion** and **discriminative models**



EmerDiff

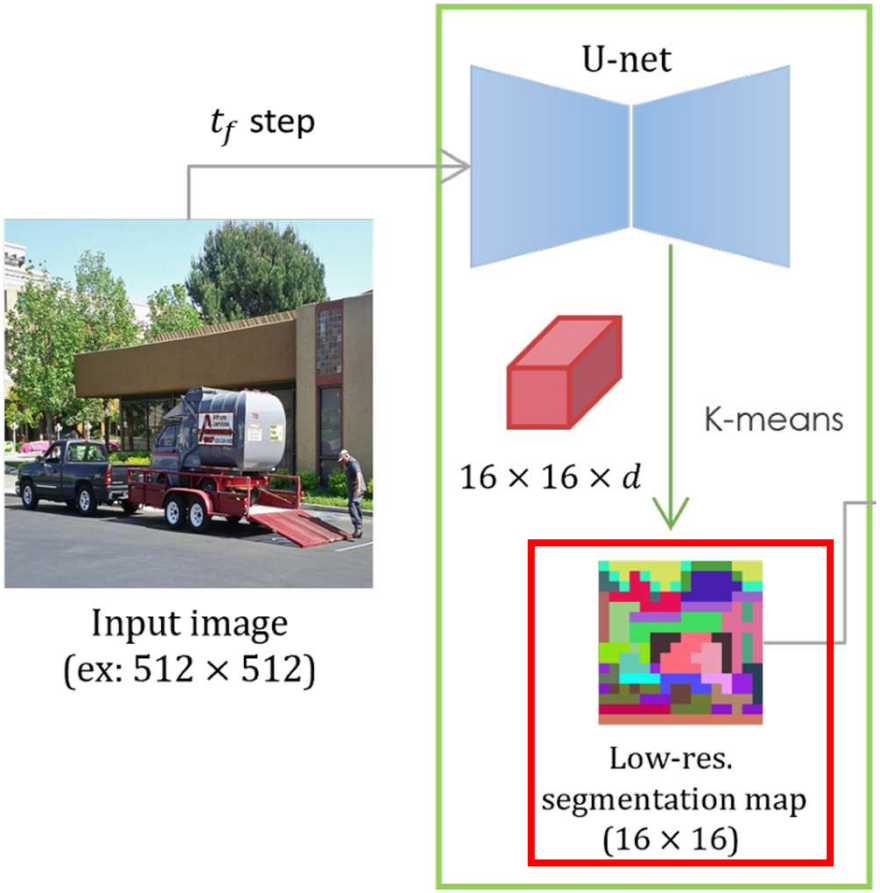
We can use diffusion features for semantic segmentation



EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models, ICLR 2024

EmerDiff

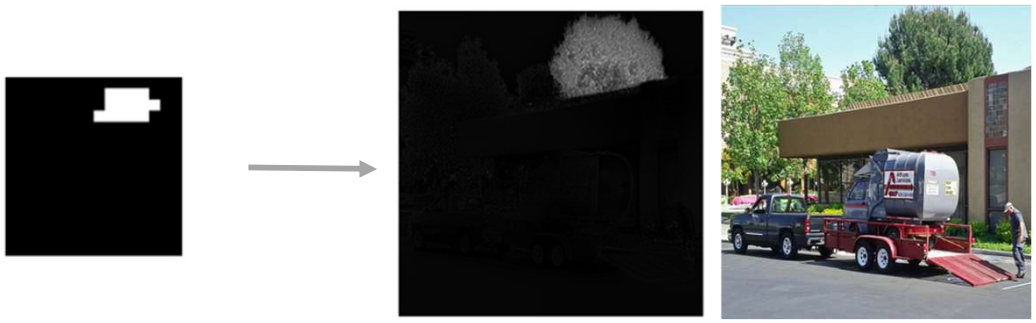
- Invert to get the diffusion feature
- Do K-means on diffusion feature
- Obtain a low-res. segmentation map

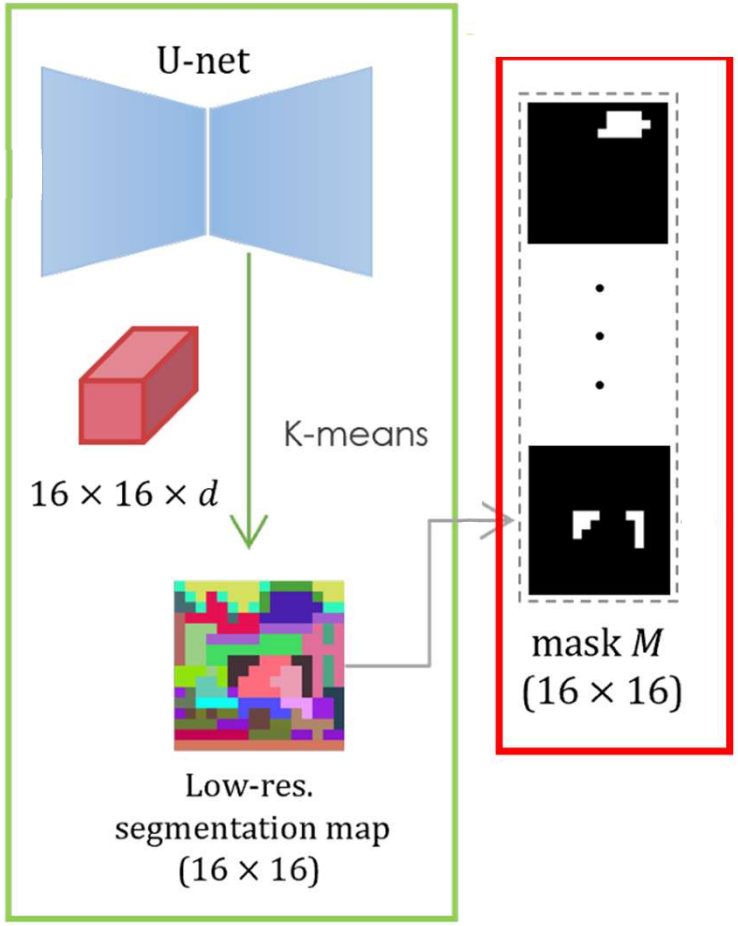


EmerDiff

We can get semantic masks from the map

How to identify **pixel–mask** correspondence?





