Zhai et al. (2022)

LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning

Presenter: Kiet Nguyen

CS 6804, 2/27/23

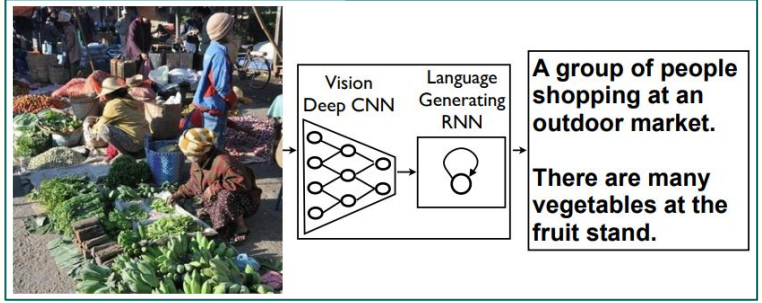
Overview

1. **Background**
2. **Motivation**
3. **Contrastive Language-Image Pretraining** (CLIP; Radford ‘21)
4. **Locked-image Tuning** (LiT🔥; Zhai ‘22)
    - Related Work
    - Methodology
    - Experiments
    - Strengths
    - Weaknesses
    - Future Work
5. **Discussion**

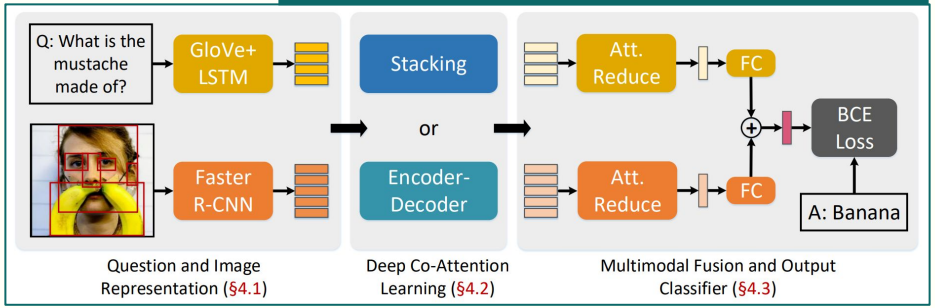
Background: V-L Models

- Task-specific training (late 2014–19)
  - **Pioneers**: COCO Captions (Chen ‘15), VQA (Antol ‘15)
  - **Tasks**: image captioning, VQA, cross-modal retrieval
  - **Limitation**: *task* generalizability

Show and Tell (Vinyals ‘15)



Modular Co-Attention Network (MCAN; Yu ‘19)

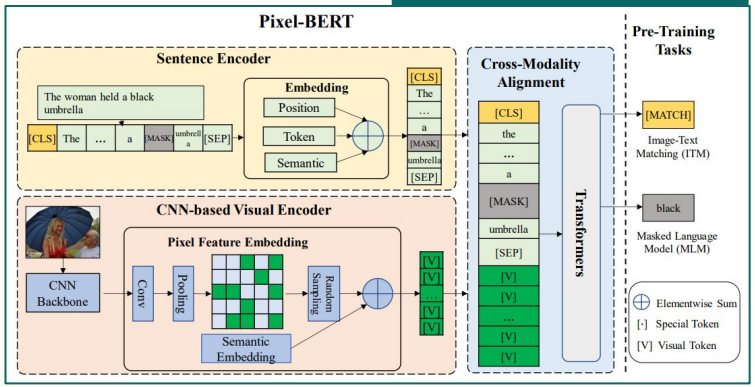


Images: From linked articles.

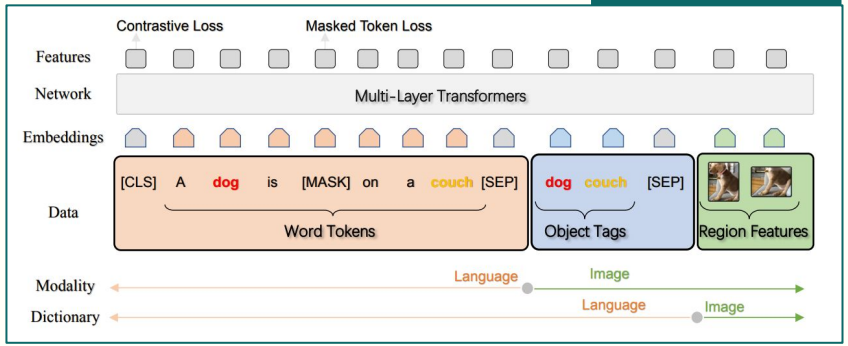
Background: V-L Models

- Medium-scale pretraining (2019–21)
  - 1Ms-10Ms of data pairs
  - **Pioneer**: BERT (Devlin ‘19)
  - **Limitation**: *data* generalizability

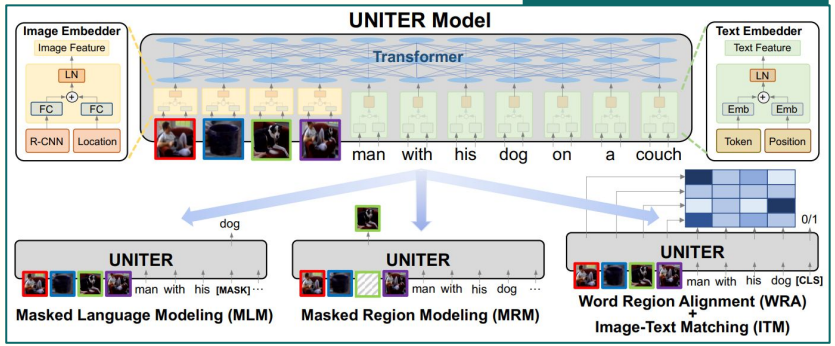
PixelBERT (Huang ‘20)



OSCAR (Li ‘20)



UNITER (Chen ‘20)

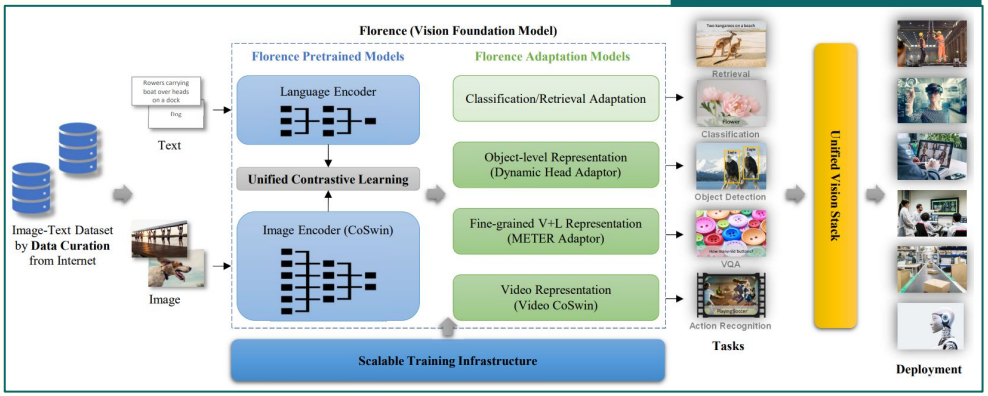


Images: From linked articles.

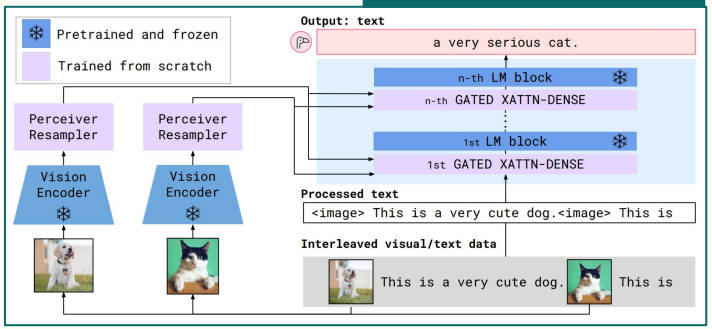
Background: V-L Models

- Large-scale pretraining (2021–present)
  - 100Ms-1Bs of data pairs
  - **Pioneers**: CLIP (Radford ‘21), ALIGN (Jia ‘21)
  - Prevalent use of contrastive learning objectives

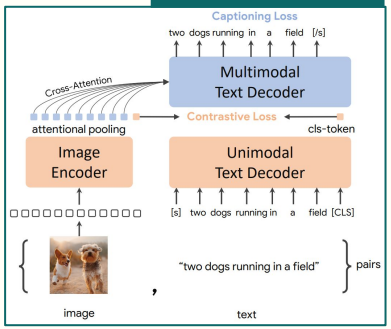
Florence (Yuan ‘21)



Flamingo (Alayrac ‘22)



CoCa (Yu ‘22)



Images: From linked articles.

