RILS: Masked Visual Reconstruction In Language Semantic Space

Shusheng Yang[1,2] Yixiao Ge[2] Kun Yi[2] Dian Li[3]
Ying Shan[2] Xiaohu Qie Xinggang Wang[1]

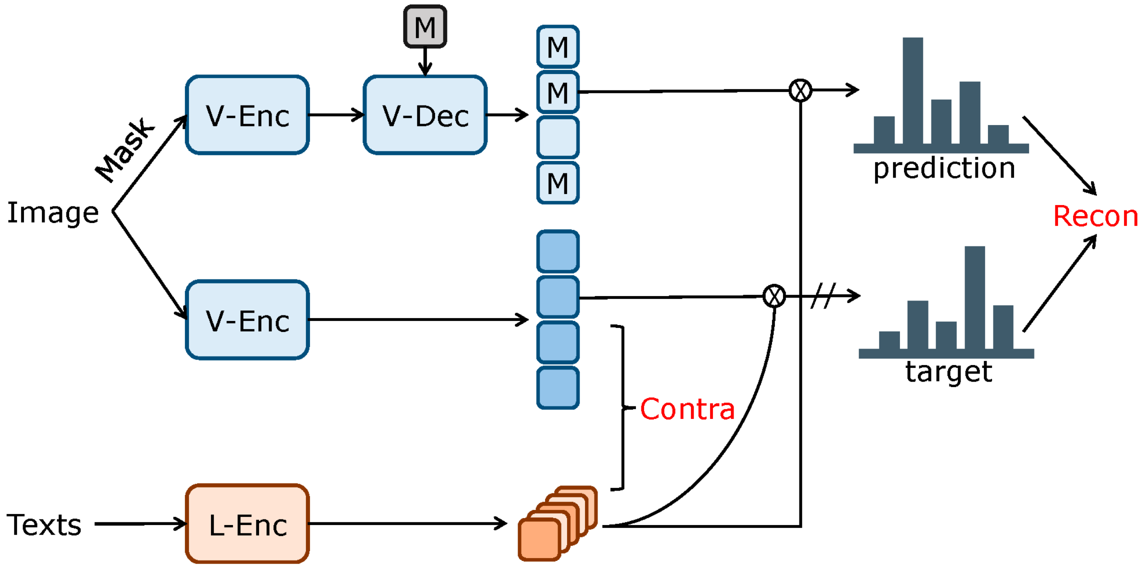
[1]School of EIC, Huazhong University of Science and Technology
[2]ARC Lab, Tencent PCG
[3]Foundation Technology Center, Tencent PCG

Paper Tag: THU-PM-258

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Quick Preview

- Better visual training by leveraging masked image modeling and image-text contrastive simultaneously
- A novel and effective pre-training method termed "Reconstruction in Language Space"
- Better transferability/zero-shot ability/few-shot ability on a wide range of downstream tasks.



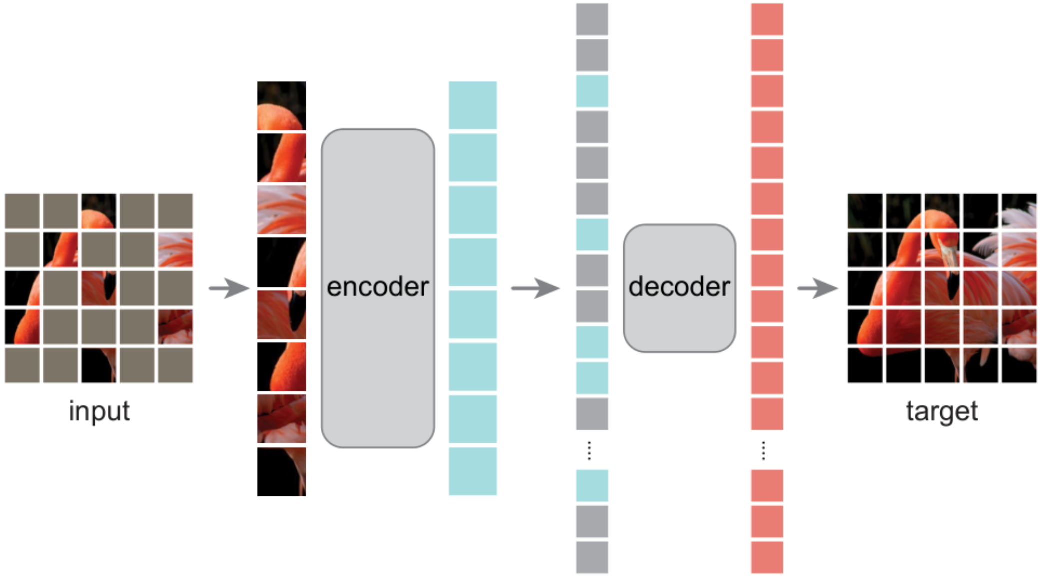
Overview of our RILS

Visual Representation Learning

- Masked Image Modeling
- Image-text Contrastive Learning

Masked Image Modeling (MIM)

- Random Mask → Reconstruct
- Fully self-supervised
- Fine-grained supervision
  - Transferability on downstream tasks