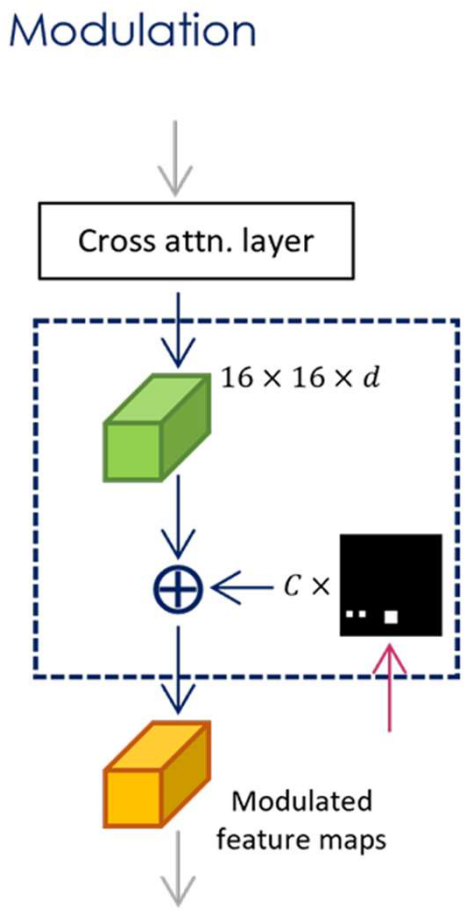
EmerDiff

Do modulation on diffusion feature by **adding the mask**

If some pixels change a lot, they are related to the mask

$$f\left(\sigma\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V\right)+cM\in\mathbb{R}^{hw\times d}$$

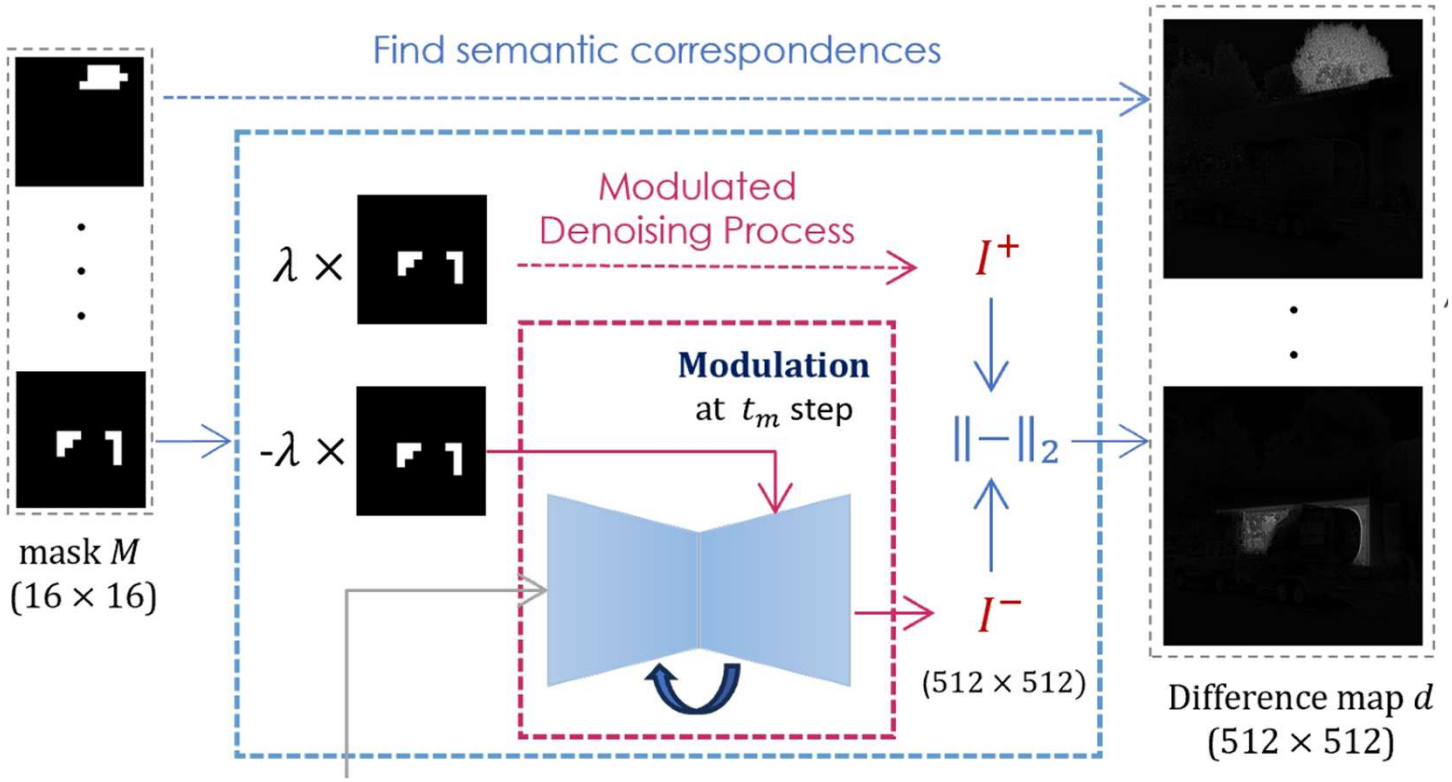
We can choose $c = +\lambda, -\lambda$ and observe the difference



EmerDiff

Compute the difference map $d = ||I^- - I^+||_2$

$$f\left(\sigma\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V\right) + cM \in \mathbb{R}^{hw\times d}$$



