STELLAR: Learning Sparse Visual Representations via Spatial-Semantic Factorization

ICML 2026

Theodore Zhengde Zhao, Sid Kiblawi, Jianwei Yang, Naoto Usuyama, Reuben Tan, Noel C Codella, Tristan Naumann, Hoifung Poon, Mu Wei

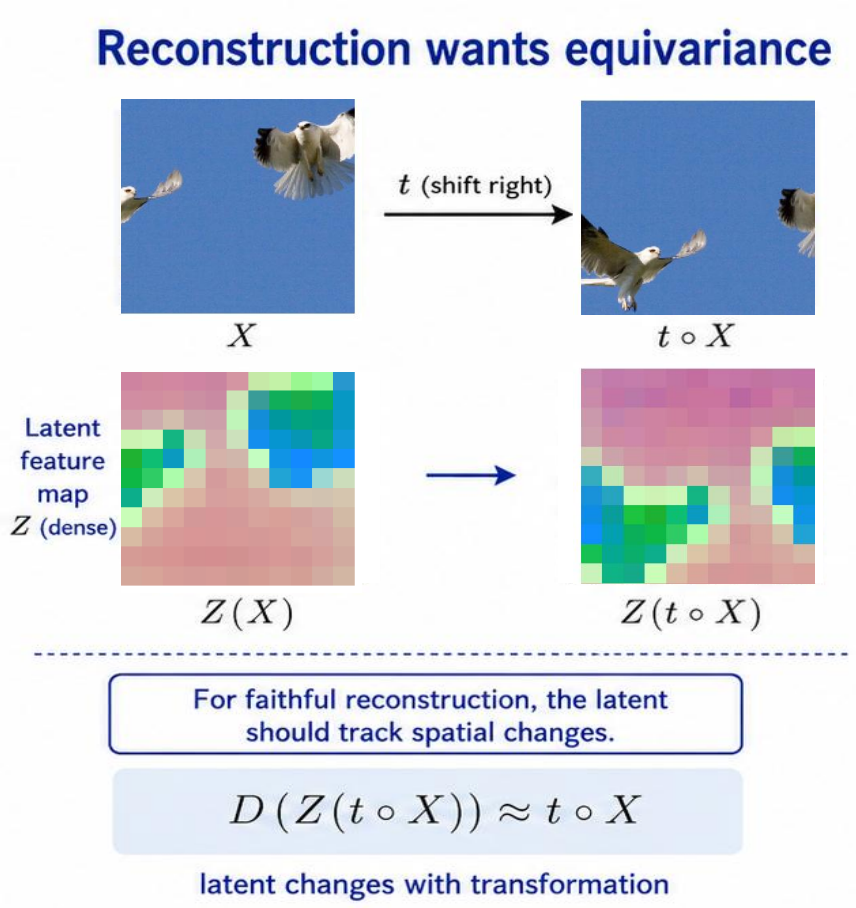
Microsoft

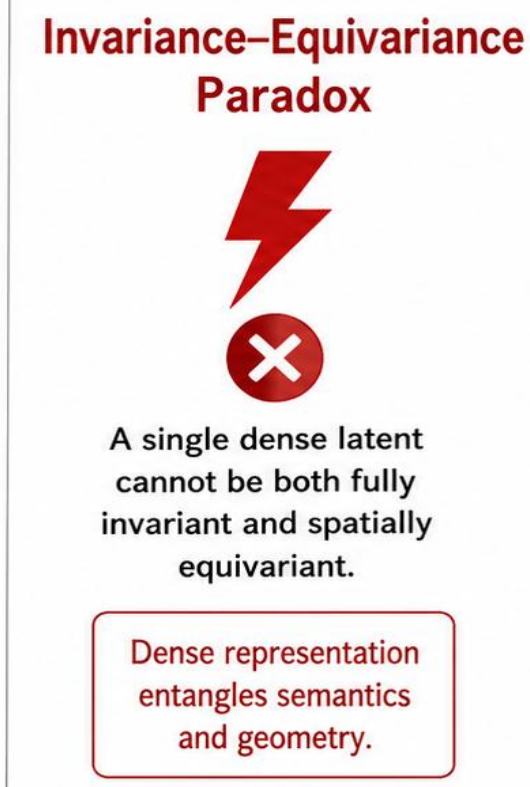
Research Question

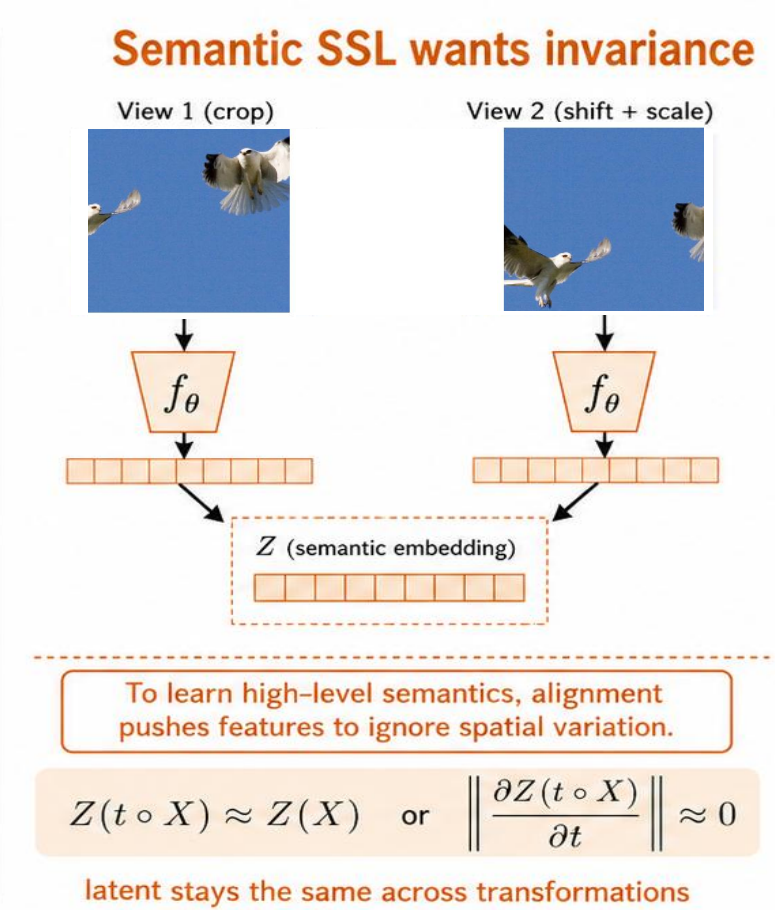
- How to learn "*good*" visual representation from image data
  - Rich abstract semantics for understanding tasks