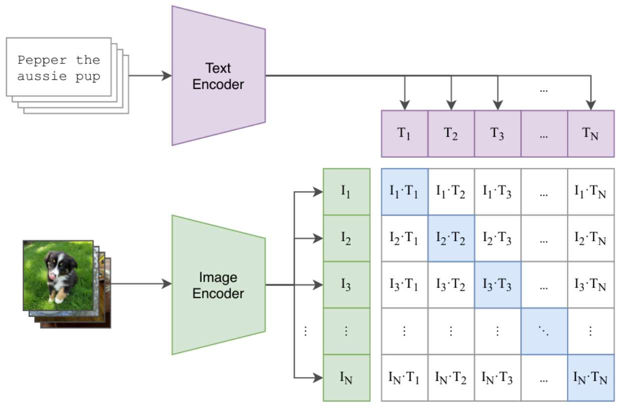


He, Kaiming, et al. [1]

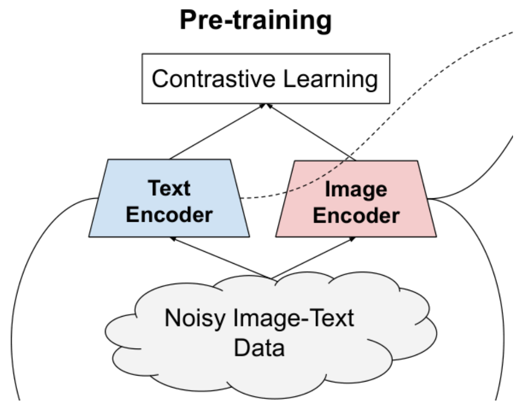
[1] Masked Autoencoders Are Scalable Vision Learners

Image-text Contrastive (ITC)

- Image-text pairs → Contrastive
- Image-text alignment
- Zero-shot Understanding
- Robustness



Radford, Alec, et al. [1]

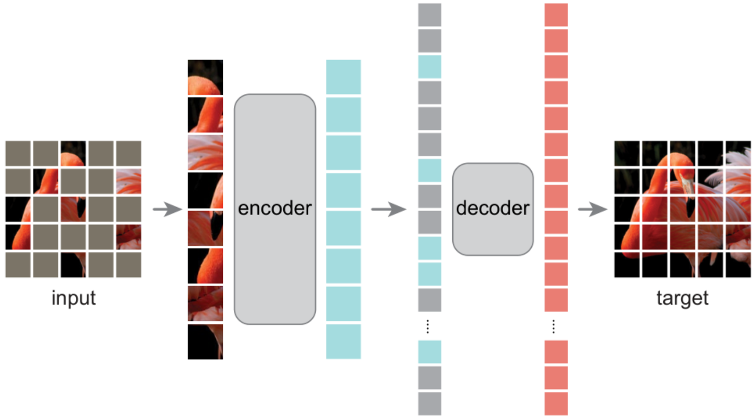


Jia, Chao, et al. [2]

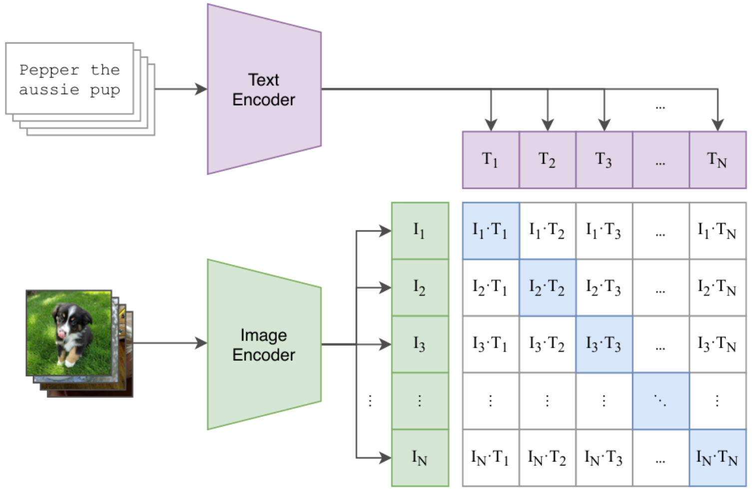
[1] Learning Transferable Visual Models From Natural Language Supervision
[2] Scaling up visual and vision-language representation learning with noisy text supervision

Motivation

MIM
&
ITC
→
Better Visual Pre-training
?



He, Kaiming, et al. [1]



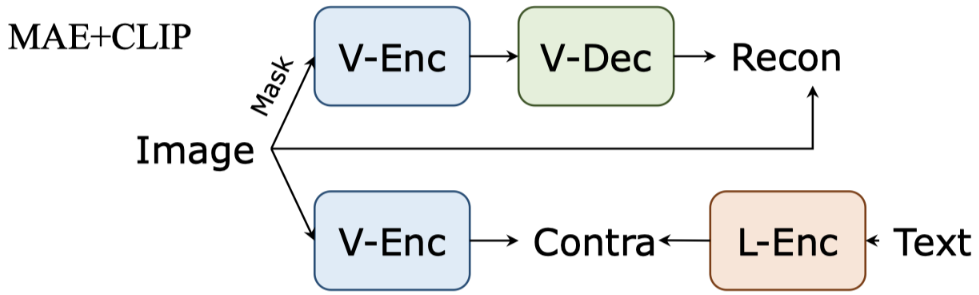
Radford, Alec, et al. [2]