EmerDiff: Emerging Pixel-level Semantic Knowledge in Diffusion Models, ICLR 2024

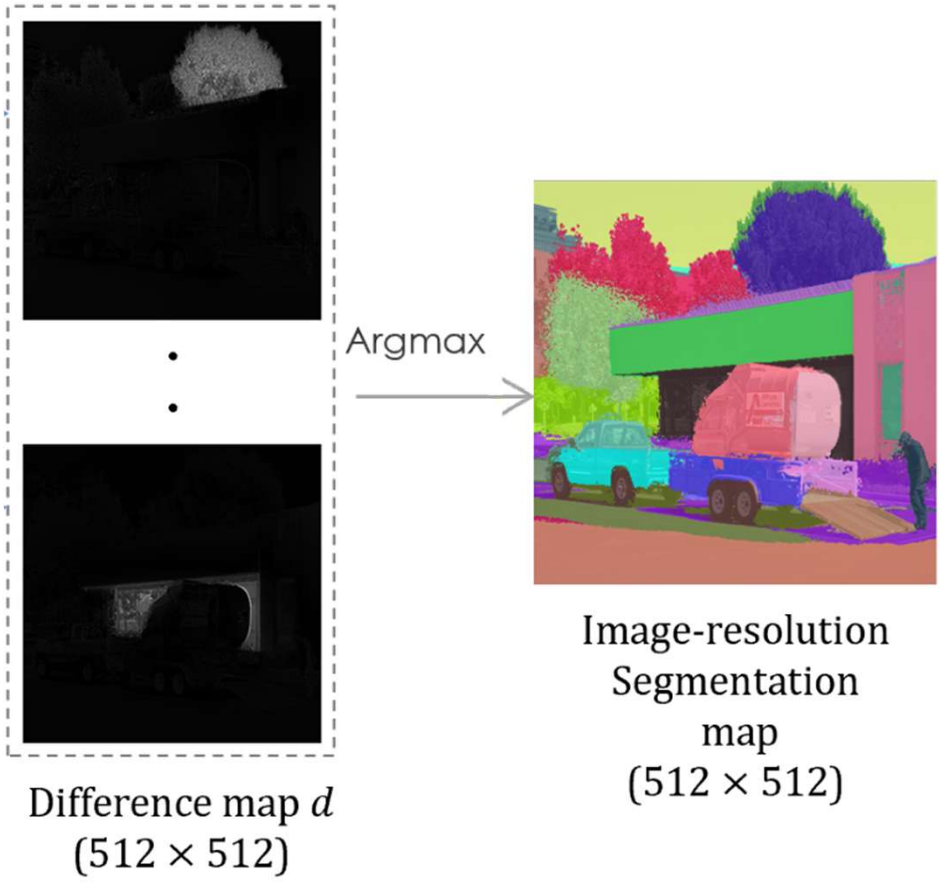
北京大学 PEKING UNIVERSITY

EmerDiff

Every pixel (x, y) belongs to a specific mask k

$$k = \text{argmax}_i \, d^i_{x,y}$$

Thus we have a high-res. segmentation map



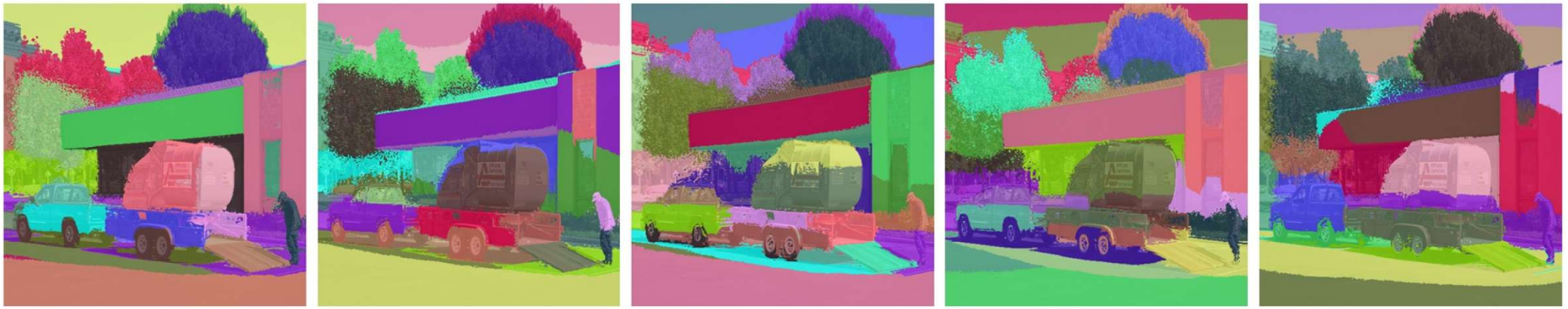
北京大学 PEKING UNIVERSITY

EmerDiff

Choice of timestep t matters

	mIoU (↑)				
Timestep	1	201	401	601	801
Unsupervised seg.	**33.1**	32.8	31.8	29.8	26.9
Open-vocabulary seg.	**15.9**	15.8	15.7	15.2	13.9

$t_f = 1$ $t_f = 201$ $t_f = 401$ $t_f = 601$ $t_f = 801$



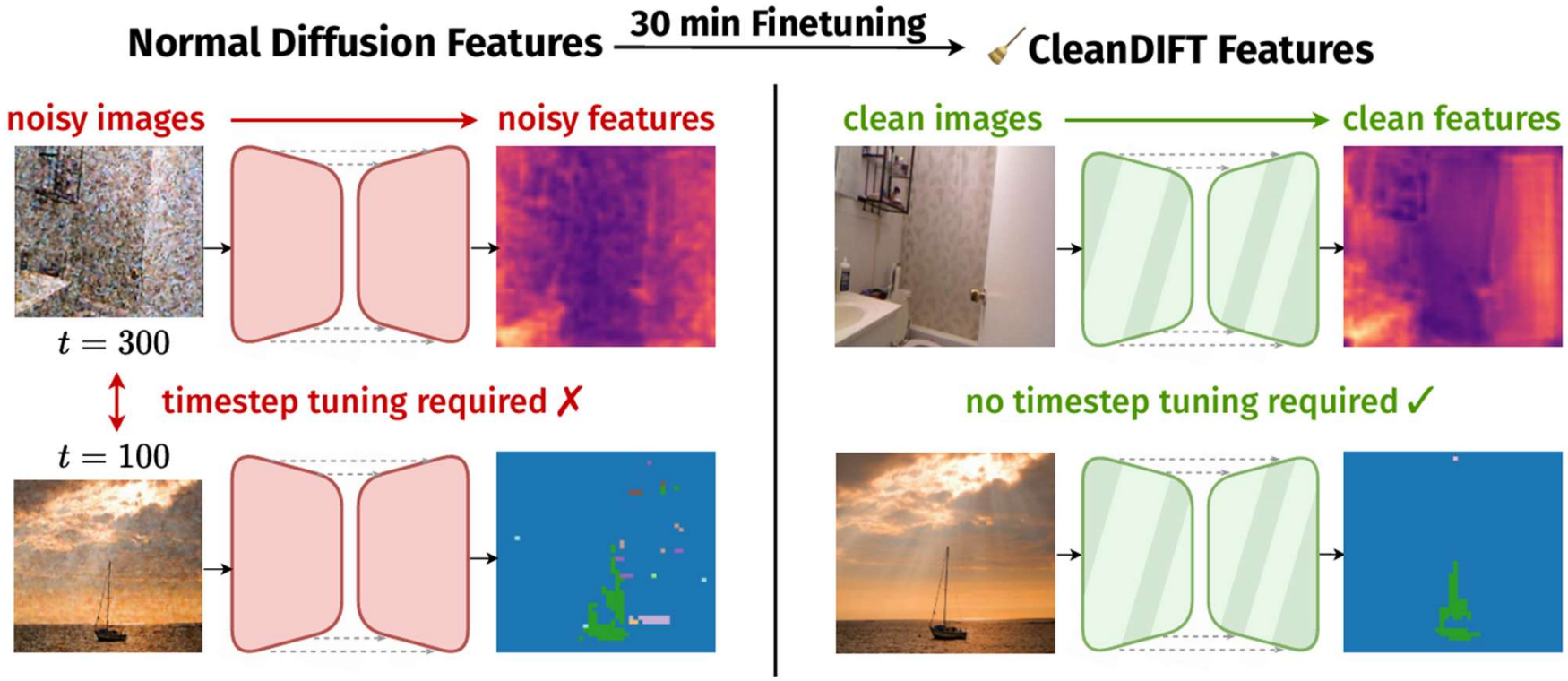
Existing Problem

- Dependence on diffusion timesteps
  - Different tasks may have different optimal timesteps

- Information loss in noisy images
  - Adding noise to clean images bottlenecks the perceptual information