  - Contains sufficient information to recover the reality (reconstruction)
- AE, VAE, MAE, DINO, Diffusion, SimCLR, BYOL…
  - Reconstruction-centric SSL: preserve details, but often weaker semantics
  - Semantic-centric SSL: strong semantics from invariance (joint-embedding), abandoning pixel reconstruction
- Take a step back:
  - 1. What's a "good" visual representation…
  - 2. … and how to learn it from image data

A Thought Experiment...

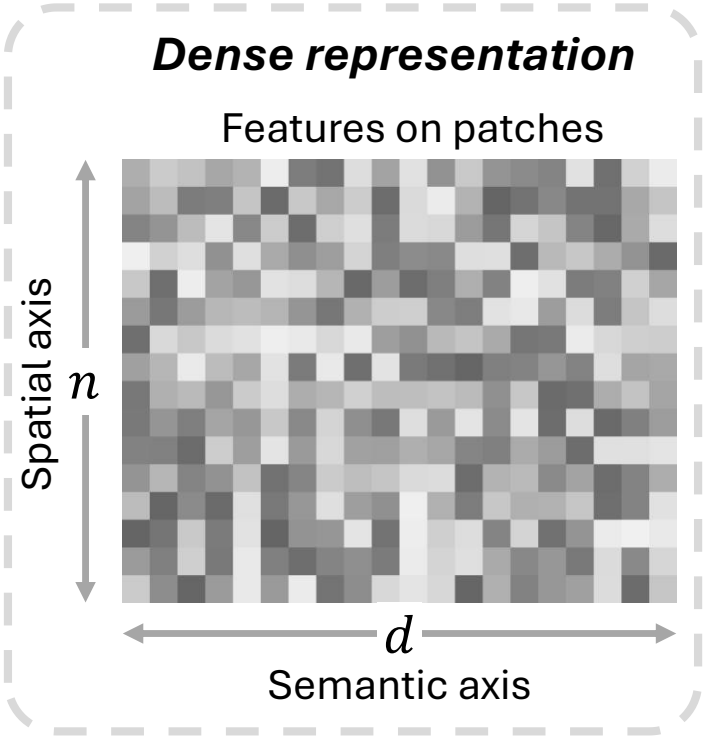
How should the latent representation change under transformation?