Motivation: Contrastive Learning

- NLP: large-scale raw text pretraining >> task-specific training
  - *E.g.*, BERT (Devlin ‘19), GPT-3 (Brown ‘20), T5 (Raffel ‘20)
  - Zero-shot transfer
  - → V-L pretraining on raw Internet data at scale?
- Scale/task tradeoff
- Contrastive > predictive learning (Tian ‘20)

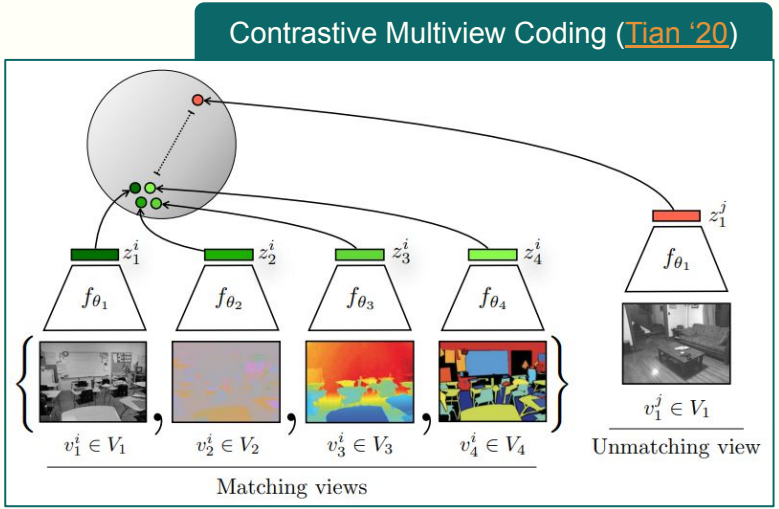


Image: Tian, Yonglong et al. “Contrastive Multiview Coding.” *Computer Vision–ECCV 2020: 16th European Conference*. Springer International Publishing, 2020.

CLIP: **C**ontrastive **L**anguage-**I**mage **P**retraining

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "**Learning Transferable Visual Models from Natural Language Supervision**." In *International Conference on Machine Learning*, pp. 8748-8763. PMLR, 2021.

- *Contributions*:
  - **CLIP**: efficient, scalable method for V-L contrastive pretraining (OpenAI)
  - **WebImageText**: dataset of 400M raw image-text pairs
  - Evaluation on 30+ tasks (linear probe and zero-shot)

CLIP: Dataset Curation

- **WebImageText**: 400M image-text pairs
- *Base queries*: all words occurring 100+ times on Wikipedia
  - \+ Bigram augmentation
  - \+ Name of Wiki. articles > some search volume
  - \+ WordNet (Miller ‘95) synonyms
  - = 500,000 queries
- Use queries to search for image-text pairs
- Up to 20,000 pairs per query → class balance

CLIP: Simplifying ConVIRT (Zhang '20)

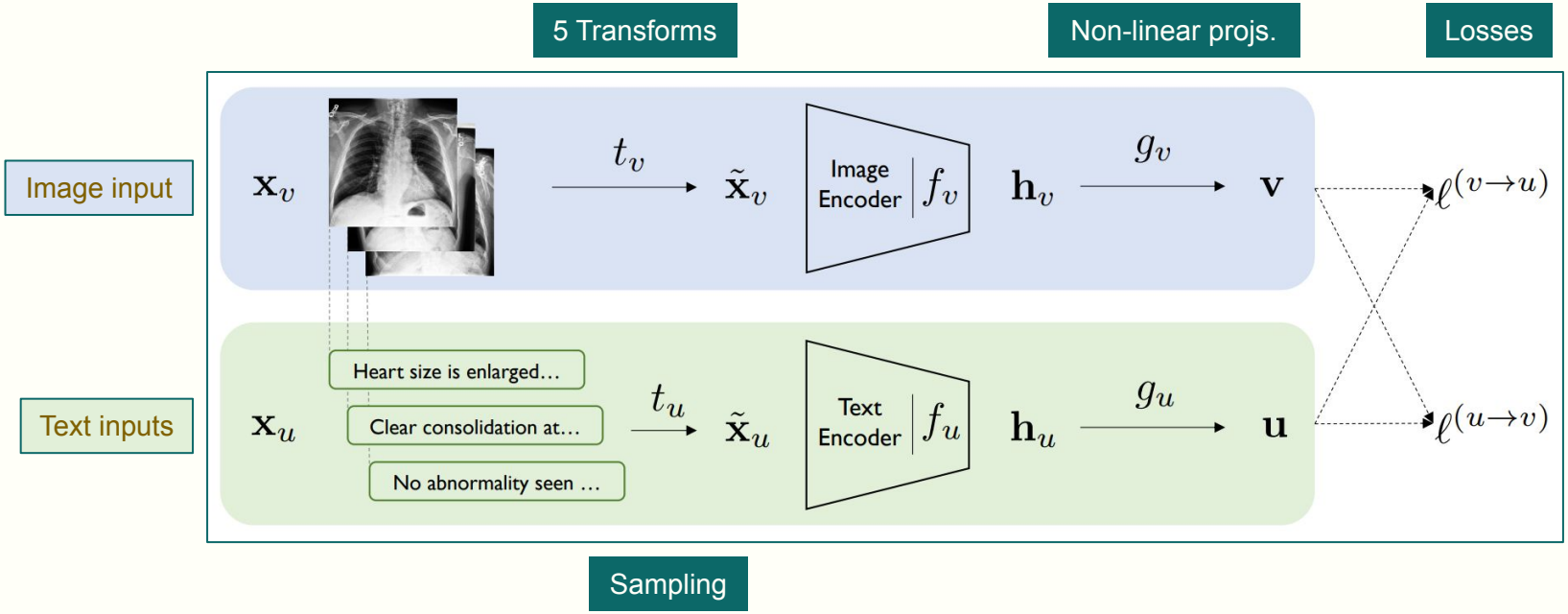


Image: Zhang, Yuhao, et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text." *Machine Learning for Healthcare Conference*. PMLR, 2022.

CLIP: Simplifying ConVIRT (Zhang '20)

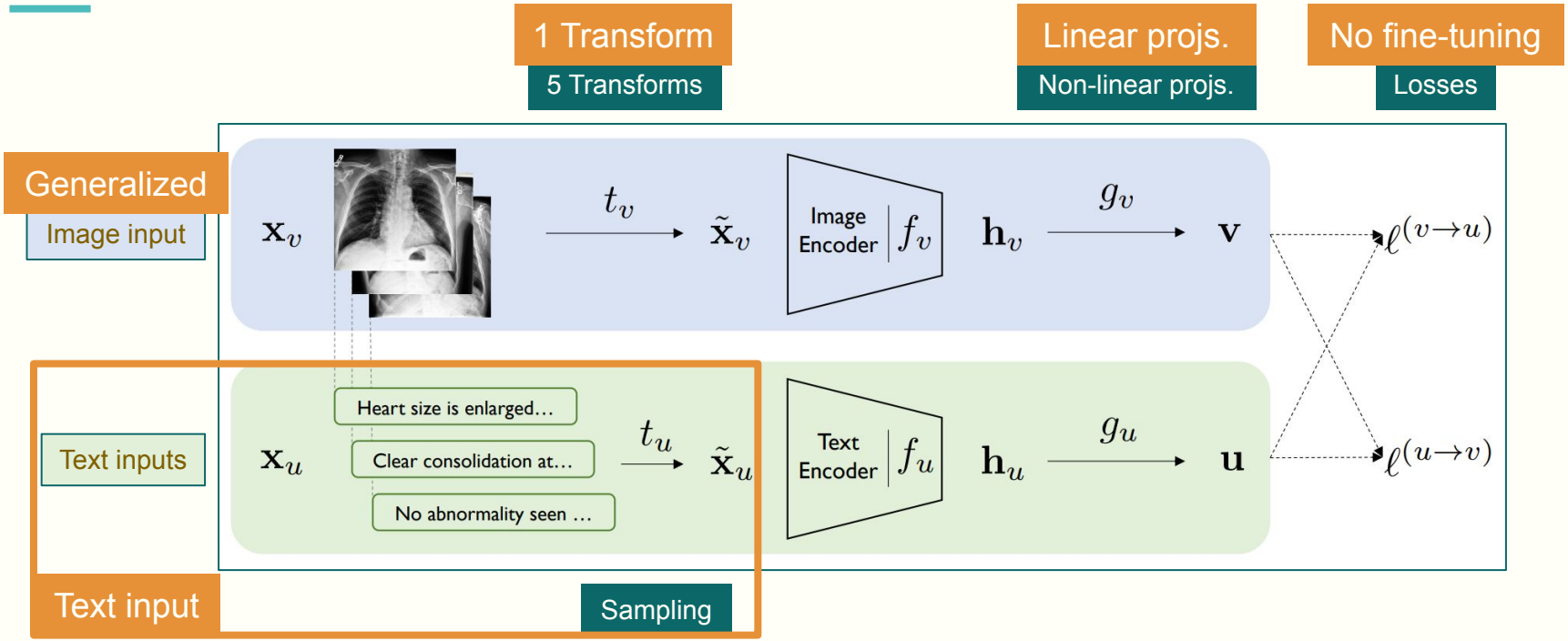


Image: Zhang, Yuhao, et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text." *Machine Learning for Healthcare Conference*. PMLR, 2022.