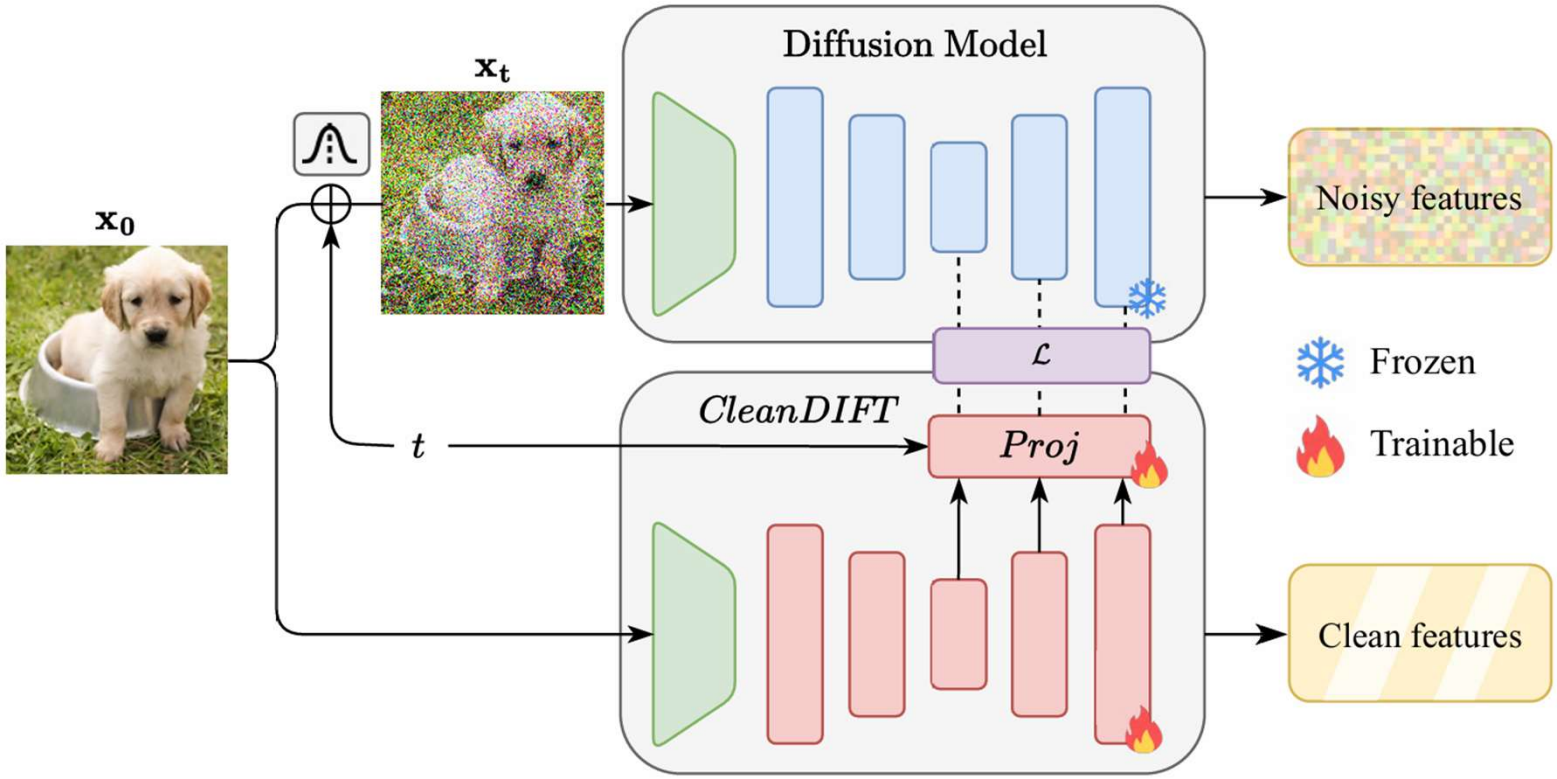
Method

CleanDIFT: noise-free, timestep-independent features



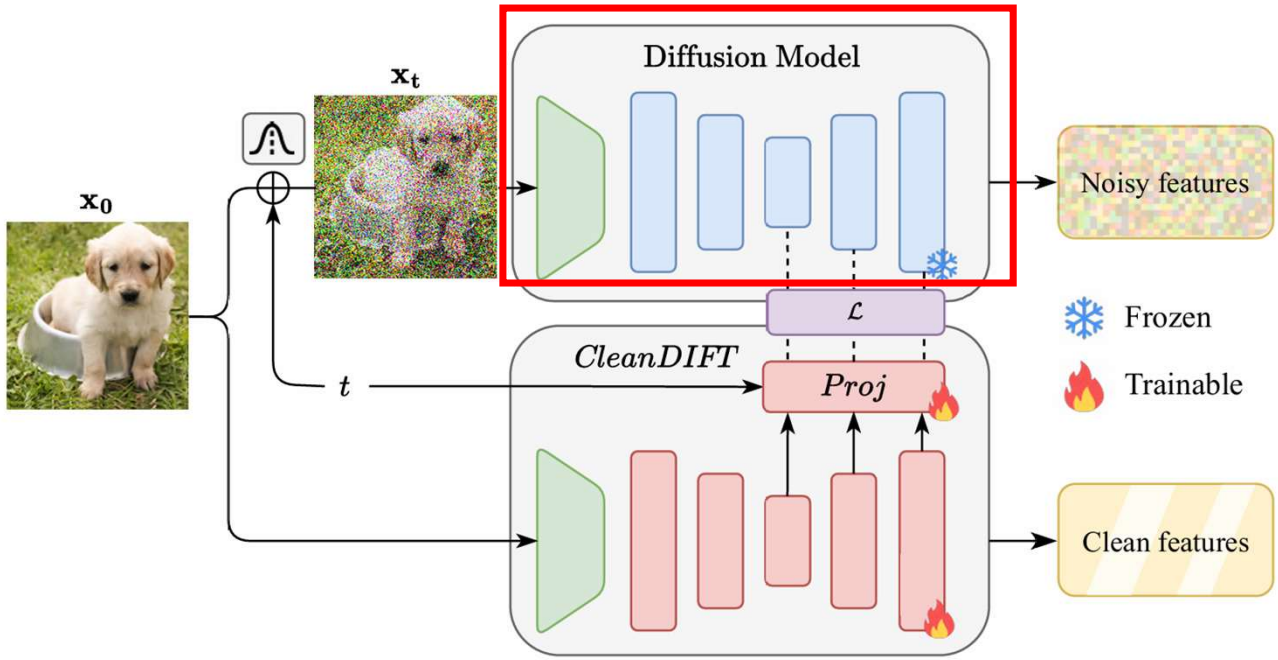
Method

Overview:



Frozen Diffusion Model

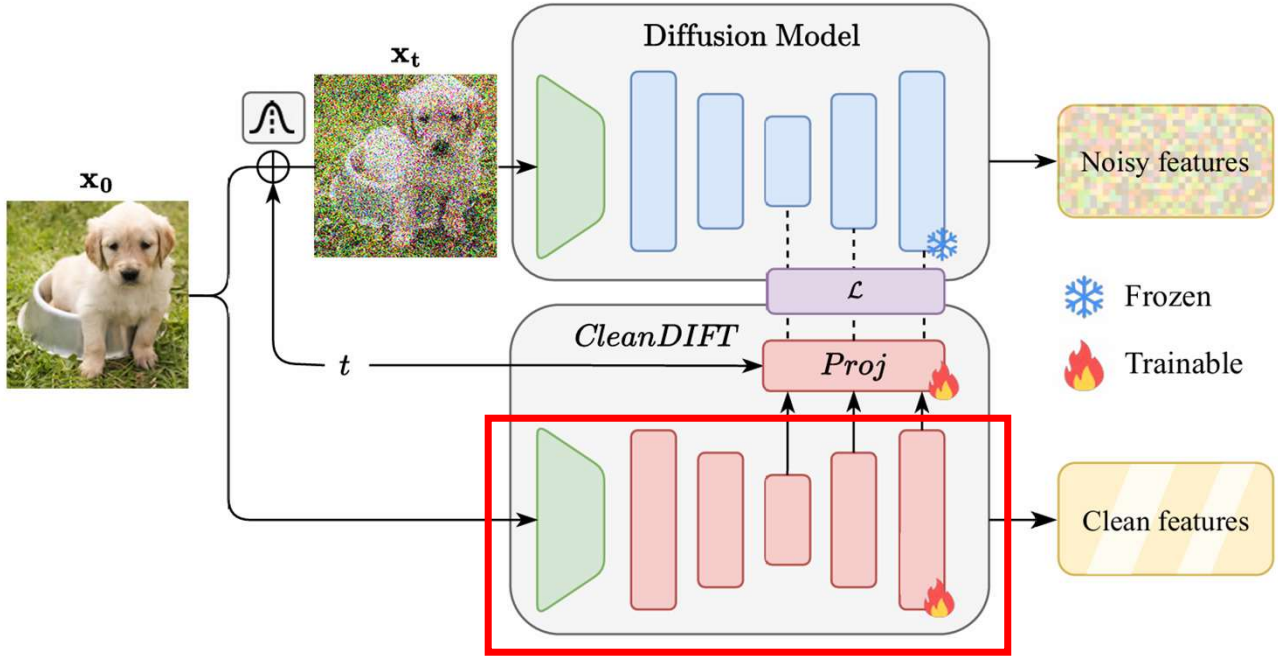
Frozen diffusion model receives the **noisy image** and **timestep** t



CleanDIFT

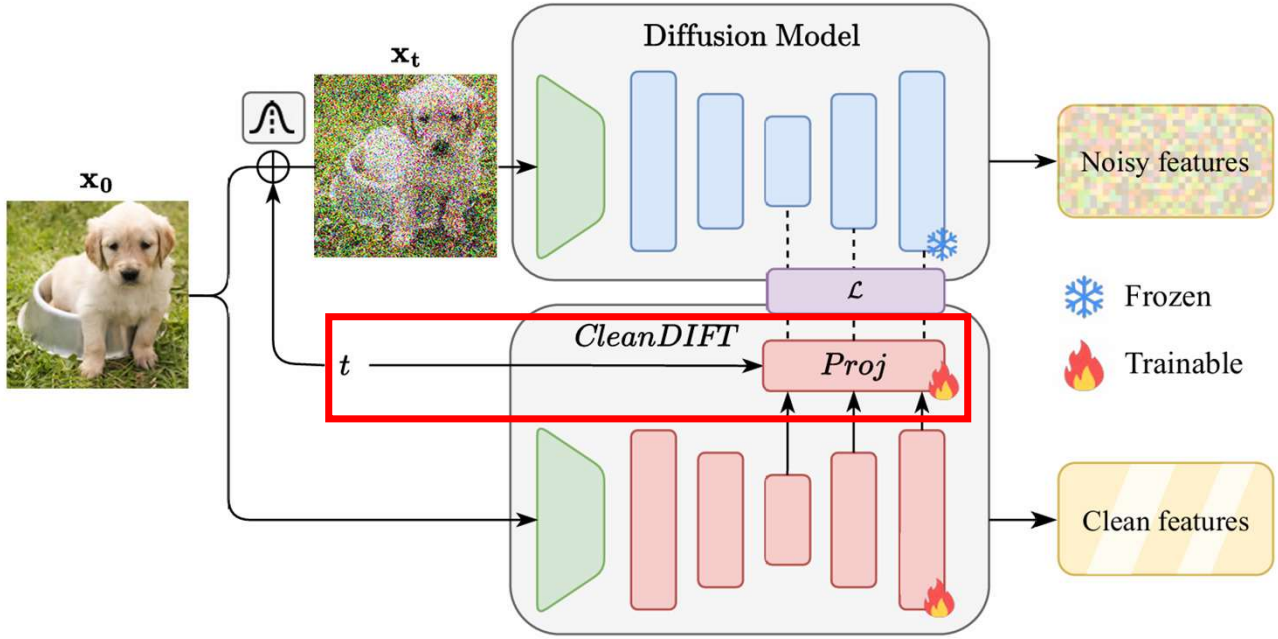
Initialize the CleanDIFT model as a trainable copy of the diffusion model

CleanDIFT only receives clean input image



Projection Heads

Projection heads receive **clean diffusion features** and **timestep *t***



Training Objective

$$\text{Stable Diffusion}$$

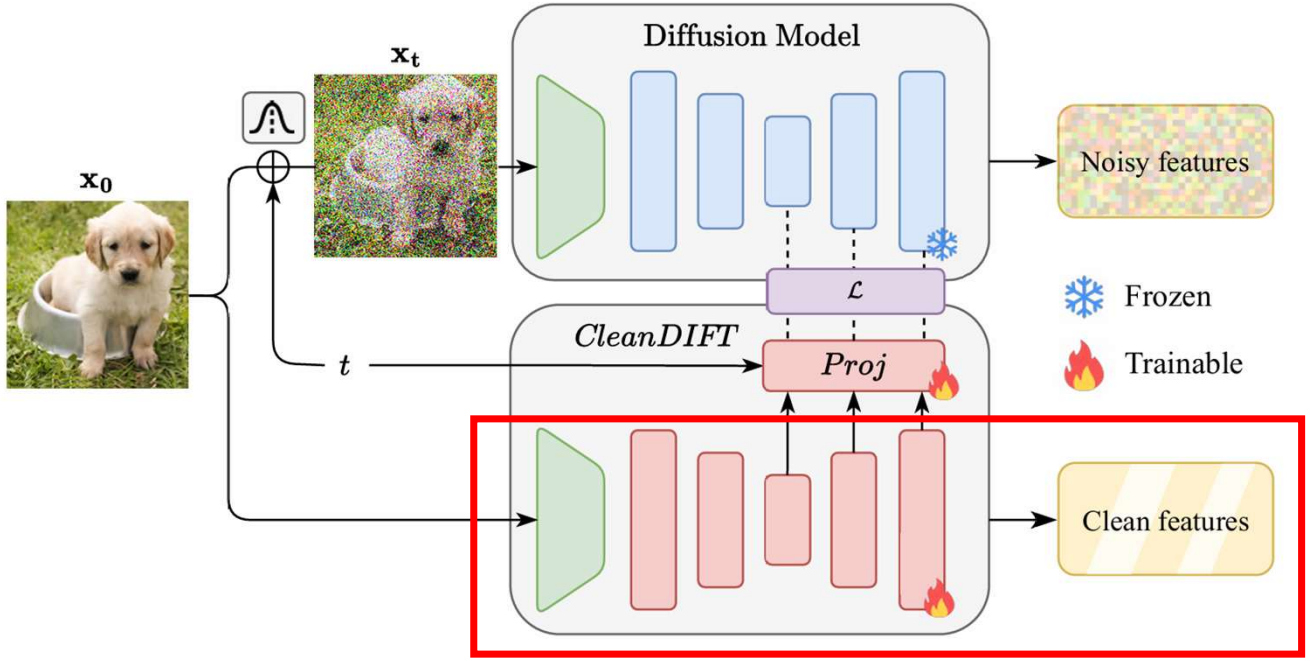
$$\text{CleanDIFT}\quad \text{feat}_c(\mathbf{x}) \rightarrow \begin{cases} \text{proj}(\cdot, t = 1) \longrightarrow \mathcal{L} \longleftarrow \text{feat}(\mathbf{x}, \boldsymbol{\epsilon}, t = 1) \\ \text{proj}(\cdot, t = 2) \longrightarrow \mathcal{L} \longleftarrow \text{feat}(\mathbf{x}, \boldsymbol{\epsilon}, t = 2) \\ \vdots \\ \text{proj}(\cdot, t = 999) \longrightarrow \mathcal{L} \longleftarrow \text{feat}(\mathbf{x}, \boldsymbol{\epsilon}, t = 999) \end{cases}$$