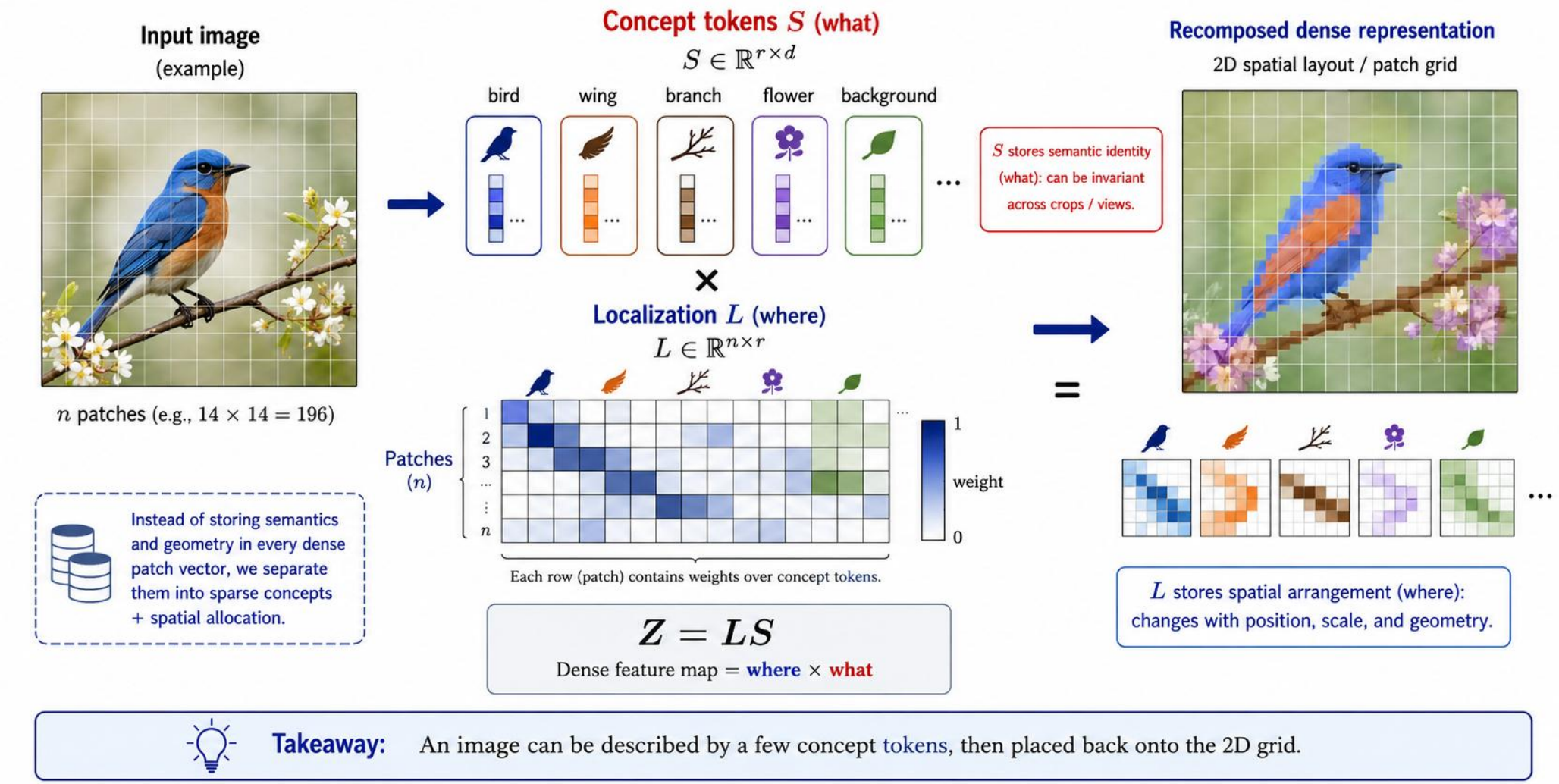




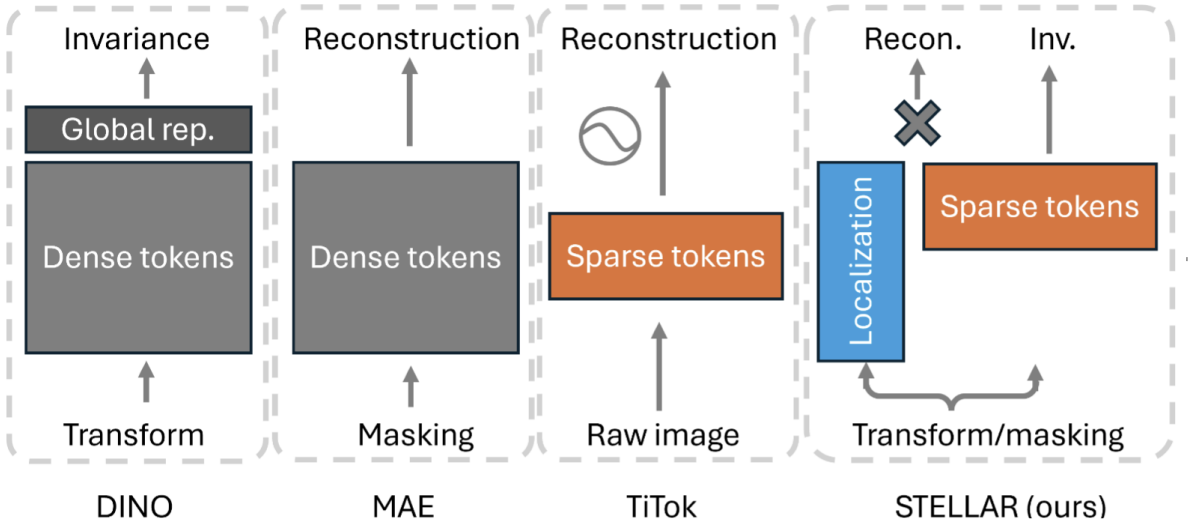
Spatial-Semantics Factorization



Modeling Image with Sparse Concepts

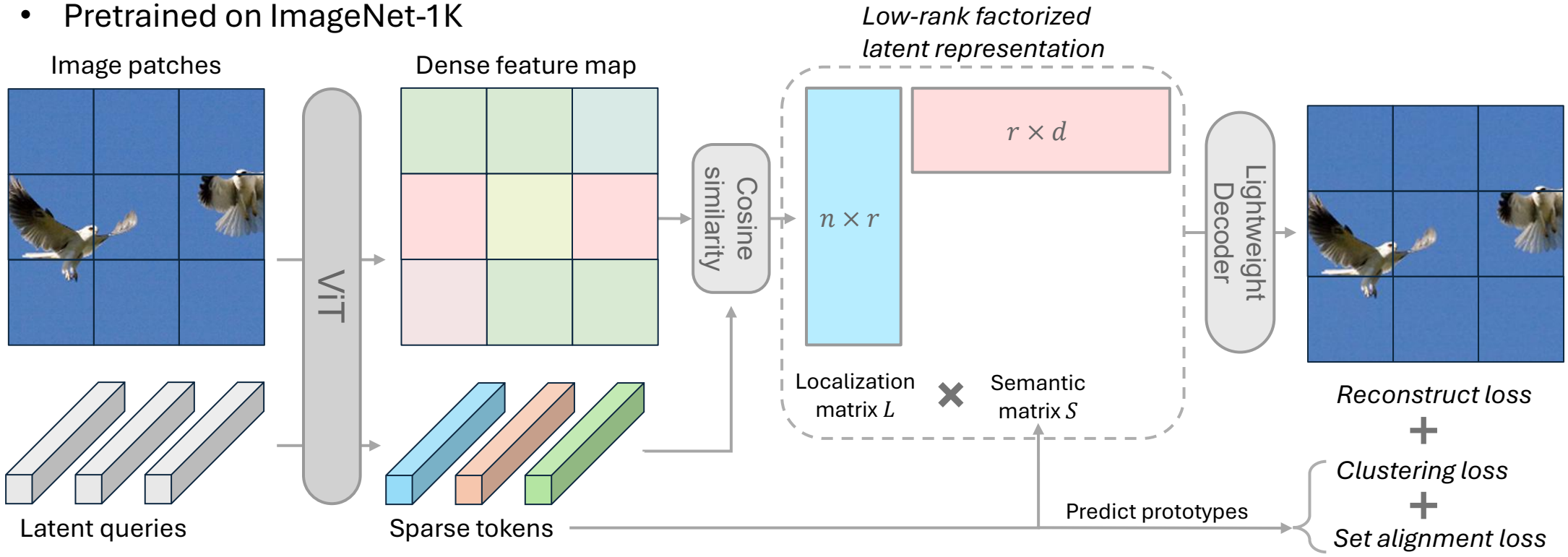


Jointly Self-supervised Learning



Simple Architecture

- Encoder: standard ViT + learnable queries
- Spatial localization from cosine similarity
- Decoder: 6-layer ViT predicting VQ-GAN code
- Pretrained on ImageNet-1K