[1] Masked Autoencoders Are Scalable Vision Learners
[2] Learning Transferable Visual Models From Natural Language Supervision

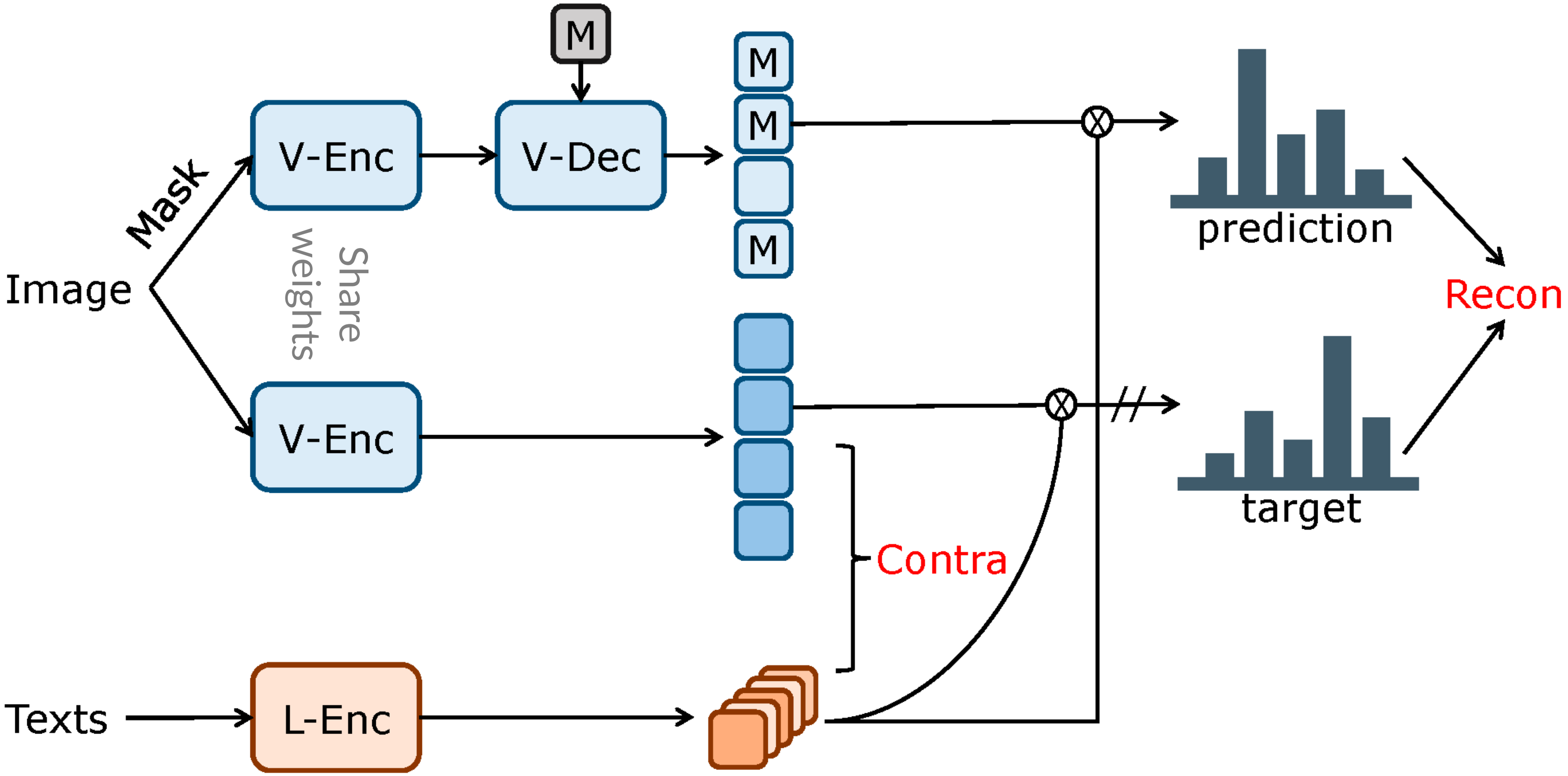
Intuition & Observation

- MIM & ITC can benefit each other
  - MIM brings local supervision, ITC brings global supervision
  - MIM excels at local relation modeling, ITC excels at global semantic alignment

- Naïve combination (MAE+CLIP) shows unsatisfactory mutual benefit
  - Reconstruction raw RGB pixels may be inconsistent with ITC
  - Two objectives should be more aligned with each other for better performance