$\mathcal{L}$ is cosine similarity loss

$$\mathcal{L} = -\sum_{k=1}^{K} \text{sim}(\text{proj}^{(k)}(\text{feat}_c^{(k)}(\mathbf{x}_0); t), \text{feat}^{(k)}(\mathbf{x}_t; t))$$

Inference

We can discard the projection heads and use CleanDIFT directly



Experiments

CleanDIFT can be used for many downstream tasks

- Unsupervised Semantic Correspondence
- Depth Estimation
- Semantic Segmentation
- Classification

北京大学 PEKING UNIVERSITY

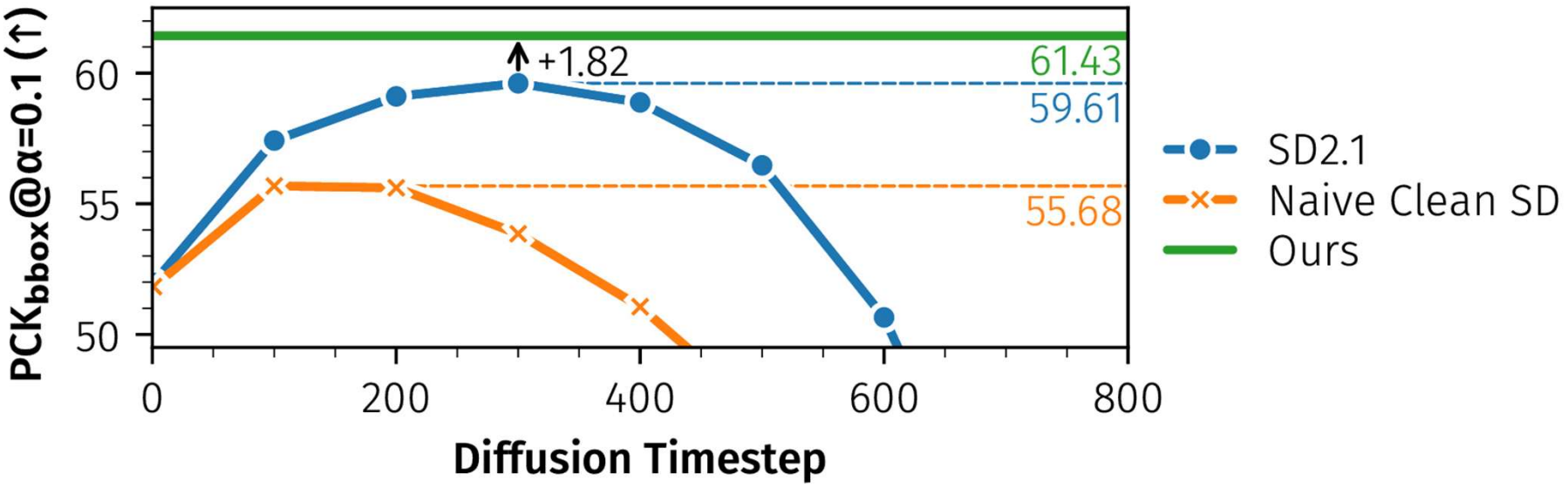
Unsupervised Semantic Correspondence

DIFT

CleanDIFT

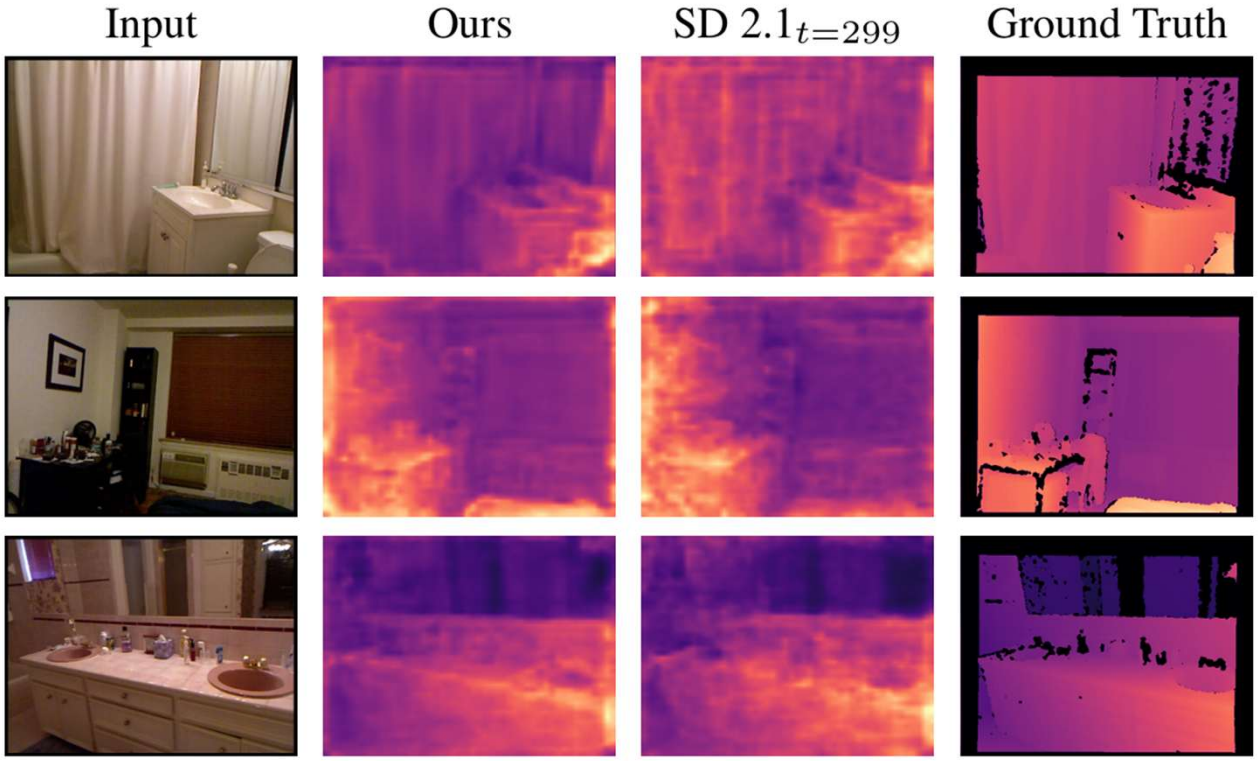