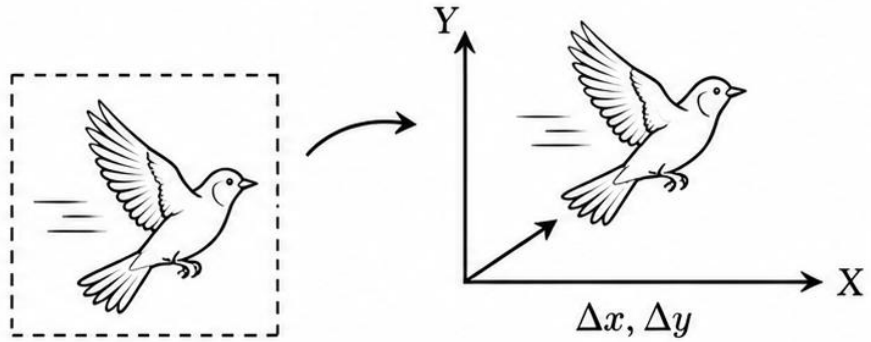


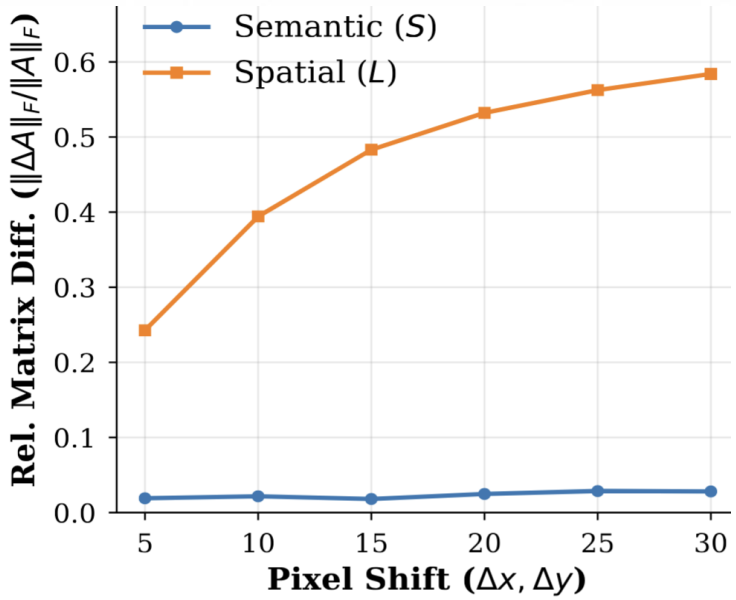
Exp: Equivariance Partition

- Measure representation change under spatial transformation

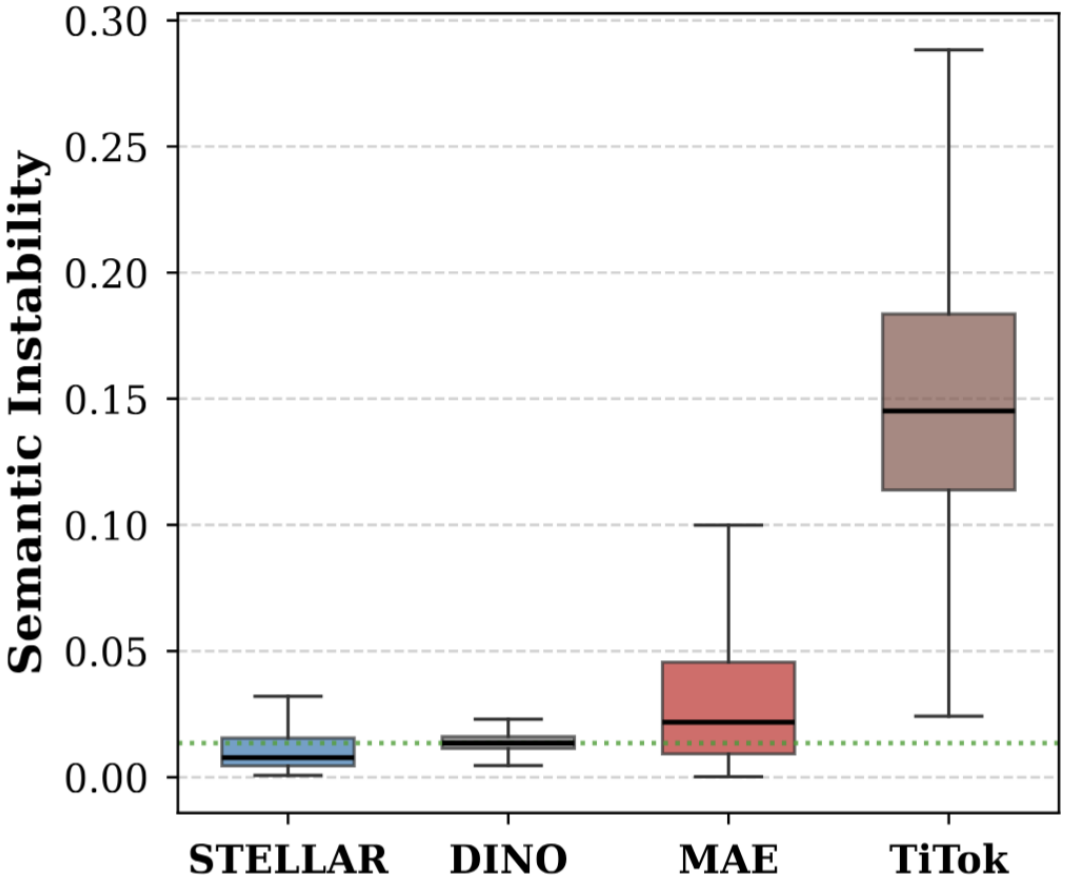
$$\frac{\partial Z}{\partial \theta} = \underbrace{\left(\frac{\partial L}{\partial \theta}\right) S}_{\text{Spatial Equivariance}} + \underbrace{L\left(\frac{\partial S}{\partial \theta}\right)}_{\text{Semantic Variance} \approx 0}$$

By disentangling spatial and semantic information, we offload spatial equivariance entirely to the localization matrix (**L**), while the semantic tokens (**S**) remain invariant.





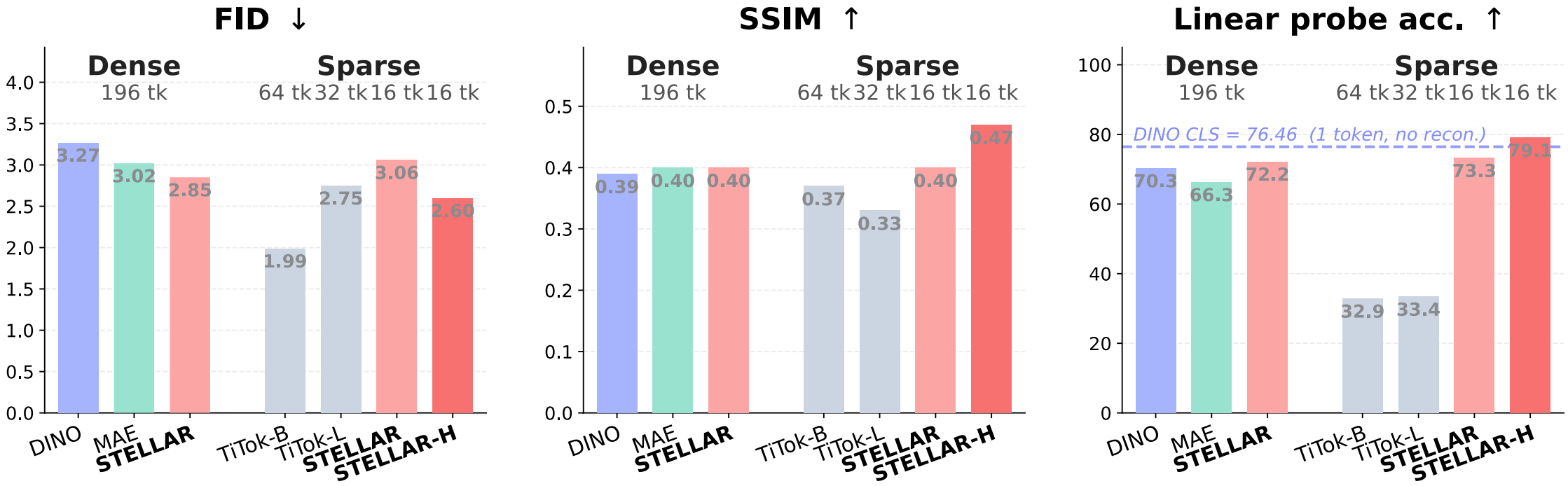
Exp: Semantic Invariance



- Perform random crop and measure change in the latent representation in terms of cosine distance
- STELLAR has DINO-level semantic invariance, while other reconstruction methods have higher variance