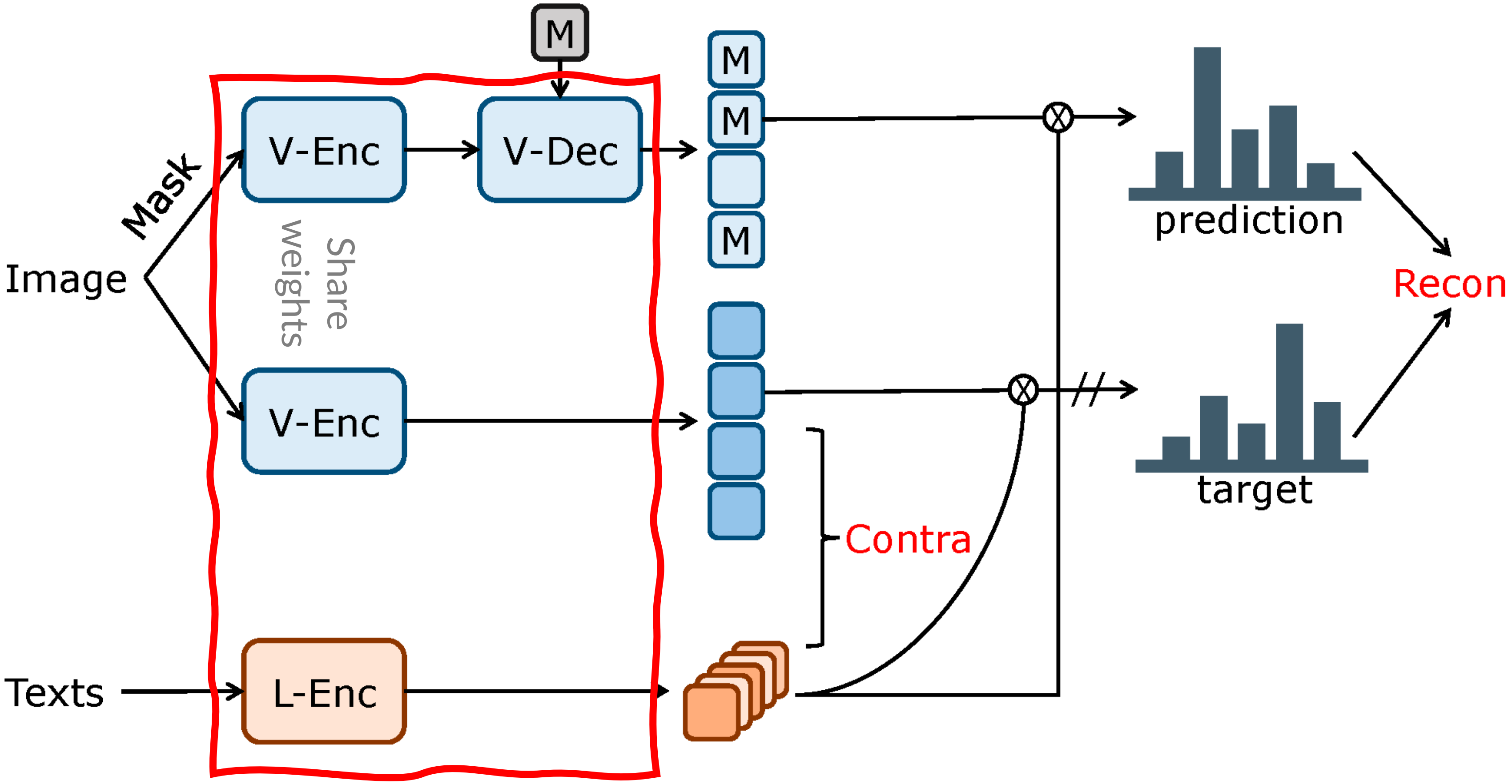


Our RILS



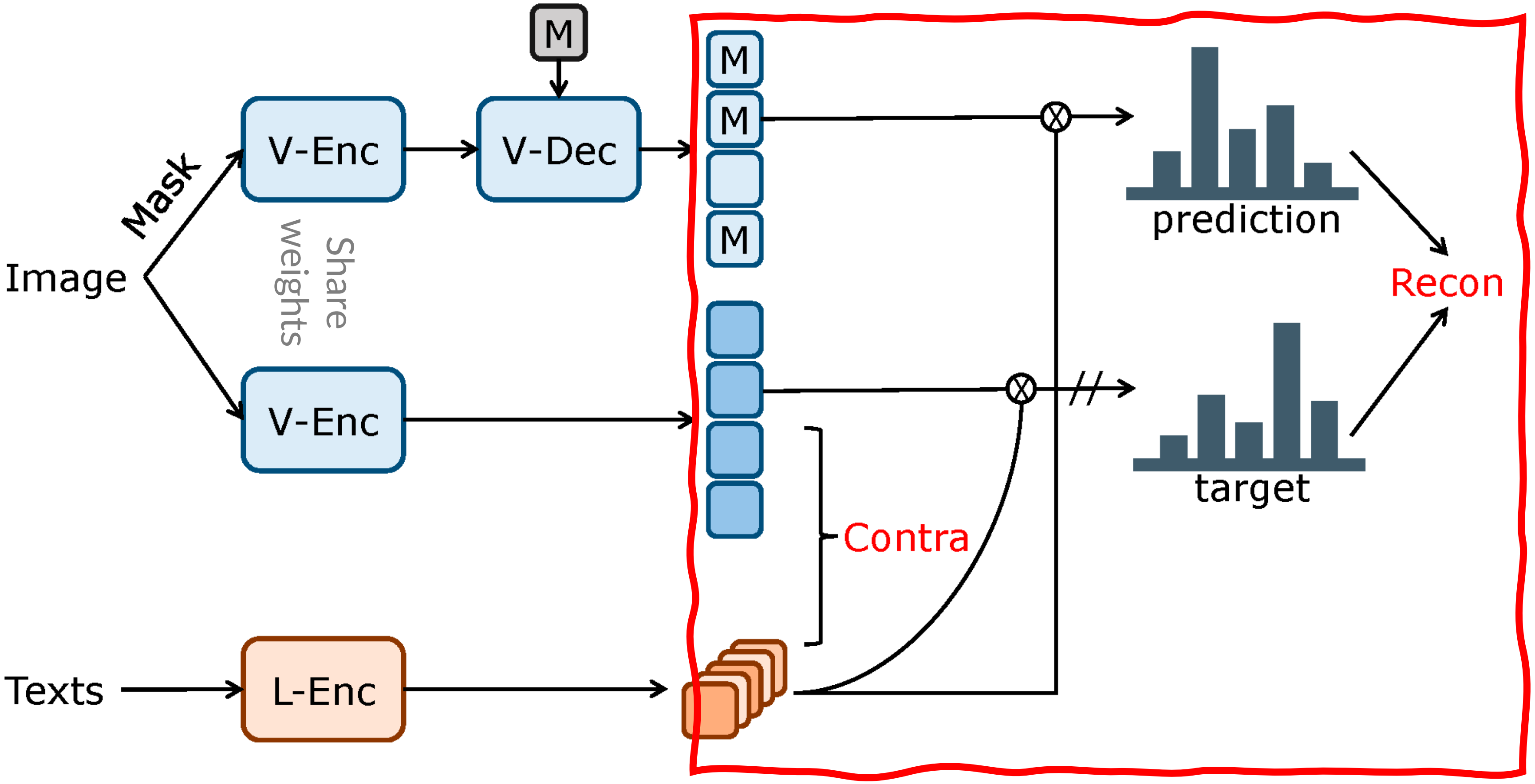
- Core insight: Reconstruction in language semantic space
- Three transformer networks