Unsupervised Semantic Correspondence

CleanDIFT outperforms DIFT even at its best timestep



Depth Estimation

Use the linear probe for depth estimation



北京大学 PEKING UNIVERSITY

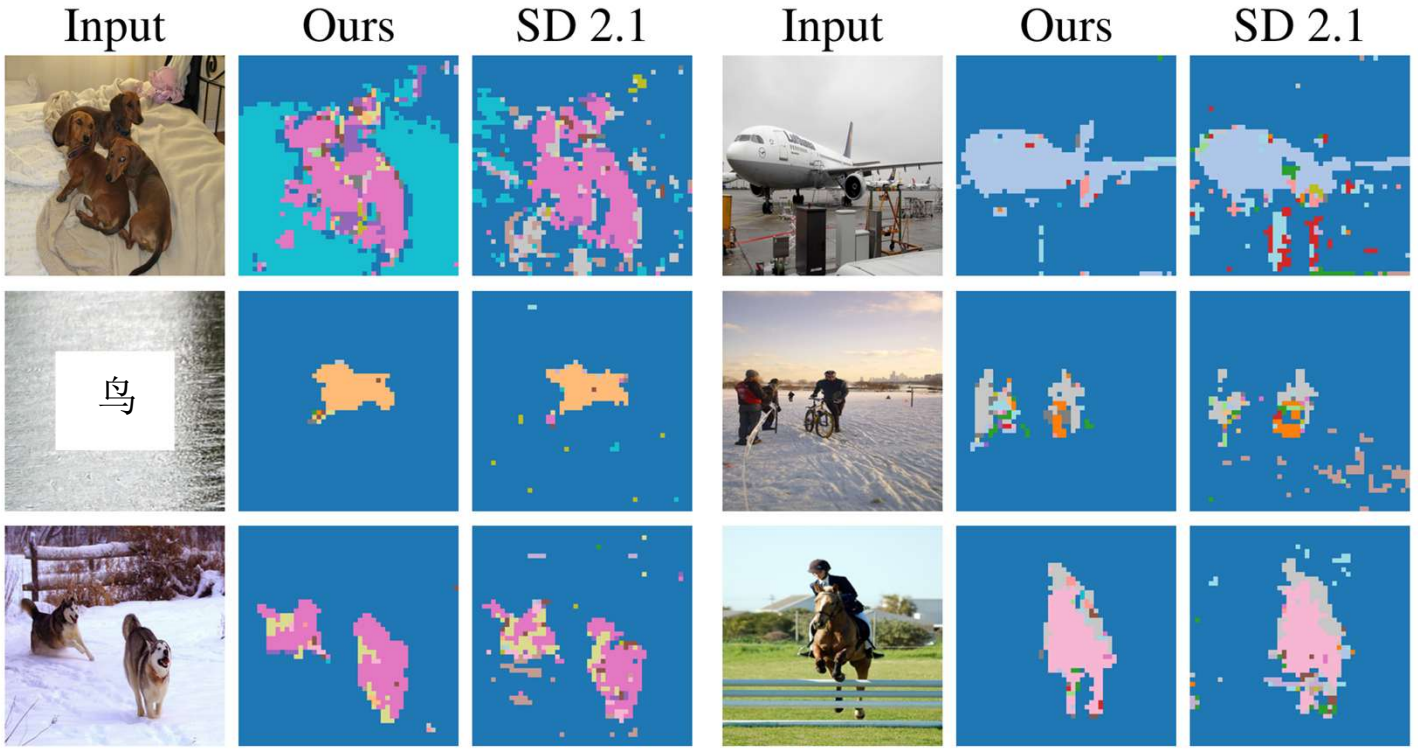
Depth Estimation

Method	Backbone	RMSE (↓)
Self-Supervised Methods		
OpenCLIP [28]	ViT-G/14	0.541
MAE [20]	ViT-H/14	0.517
DINO [6]	ViT-B/8	0.555
iBOT [68]	ViT-L/16	0.417
DINOv2 [43]	ViT-g/14	**0.344**
Diffusion Features		
DIFT-like [62]	SD 2.1 [51]	0.469
	Ours	**0.444** ▼0.025
	+ Probes from noisy features	0.453 ▼0.016