Exp: Unified Representation

- STELLAR achieves better reconstruction and spatial consistency than other dense and sparse representation
- Linear probing Acc outperformed other reconstruction feasible representations



Exp: Downstream Transfer

- Top dense prediction performance with linear probing
- Global semantics outperformed reconstruction SSL methods

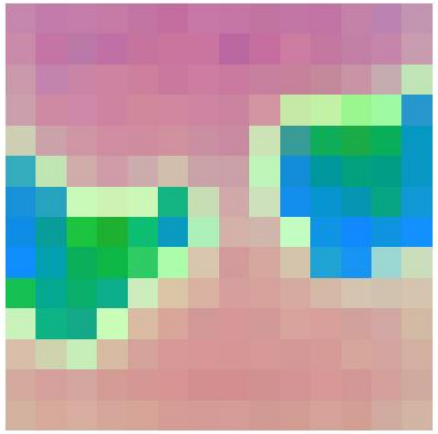
Model	Arch.	SSL Type		Segmentation (mIoU)			Classification (Acc)			
		Target	Method	ADE20K	CitySc	VOC	IN1K	Pets	Food	GlaS
Semantic-Centric (Joint Embedding / Invariance)										
BYOL	RN-50	GLOBAL	DISTILL	18.43	18.66	63.89	70.39	82.77	64.57	95.00
MOCO V3	VIT-B	GLOBAL	CONTR.	29.45	25.13	74.08	74.31	91.14	77.47	**97.50**
DINO	VIT-B	GLOBAL	DISTILL	26.87	26.82	79.29	76.46	**93.84**	79.28	95.00
MSN	VIT-B	GLOBAL	MASKING	26.66	25.39	68.59	73.65	75.91	68.93	92.50
DENSECL	RN-50	DENSE	CONTR.	23.08	18.63	70.95	61.10	72.99	59.16	85.00
DATA2VEC	VIT-B	DENSE	LAT-MIM	22.03	23.49	61.33	54.90	26.47	34.40	73.75
SIAMESEIM	VIT-B	DENSE	LAT-MIM	29.24	26.52	81.38	74.97	91.61	71.01	91.25
I-JEPA	VIT-H	DENSE	LAT-MIM	21.57	18.59	74.13	71.72	84.68	70.34	87.50
IBOT	VIT-B	GL+DE	DIST+MIM	31.78	25.69	77.06	76.40	92.40	78.08	96.25
IBOT	VIT-L	GL+DE	DIST+MIM	33.26	26.37	77.57	78.53	92.12	**81.07**	96.25
Image-Centric (Reconstruction)										
BEIT	VIT-B	DENSE	TOK MIM	11.58	18.90	27.44	32.94	36.20	54.49	90.00
BEIT	VIT-L	DENSE	TOK MIM	12.64	20.37	25.48	36.77	36.71	56.03	90.00
SIMMIM	SWIN-B	DENSE	PIX MIM	12.46	17.23	35.14	24.77	27.39	40.94	77.50
MAE	VIT-B	DENSE	PIX MIM	30.91	29.44	76.43	66.32	81.58	70.40	93.75
MAE	VIT-L	DENSE	PIX MIM	34.36	32.53	77.79	73.09	84.30	76.22	95.00
MAE	VIT-H	DENSE	PIX MIM	36.16	**35.21**	78.07	75.22	84.96	78.36	95.00
SEMMAE	VIT-B	DENSE	PIX MIM	3.52	25.48	48.33	43.84	56.99	58.90	92.50
TITOK-64	VIT-B	SPARSE	SPRS REC	–	–	–	32.87	42.06	43.68	**97.50**
TITOK-32	VIT-L	SPARSE	SPRS REC	–	–	–	33.42	27.83	38.83	78.75
Our Method (Sparse Factorized Modeling)										
STELLAR	VIT-B	SPARSE	INV+REC	31.33	27.74	81.83	73.26	89.70	74.09	95.00
STELLAR	VIT-L	SPARSE	INV+REC	34.02	31.32	**85.90**	76.94	92.53	74.78	**97.50**
STELLAR	VIT-H	SPARSE	INV+REC	**36.66**	33.30	85.66	**79.10**	92.53	77.43	92.50

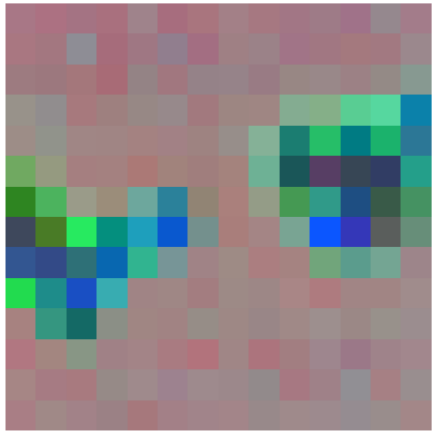
Visualization

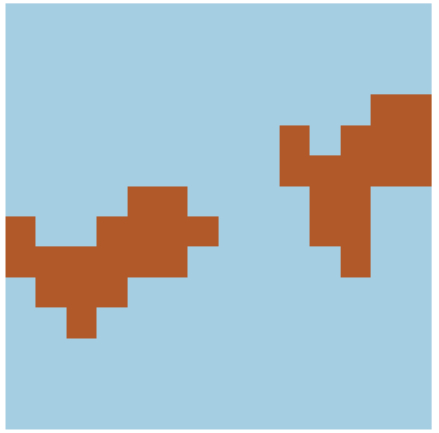
Feature map

Low-rank approx.