CLIP: Simplifying ConVIRT (Zhang '20)

- Contrastive losses:
  - Image-to-text:

$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)},$$

  - Text-to-image:

$$\ell_i^{(u \to v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}.$$

  - Final:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left(\lambda \ell_i^{(v \to u)} + (1 - \lambda) \ell_i^{(u \to v)} \right),$$

- Directly optimize τ instead of hyperparameter fine-tuning

Equations: Zhang, Yuhao, et al. "Contrastive Learning of Medical Visual Representations from Paired Images and Text." *Machine Learning for Healthcare Conference*. PMLR, 2022.

CLIP: Model Selection

- Text encoder: Transformer (Vaswani ‘17, Radford ‘19)
- Img. encoder:
  - 5 ResNets (He ‘16)
    - RN50, 101, 50x4, 50x16, 50x64
  - 3 ViTs (Dosovitskiy ‘20)
    - ViT-B/16, B/32, L/14

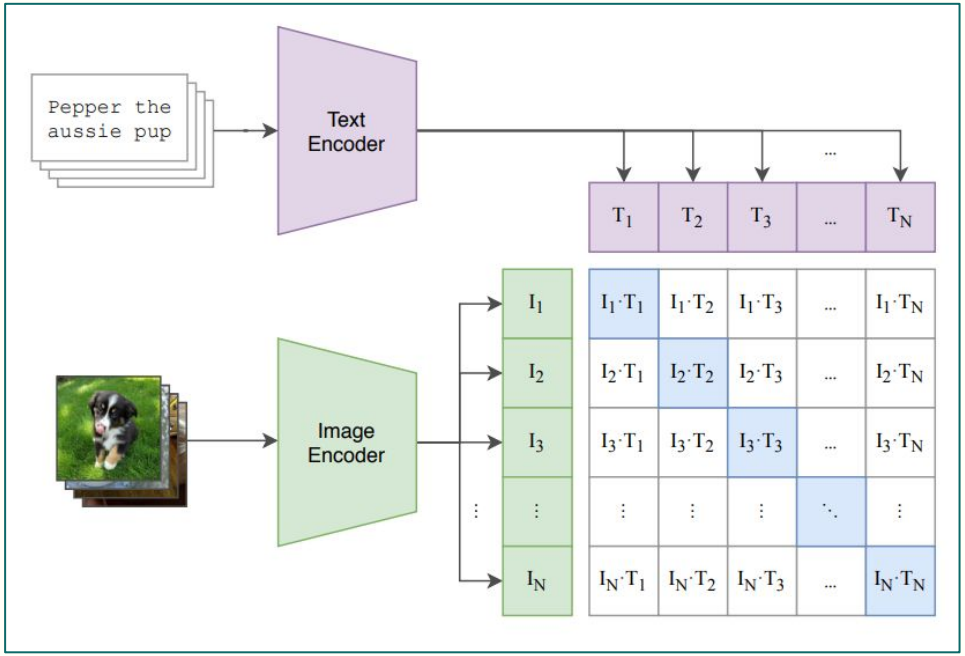


Image: Radford, Alec, et al. “Learning Transferable Visual Models from Natural Language Supervision.” *International Conference on Machine Learning*. PMLR, 2021.

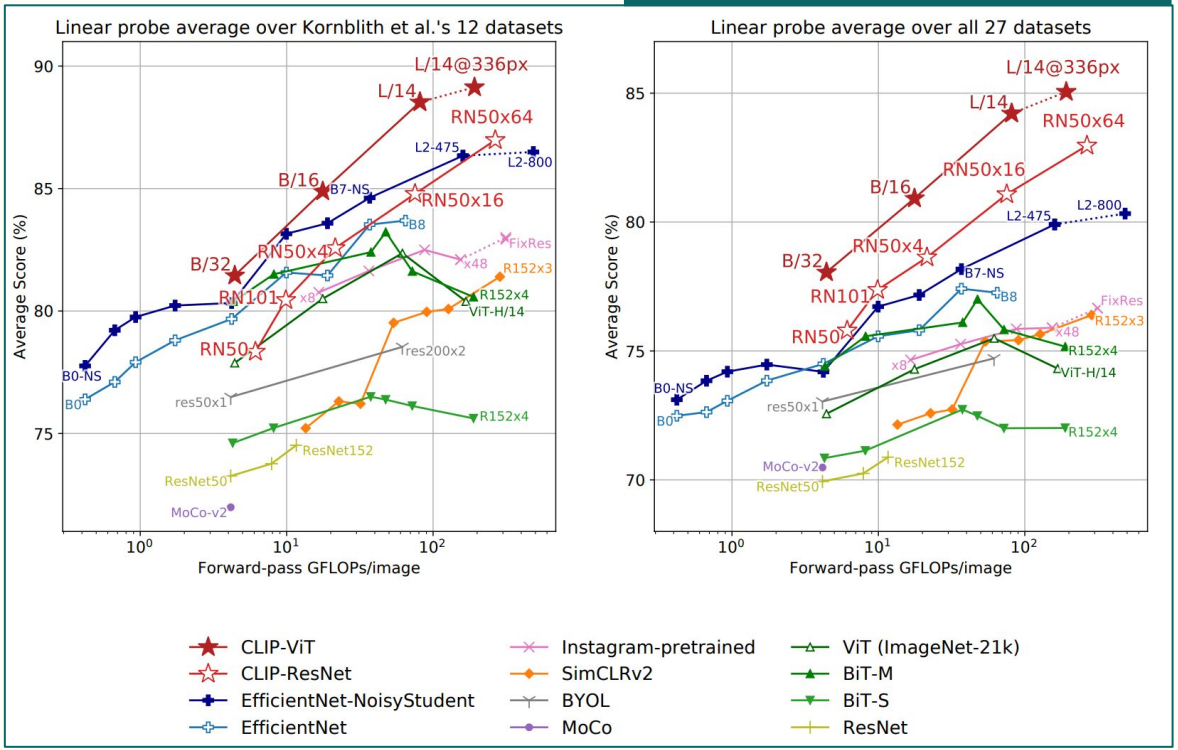
CLIP: Representation Learning

- **Linear probe**: logistic regression classifier on image features

Datasets (Kornblith ‘19)

Dataset	Classes
Food-101 [5]	101
CIFAR-10 [37]	10
CIFAR-100 [37]	10
Birdsnap [4]	500
SUN397 [72]	397
Stanford Cars [36]	196
FGVC Aircraft [48]	100
PASCAL VOC 2007 Cls. [19]	20
Describable Textures (DTD) [10]	47
Oxford-IIIT Pets [53]	37
Caltech-101 [20]	102
Oxford 102 Flowers [52]	102

CLIP linear probe results (Radford ‘21)



Images: Radford, Alec, et al. “Learning Transferable Visual Models from Natural Language Supervision.” *International Conference on Machine Learning*. PMLR, 2021.

CLIP: Zero-Shot Image Classification

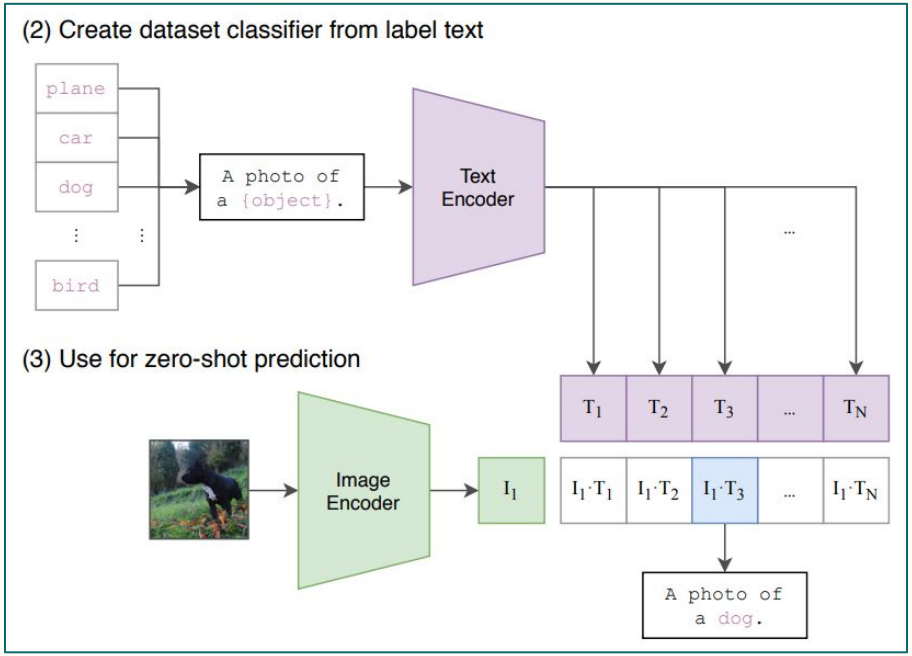


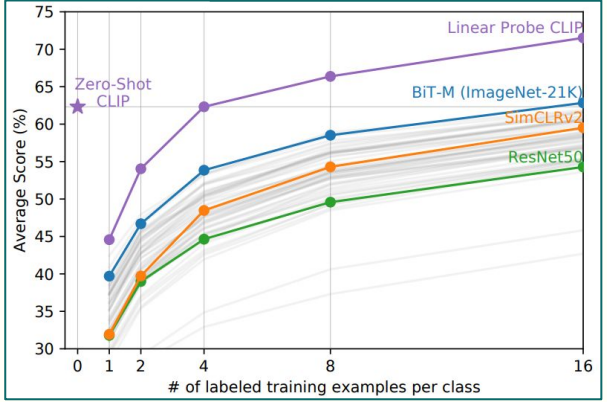
Image: Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning*. PMLR, 2021.