- Two objectives

Our RILS



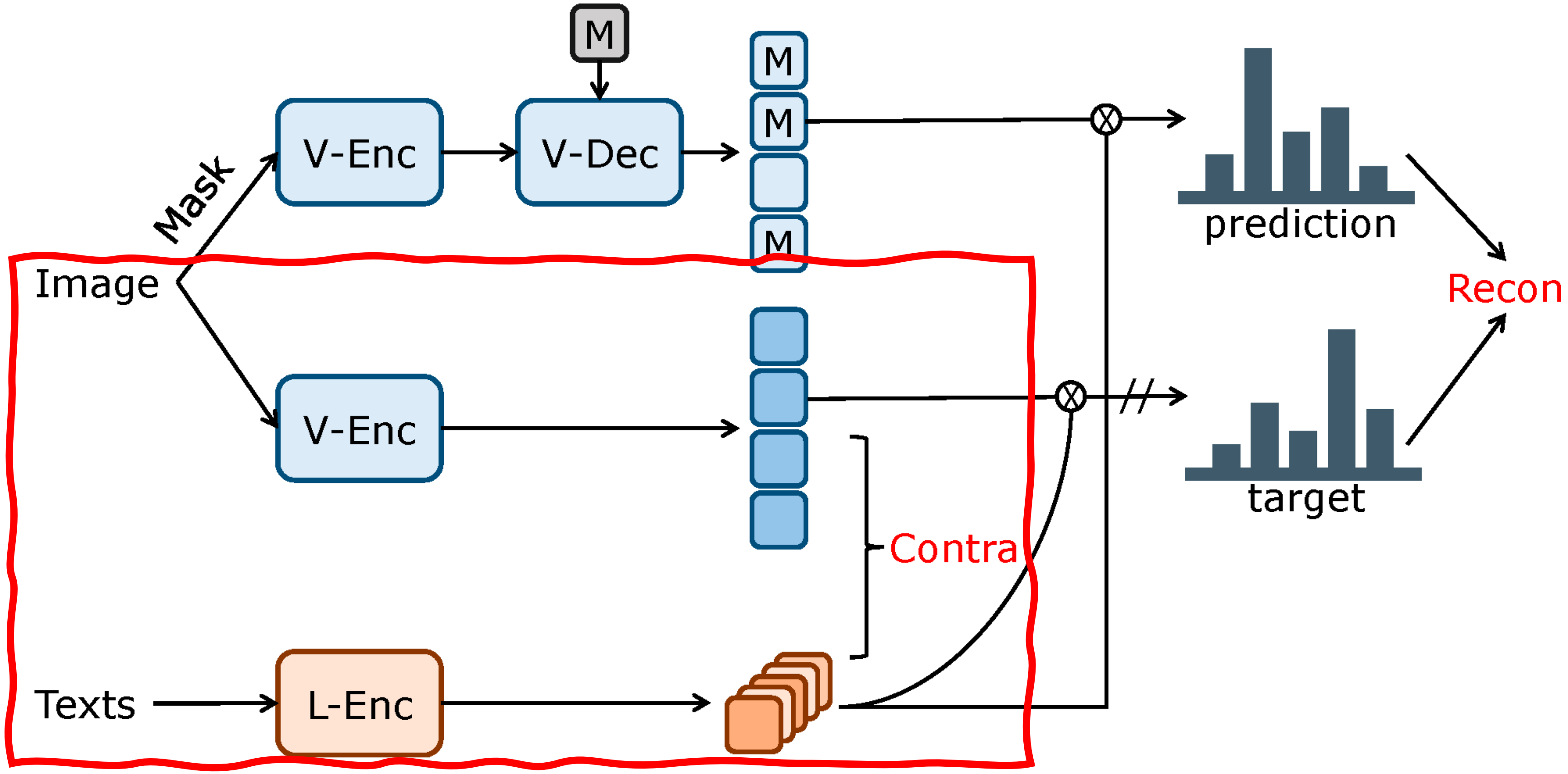
- Core insight: Reconstruction in language semantic space
- Three transformer networks
- Two objectives