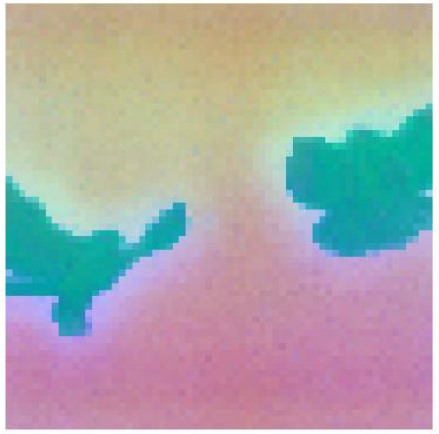
Example concept localization (threshold 1/r)

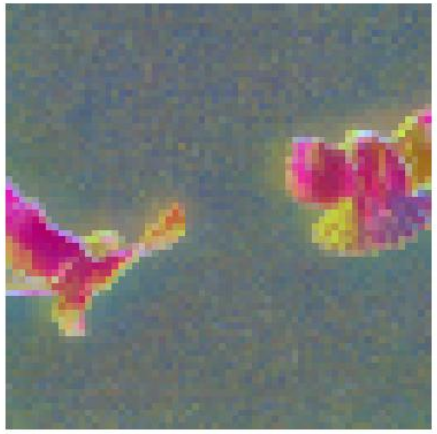


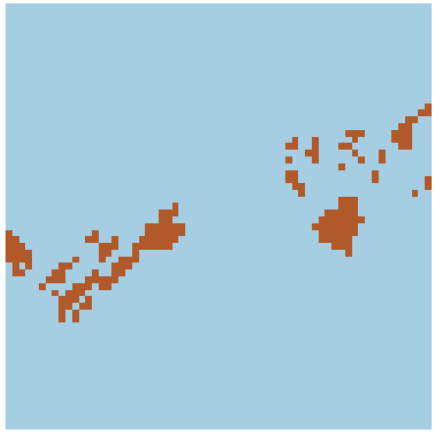


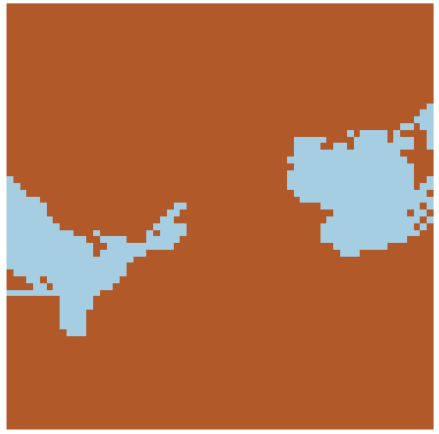


Directly scaling up res. to 1024 at inference



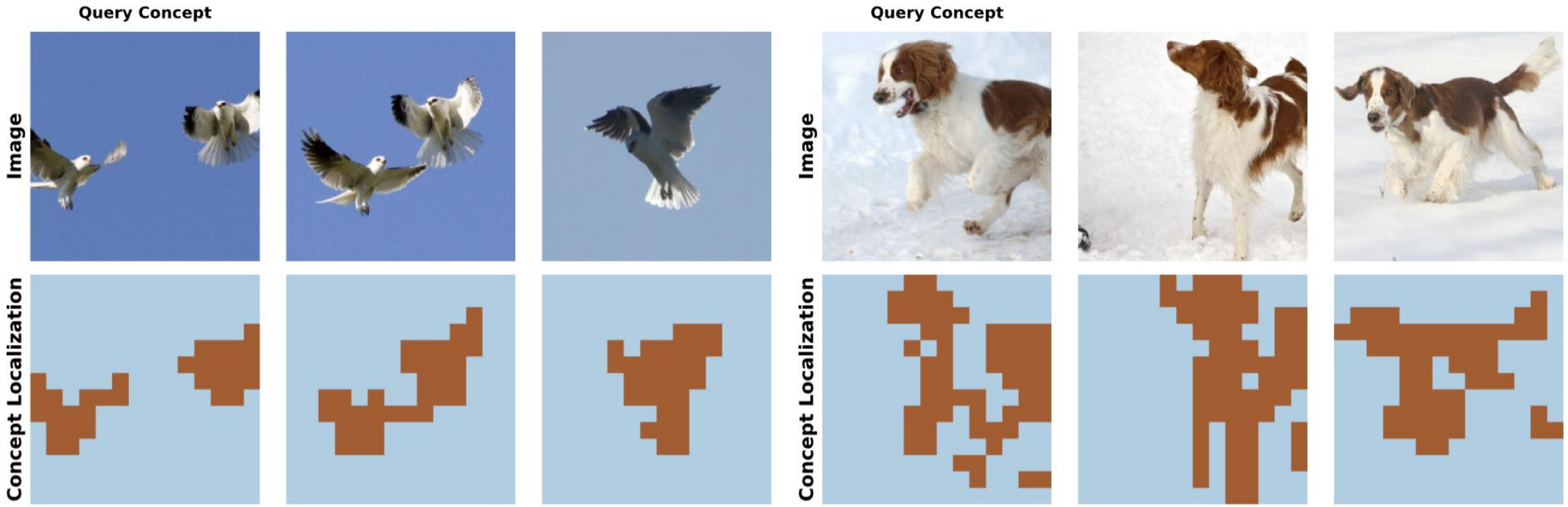






Concept Retrieval

- Retrieve the closest tokens from the query concept, and visualize the image and the localization of the token in the image