Experiments

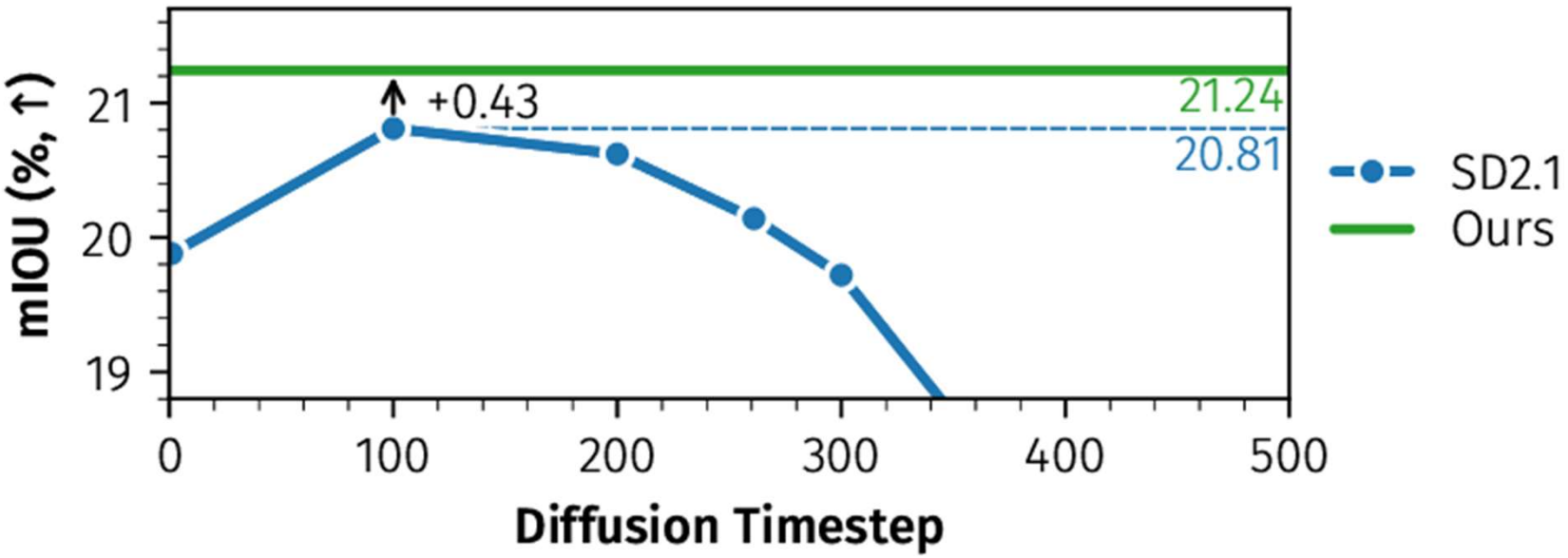
Semantic Segmentation

Use the linear probe for semantic segmentation



Semantic Segmentation

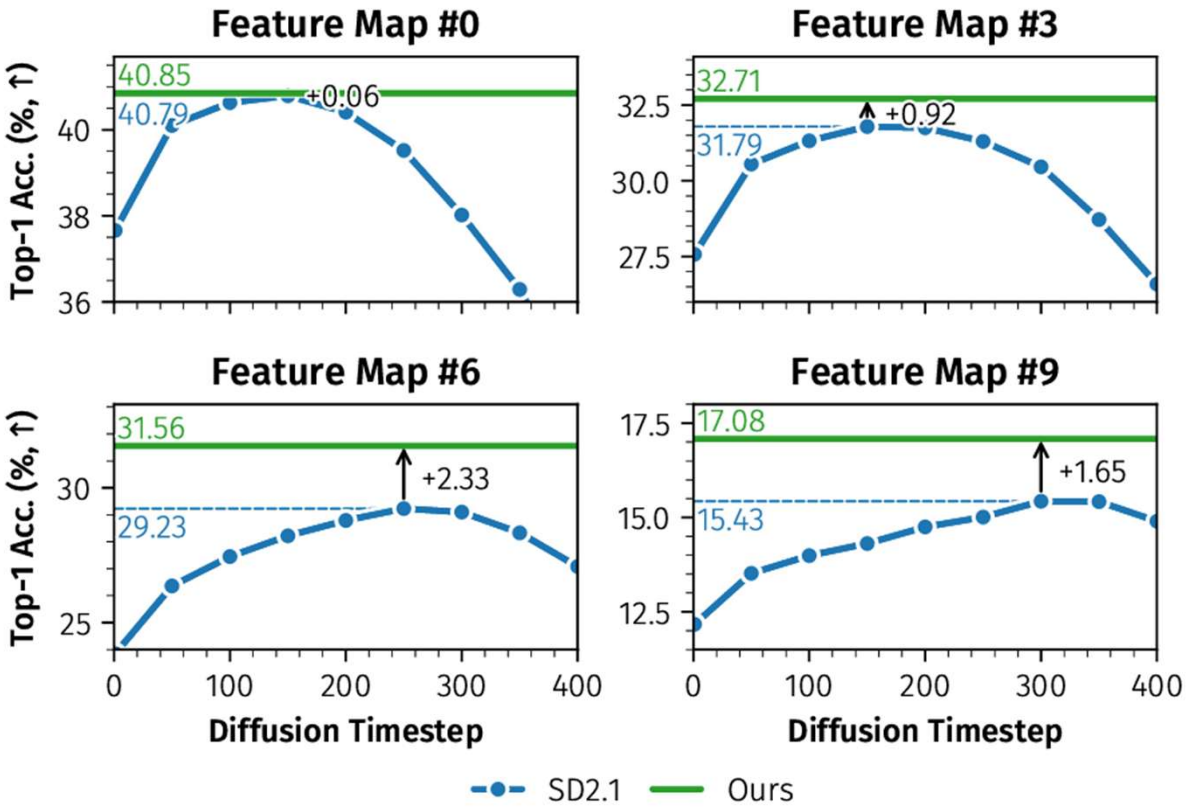
CleanDIFT outperforms DIFT even at its best timestep



Classification

CleanDIFT outperforms DIFT

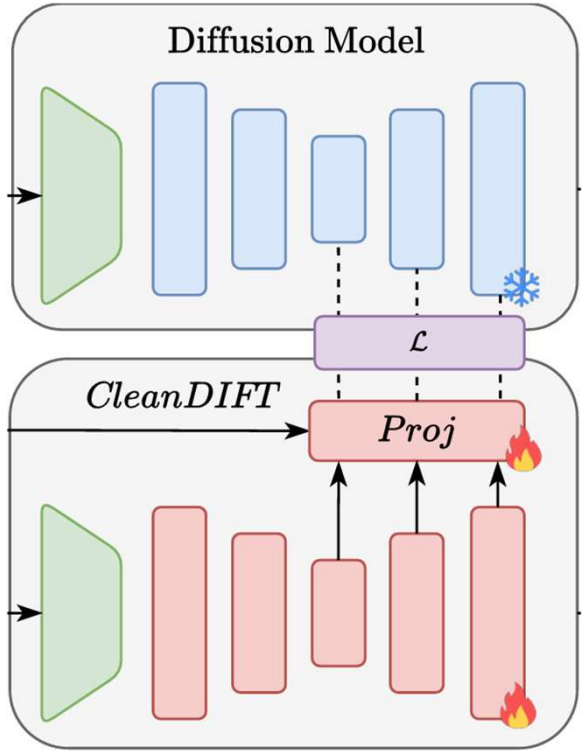
- Feature Map #0 (UpBlock1)
- Feature Map #3 (UpBlock2)
- Feature Map #6 (UpBlock3)
- Feature Map #9 (UpBlock4)



Ablation Studies

Cosine similarity with projection heads performs best

Objective	Projection Heads	PCK@α (↑)	
		$\alpha_{\text{img}} = 0.1$	$\alpha_{\text{bbox}} = 0.1$
Cosine Sim.	✓	**68.32**	**61.43**
	✗	<u>68.16</u>	<u>61.29</u>
L_2	✓	66.23	59.13
	✗	66.49	59.43
L_1	✓	66.91	60.00
	✗	66.87	59.91
SD 2.1	-	63.41	55.92



Other Diffusion Backbones

Still effective for other diffusion backbones (larger U-Net or DiT)

Backbone	PCK@α Gain ($\uparrow$)	
	$\alpha_{\text{img}} = 0.1$	$\alpha_{\text{bbox}} = 0.1$
SDXL [46]	1.7	1.6
SDXL Turbo [55]	3.2	3.7
PIXART-α [7]	2.7	2.2
Flux [14]	9.4	8.1

Conclusion

北京大学 PEKING UNIVERSITY

- Diffusion features contain both semantic information and noise
- Clean diffusion features outperform noisy diffusion features
- CleanDIFT are useful for many downstream tasks

Disscusion

Future Work

- Multi-layer Feature Aggregation
- Clean DiT features
- DiT features for downstream tasks

Thanks for listening!

Presenter: Yixuan Zou
2026.3.15