CLIP: Zero-Shot Evaluation

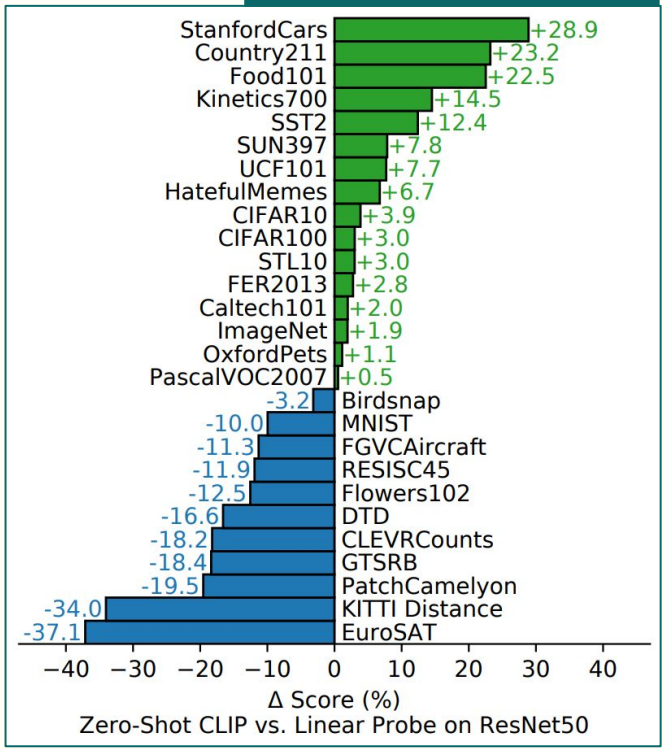
vs. Visual N-Grams (Li '17)

	aYahoo	ImageNet	SUN
Visual N-Grams	72.4	11.5	23.0
CLIP	**98.4**	**76.2**	**58.5**

vs. few-shot linear probes



vs. ResNet50 linear probes



Images: Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning*. PMLR, 2021.

CLIP: Robust to Natural Distribution Shifts

- 7 natural distribution shifts datasets (Taori ‘20):
  - ImageNet-V2 (IN-V2; Recht ‘19): redo IN data collection
  - IN-R (Hendrycks ‘20): **R**enditions (sculptures, paintings)
  - ObjectNet (Barbu ‘19): objects with overlapping classes
  - IN-Sketch (Wang ‘19): sketches with overlapping classes
  - IN-A (Hendrycks ‘19): **A**dversarial (images that ResNet fails on)
  - IN-Vid-Robust (Shankar ‘19), YTBB-Robust (Gu ‘19, Shankar ‘19): classification consistency across video frames

CLIP: Robust to Natural Distribution Shifts

- IN-Vid-Robust (Shankar ‘19), YTBB-Robust (Gu ‘19, Shankar ‘19): classification consistency across video frames

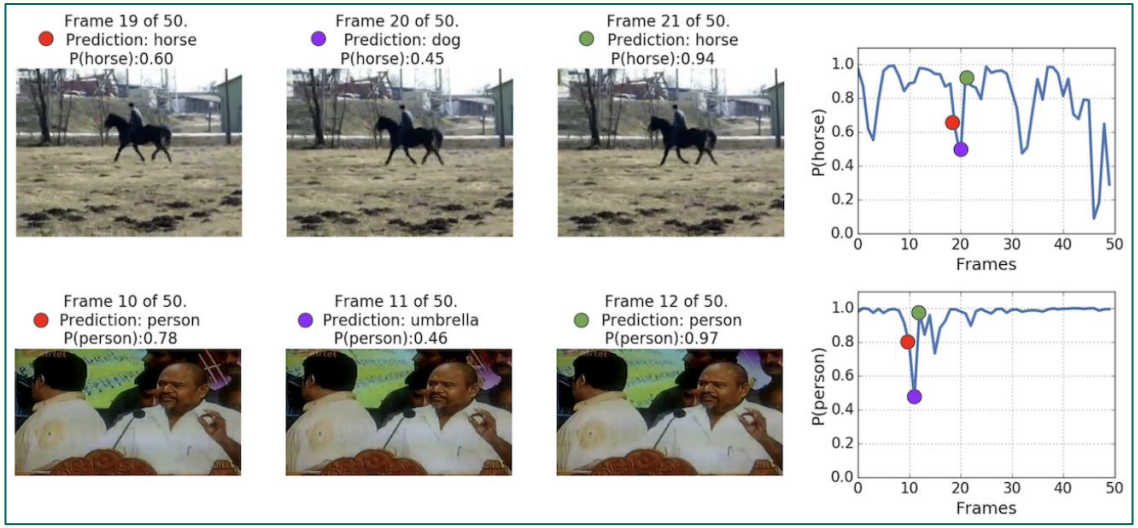


Image: Gu, Keren, et al. “Using Videos to Evaluate Image Model Robustness.” *arXiv preprint arXiv:1904.10076* (2019).

CLIP: Robust to Natural Distribution Shifts

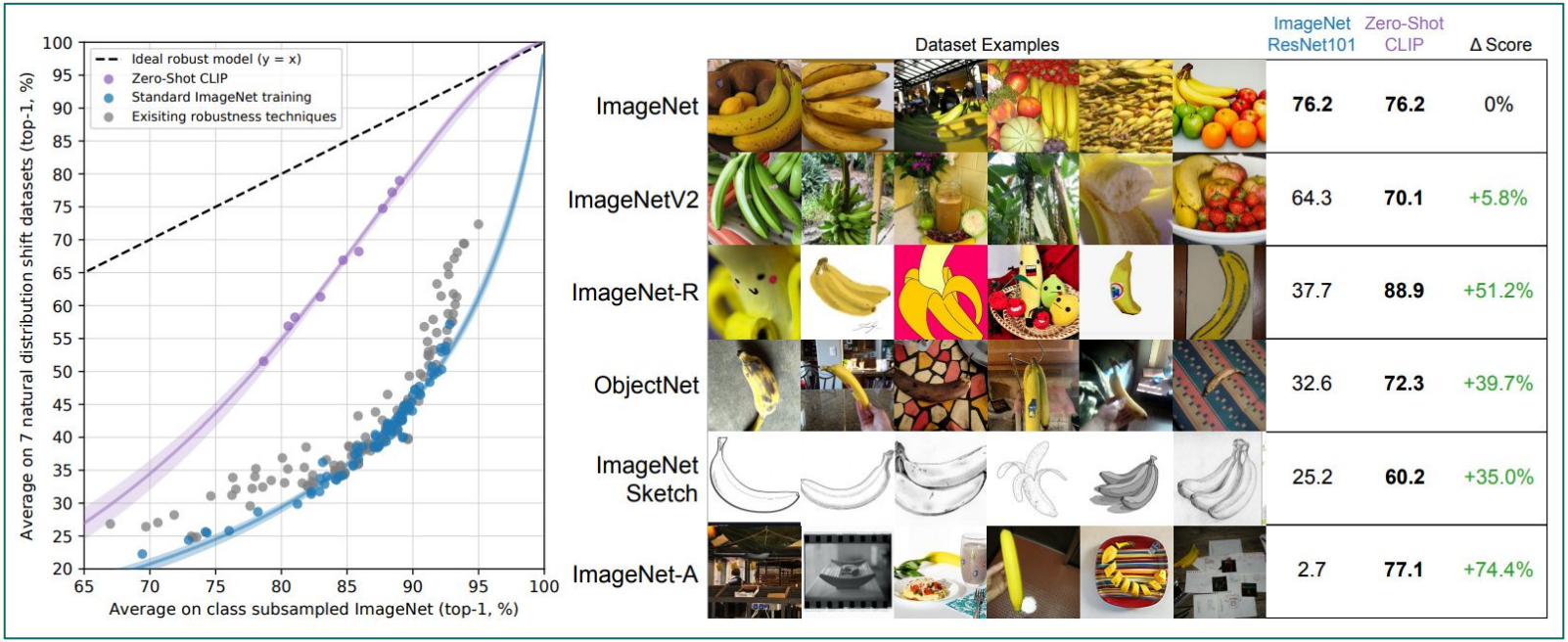


Image: Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning*. PMLR, 2021.