Our RILS



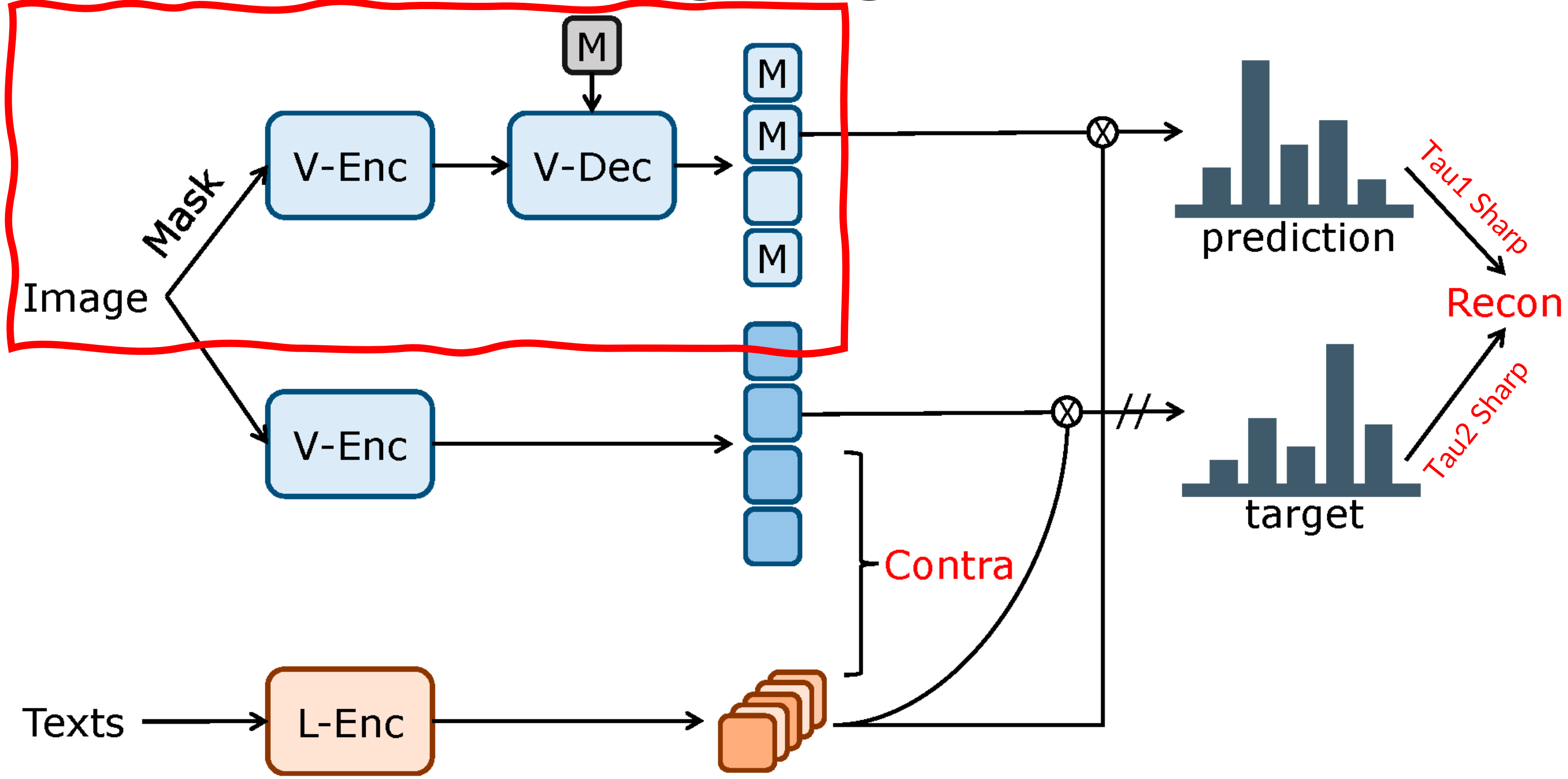
- Core insight: Reconstruction in language semantic space
- Three transformer networks
- Two objectives