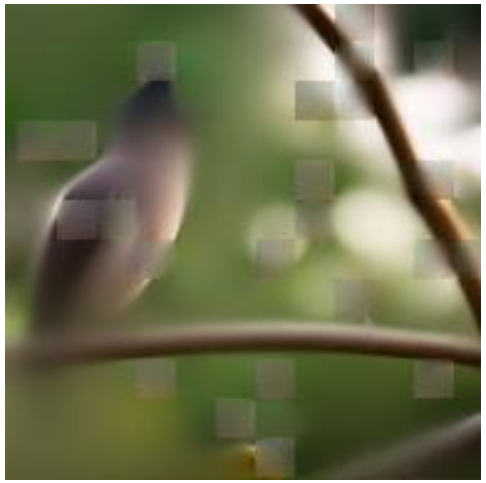
Sparse Reconstruction

Original

STELLAR 16 tokens

TiTok 32 tokens

MAE 32 tokens

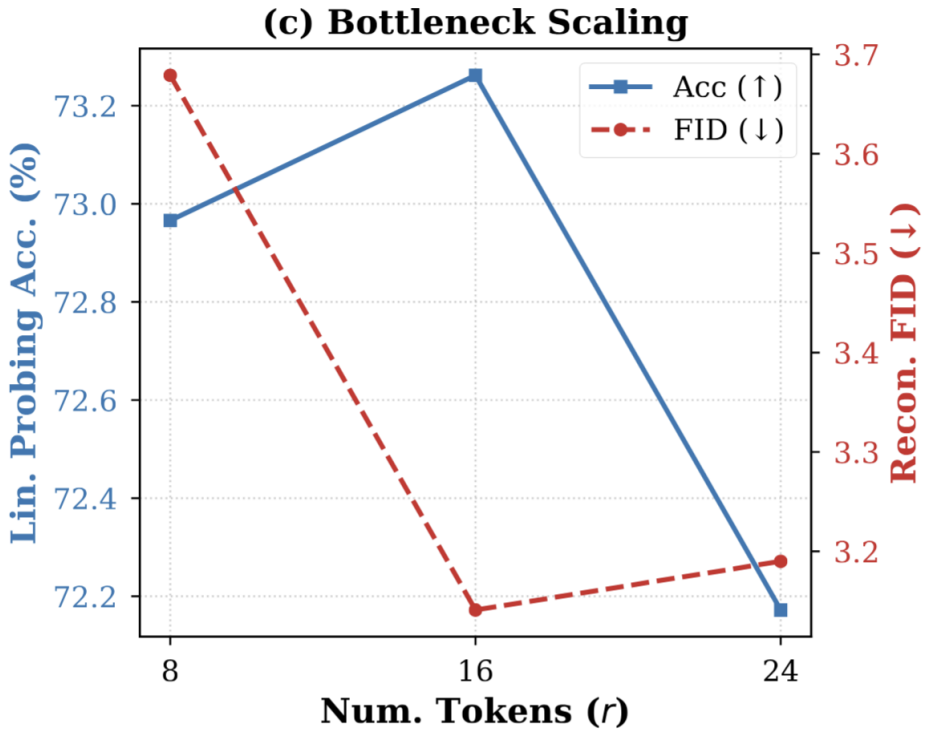


Ablation Studies

	Recon.	Cluster	Set Align	CLS Align	KoLeo	rFID ↓	IN1K ↑	ADE ↑
DEFAULT	✓	✓	✓	✓	✓	**3.14**	**73.26**	**31.33**
Impact of Individual Components								
(A)	✗	✓	✓	✓	✓	—	72.44 (-0.82)	29.94 (-1.39)
(B)	✓	✗	✗	✗	✓	3.21 (+0.07)	52.07 (-21.19)	20.46 (-10.87)
(C)	✓	✓	✗	✗	✓	8.95 (+5.81)	2.73 (-70.53)	1.93 (-29.39)
(D)	✓	✗	✓	✓	✓	3.62 (+0.48)	42.14 (-31.12)	18.90 (-12.43)
(E)	✓	✓	✓	✗	✓	3.26 (+0.12)	70.79 (-2.47)	30.20 (-1.12)
(F)	✓	✓	✓	✓	✗	3.25 (+0.11)	72.05 (-1.21)	30.10 (-1.23)

Effect of Number of Tokens/Rank

- Semantic favors less tokens (highly concentrate)
- Reconstruction favors more tokens(expressiveness)



Takeaways

- Factorized latent supports joint learning of semantics and reconstruction
- Represent one image using **16 tokens** with **2.60 FID** and **79.1 lin. Acc**.
- Efficient latent representation (90% token reduction)

Thanks!

Q&A