CLIP vs. ALIGN (Jia ‘22)

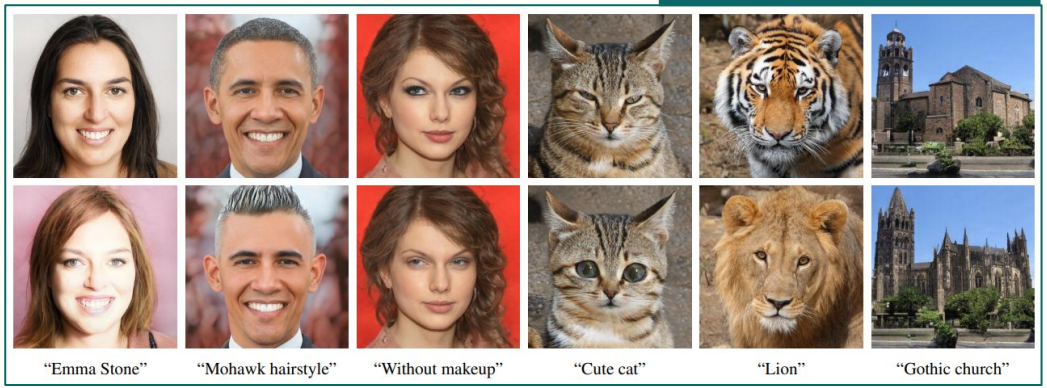
- **ALIGN**: **A** **L**arge-Scale **I**ma**G**e and **N**oisy-Text Embedding
- Differences vs. CLIP:
  - Dataset:
    - Size: **1.8B raw image-text pairs** vs. 400M
    - **Minimal filtering** → larger, noisier dataset
  - Architecture:
    - Image encoder: **EfficientNet** (Tan ‘19) vs. ResNet/ViT
    - Text encoder: **BERT** vs. Transformer
- Significantly better performance in **text-to-image retrieval** (7-8% R@1 increase); otherwise comparable

CLIP: Strengths

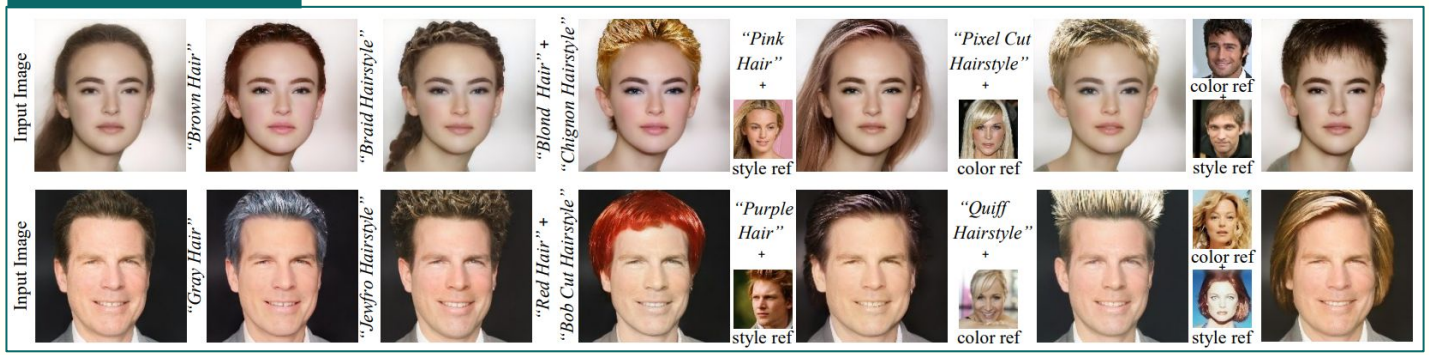
- Efficiency
  - Contrastive loss
  - ViT vs. ResNet
- Generalizability
- Strong **global** V-L representations
  - Applications in text-guided image generation

CLIP: Applications

StyleCLIP (Patashnik ‘21)



HairCLIP (Wei ‘22)



Images: From linked articles.

CLIP: Applications

DiffusionCLIP (Kim '22)

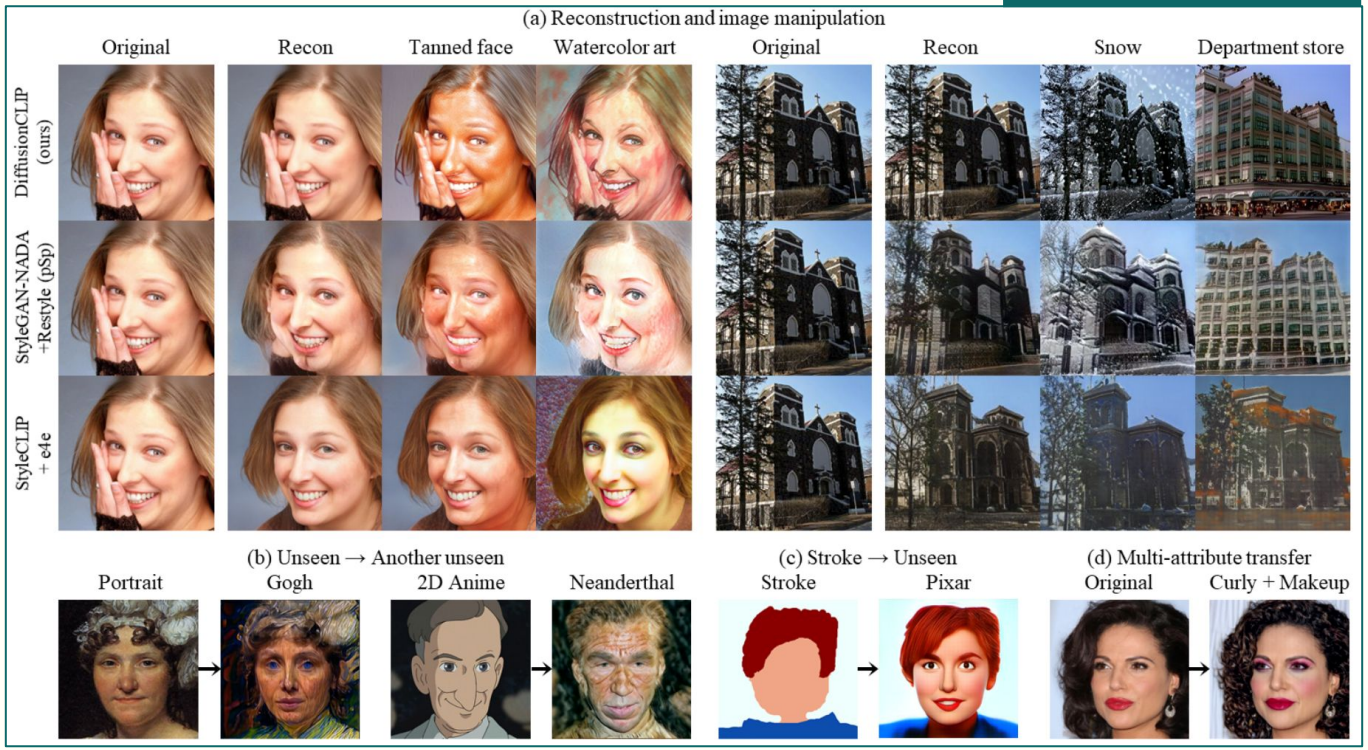
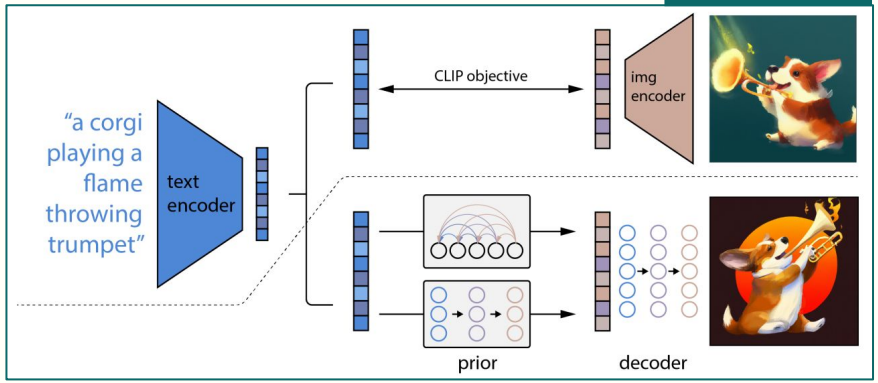


Image: Kim, Gwanghyun, et al. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