Image-text Contrastive



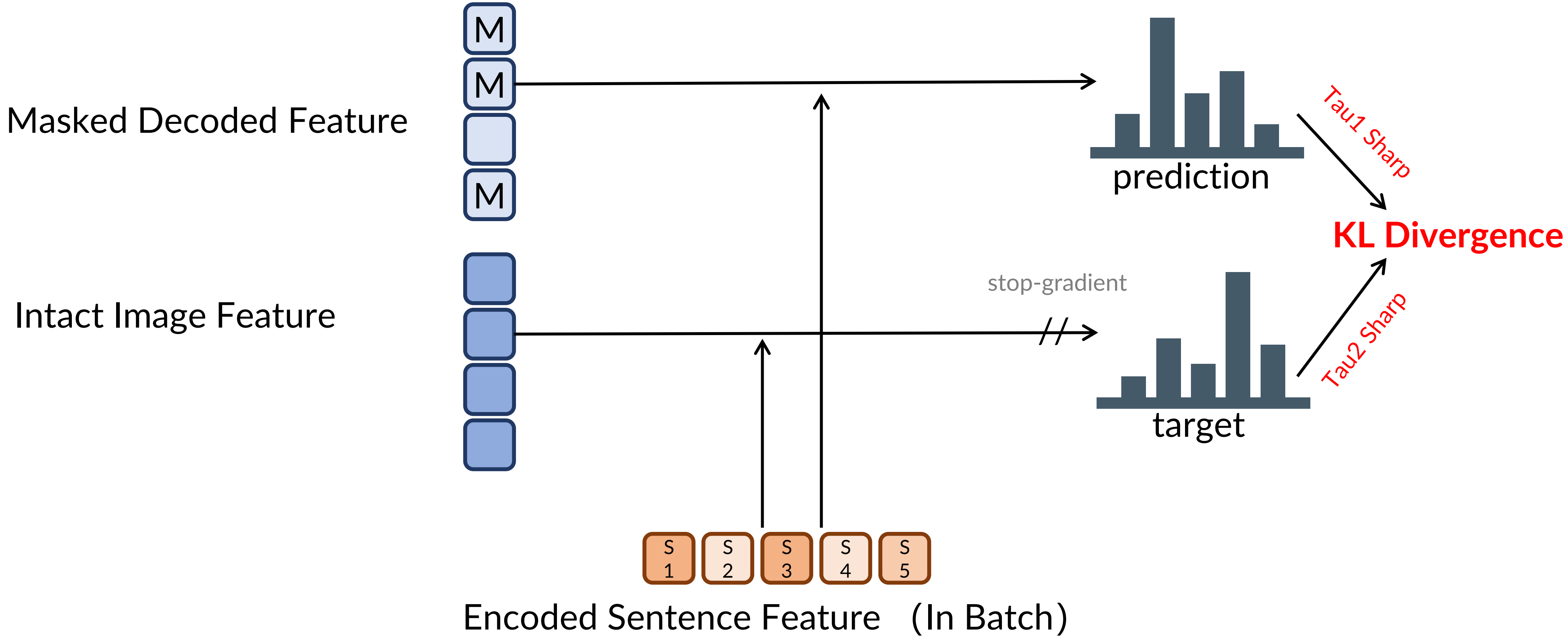
- Original Images and texts are fed into vision encoder and text encoder
- Contrastive learning on encoded image features and encoded text features

Reconstruct in Language Space



- Asymmetric encoder-decoder design
- Masked image is fed into V-Enc and V-Dec to extract features and reconstruct visual signals

Reconstruct in Language Space



- Masked decoded features and original encoded features are mapped to probabilistic distribution over in-batch text features (patch-sentence prob)
- Minimize the KL divergence between prediction and target

Training Objective

$$\mathcal{L}_{\mathrm{I2T}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_i^I, z_i^T \rangle/\sigma)}{\sum_{j=1}^{B} \exp(\langle z_i^I, z_j^T \rangle/\sigma)},$$

$$\mathcal{L}_{\mathrm{T2I}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\langle z_i^T, z_i^I \rangle/\sigma)}{\sum_{j=1}^{B} \exp(\langle z_i^T, z_j^I \rangle/\sigma)},$$

Image-text Contrastive Loss (InfoNCE)

$$\boldsymbol{p}_i^k = \{\frac{\exp(\langle \tilde{f}_i^k, z_l^T \rangle/\tau_1)}{\sum_{j=1}^{B} \exp(\langle \tilde{f}_i^k, z_j^T \rangle/\tau_1)} \mid l \in [1, B]\},$$

$$\boldsymbol{q}_i^k = \{\frac{\exp(\langle \tilde{g}_i^k, z_l^T \rangle/\tau_2)}{\sum_{j=1}^{B} \exp(\langle \tilde{g}_i^k, z_j^T \rangle/\tau_2)} \mid l \in [1, B]\},$$

$$\mathcal{L}_{\mathrm{Recon}} = \frac{1}{\mathcal{C} \cdot ||\mathcal{M}||} \sum_{i \in \mathcal{C}} \sum_{k \in \mathcal{M}} -\mathrm{sg}[\boldsymbol{p}_i^k] \log \boldsymbol{q}_i^k,$$

Reconstruction Loss (KL Divergence)

$$\mathcal{L}_{\mathrm{RILS}} = \lambda_1 \cdot \mathcal{L}_{\mathrm{Contra}} + \lambda_2 \cdot \mathcal{L}_{\mathrm{Recon}}.$$