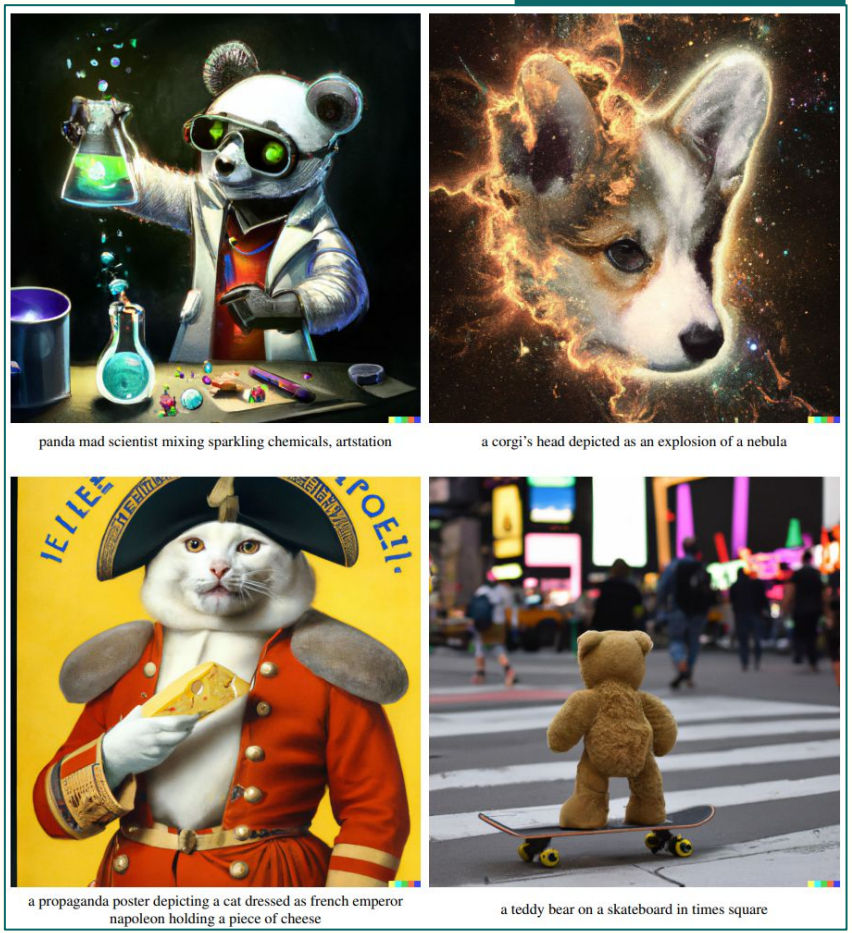
CLIP: Applications

- DALL·E 2 (Ramesh ‘22)

Architecture



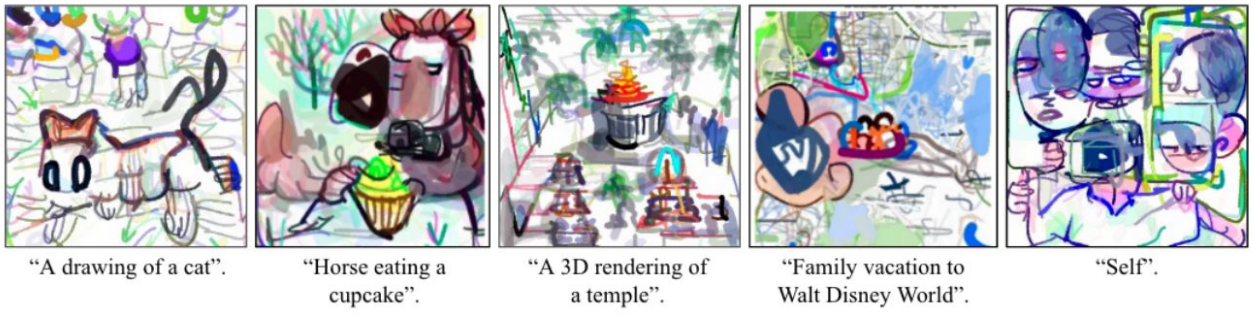
Qualitative examples



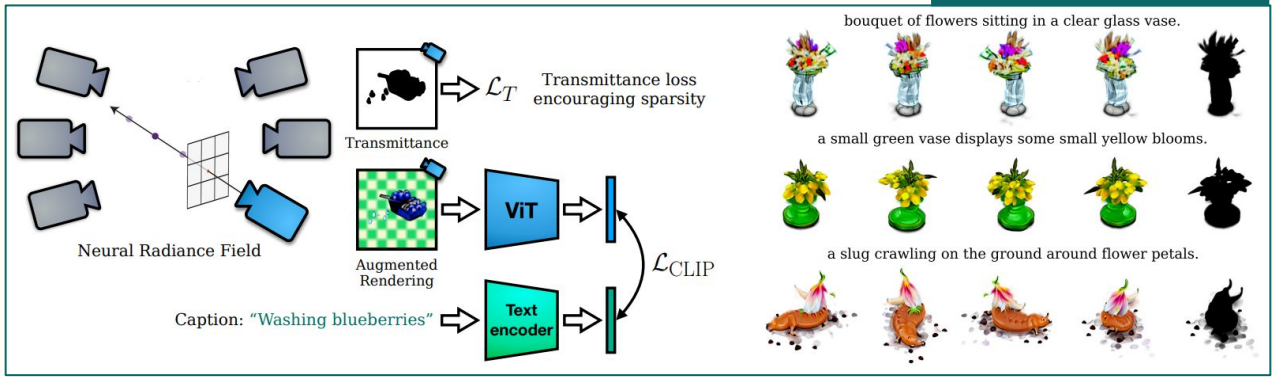
Images: Ramesh, Aditya, et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents." *arXiv preprint arXiv:2204.06125* (2022).

CLIP: Applications

CLIPDraw (Frans ‘21)



Dream Field (Jain ‘22)



Images: From linked articles.

CLIP: Weaknesses

- Zero-shot performance well worse than fine-tuned SotA
- Does not work well with image regions
- Struggles on:
  - More abstract tasks (*e.g.*, counting)
  - Reasoning tasks (*e.g.*, VQA)
  - Fine-grained tasks (*e.g.*, car models)
  - Out-of-distribution datasets (*e.g.*, MNIST)
- Sensitive to text wording

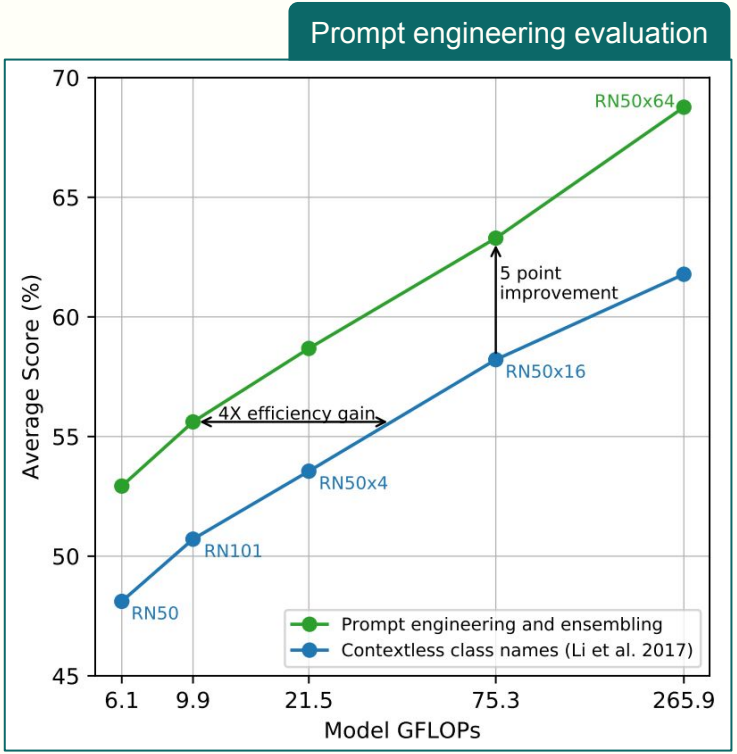


Image: Radford, Alec, et al. "Learning Transferable Visual Models from Natural Language Supervision." *International Conference on Machine Learning*. PMLR, 2021.

CLIP: More Weaknesses

- CLIP and ALIGN: both trained **from scratch**
  - → Data- and compute-inefficient
- Leverage performant pretrained architectures?
  - *e.g.*, **ViT-G/14** (Zhai '22)

Scaled ViT sizes

Name	Width	Depth	MLP	Heads	Mio. Param	GFLOPs 224^2	GFLOPs 384^2
s/28	256	6	1024	8	5.4	0.7	2.0
s/16	256	6	1024	8	5.0	2.2	7.8
S/32	384	12	1536	6	22	2.3	6.9
Ti/16	192	12	768	3	5.5	2.5	9.5
B/32	768	12	3072	12	87	8.7	26.0
S/16	384	12	1536	6	22	9.2	31.2
B/28	768	12	3072	12	87	11.3	30.5
B/16	768	12	3072	12	86	35.1	111.3
L/16	1024	24	4096	16	303	122.9	382.8
g/14	1408	40	6144	16	1011	533.1	1596.4
G/14	1664	48	8192	16	1843	965.3	2859.9

ViT-G/14 performance

Benchmark	ImageNet	INet V2	INet ReaL	ObjectNet	VTAB (light)
NS (Eff.-L2) [48]	88.3	80.2	-	68.5	-
MPL (Eff.-L2) [28]	90.2	-	**91.02**	-	-
CLIP (ViT-L/14) [30]	85.4	75.9	-	**72.3**	-
ALIGN (Eff.-L2) [20]	88.6	70.1	-	-	-
BiT-L (ResNet) [22]	87.54	-	90.54	58.7	76.29
ViT-H/14 [15]	88.55	-	90.72	-	77.63
Our ViT-G/14	**90.45±0.03**	**83.33±0.03**	90.81±0.01	70.53±0.52	**78.29±0.53**

Tables: Zhai, Xiaohua, et al. "Scaling Vision Transformers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

LiT🔥: Locked-Image Tuning

Zhai, Xiaohua, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. "**LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning**." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123-18133. 2022.

- *Contributions*:
  - **Contrastive tuning** (CT): contrastive learning strategy using pretrained models
  - **LiT🔥**: an instance of CT, using frozen image features to tune text features
  - Evaluations and ablations on CT design choices

Related Work

- **Transfer learning**, *i.e.*, the pretraining-finetuning paradigm
  - *e.g.*, ViT (Dosovitskiy ‘20), Big Transfer (BiT; Kolesnikov ‘20)
  - Scaling up pretraining → increased performance (esp. **low-shot**) and robustness
- **Zero-shot transfer**, *i.e.*, pretraining without finetuning:
  - Fusion encoders with non-contrastive objectives (MLM, ITM, etc.)
    - *e.g.*, METER (Dou ‘22)
  - Separate encoders with **contrastive learning** objectives
    - *e.g.*, **CLIP**, ALIGN, ConVIRT
  - Hybrid, *e.g.*, ALBEF (Li ‘22)

Methodology: Contrastive Tuning

- Non-contrastive approach: learn performant embeddings from high-quality image datasets
  - *e.g.*, ImageNet-21k (Deng ‘09), JFT-300M (Sun ‘17)
- However: limited to predefined categories, unlike real-world data

→ **Combine both using pretrained models in contrastive pretraining**

Methodology: Contrastive Tuning

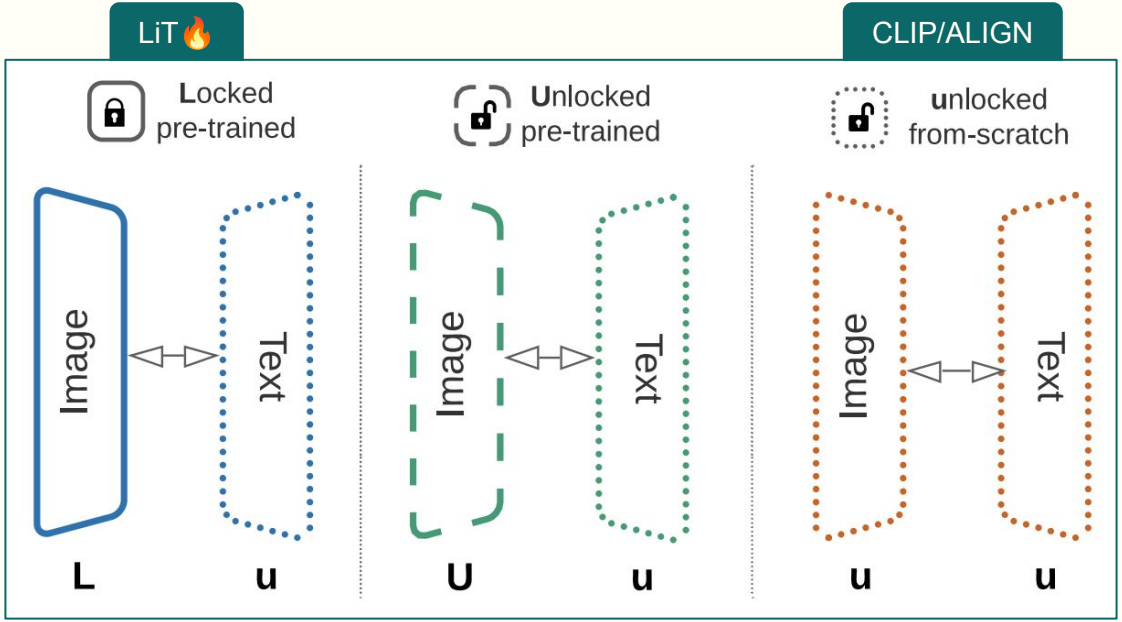


Image: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Implementation Details

- Pretraining datasets:
  - Conceptual Captions 12M (CC12M; Sharma ‘18, Changpinyo ‘21)
  - Yahoo Flickr Creative Commons (YFCC100m; Thomee ‘16)
    - 15M filtered subset: $\text{YFCC100m}_{\text{CLIP}}$
  - LiT🔥 dataset: 4B image-text pairs
    - Filtered similarly to ALIGN with more relaxed text filtering
- Image encoder: ViT-g/14 trained on JFT-3B (Zhai ‘22)
- Tasks:
  - Zero-shot ImageNet classification
  - COCO I→T and T→I retrieval

Experiments: Evaluation Datasets

- *Standard*: ImageNet

Experiments: Evaluation Datasets

- *Standard*: ImageNet
- *Robustness*:
  - IN-V2, IN-R, IN-A, ObjectNet
  - IN-ReaL (Beyer ‘20): **Rea**ssessed **L**abels

Image: Beyer, Lucas, et al. “Are We Done with ImageNet?” *arXiv preprint arXiv:2006.07159* (2020).

Experiments: Evaluation Datasets

- *Standard*: ImageNet
- *Robustness*:
  - IN-V2, IN-R, IN-A, ObjectNet
  - IN-ReaL (Beyer ‘20)

Experiments: Evaluation Datasets

- *Standard*: ImageNet
- *Robustness*:
  - IN-V2, IN-R, IN-A, ObjectNet
  - IN-ReaL (Beyer ‘20)
- *Diversity*: VTAB-Natural (Zhai ‘19)

VTAB (Zhai ‘19)

Category	Dataset	Train size	Classes	Reference
• Natural	Caltech101	3,060	102	(Li et al., 2006)
• Natural	CIFAR-100	50,000	100	(Krizhevsky, 2009)
• Natural	DTD	3,760	47	(Cimpoi et al., 2014)
• Natural	Flowers102	2,040	102	(Nilsback & Zisserman, 2008)
• Natural	Pets	3,680	37	(Parkhi et al., 2012)
• Natural	Sun397	87,003	397	(Xiao et al., 2010)
• Natural	SVHN	73,257	10	(Netzer et al., 2011)
• Specialized	EuroSAT	21,600	10	(Helber et al., 2019)
• Specialized	Resisc45	25,200	45	(Cheng et al., 2017)
• Specialized	Patch Camelyon	294,912	2	(Veeling et al., 2018)
• Specialized	Retinopathy	46,032	5	(Kaggle & EyePacs, 2015)
• Structured	Clevr/count	70,000	8	(Johnson et al., 2017)
• Structured	Clevr/distance	70,000	6	(Johnson et al., 2017)
• Structured	dSprites/location	663,552	16	(Matthey et al., 2017)
• Structured	dSprites/orientation	663,552	16	(Matthey et al., 2017)
• Structured	SmallNORB/azimuth	36,450	18	(LeCun et al., 2004)
• Structured	SmallNORB/elevation	36,450	9	(LeCun et al., 2004)
• Structured	DMLab	88,178	6	(Beattie et al., 2016)
• Structured	KITTI/distance	5,711	4	(Geiger et al., 2013)

Table: Zhai, Xiaohua, et al. “A Large-Scale Study of Representation Learning with the Visual Task Adaptation Benchmark.” *arXiv preprint arXiv:1910.04867* (2019).

Experiments: Evaluation Datasets

- *Standard*: ImageNet
- *Robustness*:
  - IN-V2, IN-R, IN-A, ObjectNet
  - IN-ReaL (Beyer ‘20)
- *Diversity*: VTAB-Natural (Zhai ‘19)

VTAB (Zhai ‘19)

Category	Dataset	Train size	Classes	Reference
• Natural	Caltech101	3,060	102	(Li et al., 2006)
• Natural	CIFAR-100	50,000	100	(Krizhevsky, 2009)
• Natural	DTD	3,760	47	(Cimpoi et al., 2014)
• Natural	Flowers102	2,040	102	(Nilsback & Zisserman, 2008)
• Natural	Pets	3,680	37	(Parkhi et al., 2012)
• Natural	Sun397	87,003	397	(Xiao et al., 2010)
• Natural	SVHN	73,257	10	(Netzer et al., 2011)
• Specialized	EuroSAT	21,600	10	(Helber et al., 2019)
• Specialized	Resisc45	25,200	45	(Cheng et al., 2017)
• Specialized	Patch Camelyon	294,912	2	(Veeling et al., 2018)
• Specialized	Retinopathy	46,032	5	(Kaggle & EyePacs, 2015)
• Structured	Clevr/count	70,000	8	(Johnson et al., 2017)
• Structured	Clevr/distance	70,000	6	(Johnson et al., 2017)
• Structured	dSprites/location	663,552	16	(Matthey et al., 2017)
• Structured	dSprites/orientation	663,552	16	(Matthey et al., 2017)
• Structured	SmallNORB/azimuth	36,450	18	(LeCun et al., 2004)
• Structured	SmallNORB/elevation	36,450	9	(LeCun et al., 2004)
• Structured	DMLab	88,178	6	(Beattie et al., 2016)
• Structured	KITTI/distance	5,711	4	(Geiger et al., 2013)

Table: Zhai, Xiaohua, et al. “A Large-Scale Study of Representation Learning with the Visual Task Adaptation Benchmark.” *arXiv preprint arXiv:1910.04867* (2019).