Pre-training

- Vanilla ViT as vision encoder
- 1-layer ViT block as vision decoder
- 20M image-text pairs sample from Laion-400M
- 25 epochs + 32 gpus

ImageNet Classification

Method	PT Dataset	PT Epoch	Lin. Probe	Fine-tuning
MAE	Laion 20M	25(~400)	44.3	82.1
CLIP			67.8	82.7
MAE+CLIP			64.5	82.9
RILS			71.5	83.3
MAE	IN-1K	1600	67.8	83.6
RILS	Laion 50M	25(~1000)	71.9	83.6

Better performance on linear probe and end-to-end fine-tuning

Detection & Segmentation

Method	COCO		LVIS		ADE20K
	Det	Inst Seg	Det	Inst Seg	Sem Seg
MAE	48.1	42.4	31.0	29.6	44.2
CLIP	47.7	42.0	32.3	30.5	45.2
MAE+CLIP	48.1	42.4	32.6	30.7	45.3
RILS	48.5	42.6	33.8	31.6	48.1

80 Categories **>1000 Categories** **150 Categories**

Obviously better results on complex and fine-grained image understanding

Label Efficient Transfer

Method	IN1K (images per class)			COCO (sampling ratio)		
	1	2	10	2%	10%	20%
MAE	3.4	5.2	14.8	6.10	23.16	29.78
CLIP	19.4	29.2	46.3	5.05	22.49	29.88
MAE+CLIP	17.7	27.2	46.4	5.28	23.72	29.53
RILS	24.0	34.6	51.8	6.46	24.69	31.97

Strong out-of-the-box capacity by performing reconstruction in language semantic space

Zero-shot Classification and Retrieval

Method	Food101	CIFAR10	CIFAR100	CUB200	SUN397	Cars	Aircraft	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	Country211	CLEVR	SST2	ImageNet	**Average**	**# Wins.**
CLIP [47]	55.7	76.0	46.9	**24.4**	50.7	17.8	4.8	31.5	53.7	78.4	31.8	26.8	37.6	89.0	22.7	36.9	**24.1**	6.8	20.0	49.1	40.3	39.3	2
SLIP [43]	56.7	73.4	43.2	22.6	51.6	17.7	4.9	32.4	52.5	79.1	33.3	**29.4**	33.5	89.5	17.8	36.2	17.8	6.8	**23.4**	49.7	41.6	38.7	2
MAE+CLIP	57.8	78.2	52.4	23.9	51.6	18.1	4.6	31.5	55.8	78.4	32.0	27.6	32.7	89.8	27.0	39.4	22.9	7.2	14.7	49.3	42.3	39.9	0
RILS	**58.9**	**86.2**	**55.1**	23.4	**51.8**	**19.5**	**5.9**	**32.8**	**59.2**	**80.7**	**33.5**	22.6	**40.1**	**93.2**	**28.8**	**40.2**	19.1	**7.8**	16.8	**50.0**	**45.0**	**42.3**	**17**

RILS wins 17 over 21 classification datasets

Method	Z.S. COCO Retrieval			
	I2T R@1	I2T R@5	T2I R@1	T2I R@5
CLIP	41.82	69.50	30.54	57.10
SLIP	44.54	72.20	33.26	59.66
MAE+CLIP	42.72	70.66	31.40	57.50
RILS	45.06	73.38	34.86	61.36

Better image-text alignment

Robustness on OOD classification

Method	IN-A	IN-R	IN-Sketch	IN-V2	ObjectNet	Avg.
CLIP	9.3	51.2	28.1	39.8	17.7	32.3
SLIP	10.5	49.8	26.7	41.3	20.4	33.1
MAE+CLIP	11.6	53.9	31.1	41.6	19.4	34.4
RILS	**12.1**	**55.7**	**31.4**	**43.3**	**21.0**	**35.7**

RILS wins on all 5 ImageNet1K out-of-distribution variants

Comparisons with counterparts

Method	ZS.	Lin.	FT.
MAE [28]	–	43.4	81.5
CLIP [47]	32.1	64.1	82.0
MIM→LiT [70]	13.2	43.4	81.5
MIM→CLIP	34.4	64.8	82.2
CLIP→MIM [34, 44, 63]	–	66.2	82.4
RILS (E2E)	**37.5**	**68.5**	**82.7**

Reconstruction Space	ZS.	Lin.	FT.
Raw Pixel Space (MAE+CLIP)	34.2	61.9	82.2
High-level Vision Space [12, 74]	34.8	67.7	82.4
Language Semantic Space (RILS)	**37.5**	**68.5**	**82.7**

All models are trained on exact the same dataset

RILS outperforms its two-stage counterparts

Reconstruction space matters

Summary

- An end-to-end visual pre-training method by leveraging MIM + ITC
- To achieve better mutual benefit between MIM and ITC, we propose to perform masked reconstruction in language semantic space
- Local- and global- supervision → better performance on fine-/coarse-grained tasks
- Reconstruct in language space → better vision-language alignment → Better performance on complex task and zero-shot/low-shot ability.

Thanks For Your Attention!