Experiments: Zero-Shot Performance

Zero-shot classification performance

Dataset	Method	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
Private	CLIP [45]	76.2	70.1	88.9	77.2	72.3	-	-
	ALIGN [30]	76.4	70.1	92.2	75.8	-	-	-
	LiT	**84.5**	**78.7**	**93.9**	**79.4**	**81.1**	88.0	72.6
Public	CLIP [45]	31.3	-	-	-	-	-	-
	OpenCLIP [28]	34.8	30.0	-	-	-	-	-
	LiT	**75.7**	**66.6**	60.4	37.8	54.5	82.1	63.1
*	ResNet50 [25]	75.8	63.8	36.1	0.5	26.5	82.5	72.6

Table: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Zero-Shot Performance

Zero-shot classification performance

Dataset	Method	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
Private	CLIP [45]	76.2	70.1	88.9	77.2	72.3	-	-
	ALIGN [30]	76.4	70.1	92.2	75.8	-	-	-
	LiT	**84.5**	**78.7**	**93.9**	**79.4**	**81.1**	88.0	72.6
Public	CLIP [45]	31.3	-	-	-	-	-	-
	OpenCLIP [28]	34.8	30.0	-	-	-	-	-
	LiT	**75.7**	**66.6**	60.4	37.8	54.5	82.1	63.1
*	ResNet50 [25]	75.8	63.8	36.1	0.5	26.5	82.5	72.6

ViT-g/14

ViT-L/16

Pretrained on full dataset

Pretrained on YFCC100m$_{CLIP}$ + CC12M

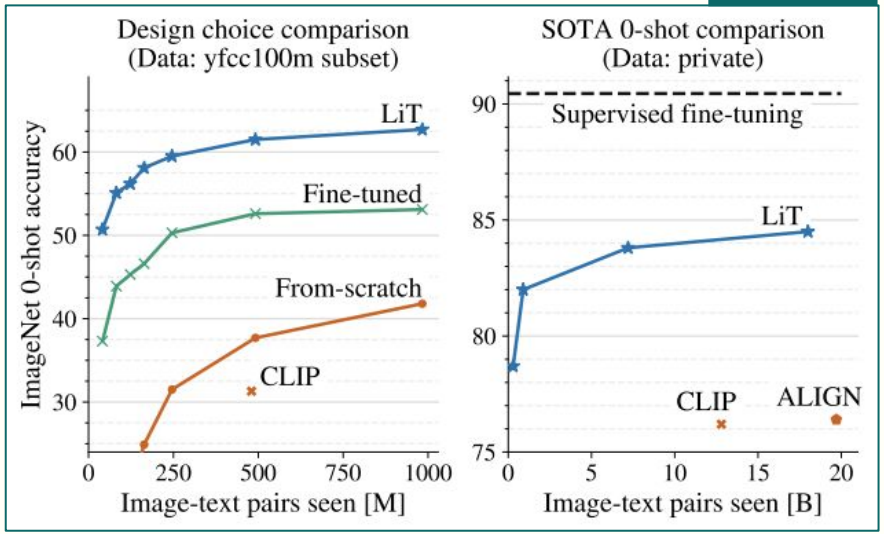
Table: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Zero-Shot Performance

Zero-shot classification performance

Dataset	Method	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
Private (ViT-g/14)	CLIP [45]	76.2	70.1	88.9	77.2	72.3	-	-
	ALIGN [30]	76.4	70.1	92.2	75.8	-	-	-
	LiT	**84.5**	**78.7**	**93.9**	**79.4**	**81.1**	88.0	72.6
Public (ViT-L/16)	CLIP [45]	31.3	-	-	-	-	-	-
	OpenCLIP [28]	34.8	30.0	-	-	-	-	-
	LiT	**75.7**	**66.6**	60.4	37.8	54.5	82.1	63.1
*	ResNet50 [25]	75.8	63.8	36.1	0.5	26.5	82.5	72.6

Efficiency



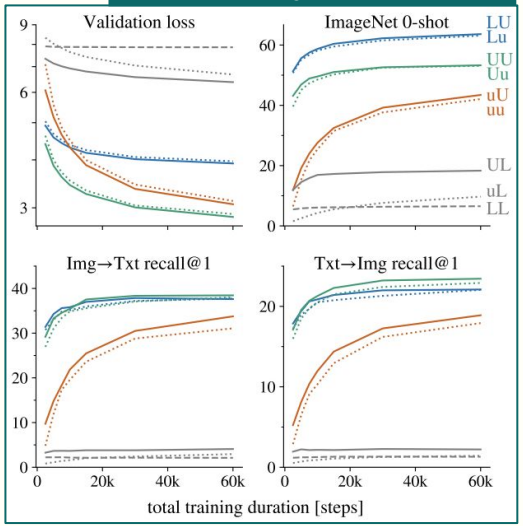
Table, Image: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Design Choices

Full dataset (3.6B pairs)

Method	ImgNet	ImgNet-v2	Cifar100	Pets
Lu	70.1	61.7	70.9	88.1
Uu	57.2	50.2	62.1	74.8
uu	50.6	43.3	47.9	70.3

YFCC100m$_{CLIP}$ (15M pairs)



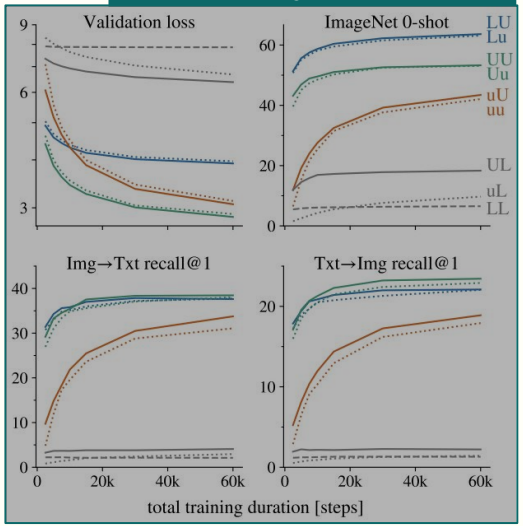
Table, Image: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Design Choices

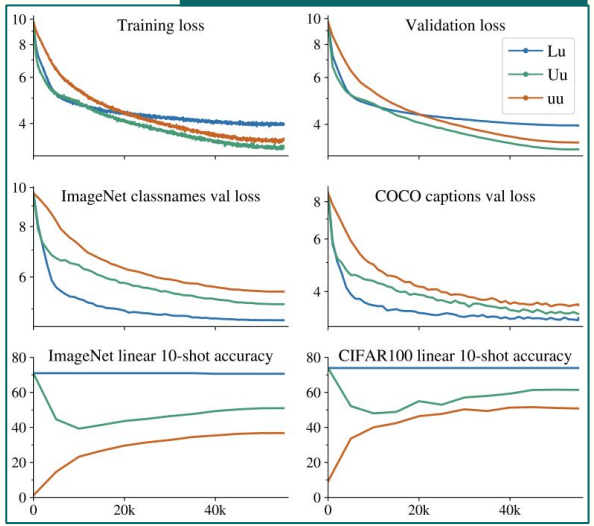
Full dataset (3.6B pairs)

Method	ImgNet	ImgNet-v2	Cifar100	Pets
Lu	70.1	61.7	70.9	88.1
Uu	57.2	50.2	62.1	74.8
uu	50.6	43.3	47.9	70.3

YFCC100m$_{CLIP}$ (15M pairs)



Out-of-distribution datasets



Table, Images: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Scope

Table: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Supervision

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Similar performance

Table: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Good performance on narrow task; does not generalize well

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Model Specificity

Differently pretrained image encoders

Model: ViT-B/16	Pre-training				LiT		
	Dataset	Labels?	Full IN	10-shot	0-shot	I→T	T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Transformer vs. other architectures

Model	0shot	Adapt	I→T	T→I	Param	Speed	FLOPs
ViT-B/32	60.7	79.1	41.3	25.0	197 M	2855	12 G
Mixer-B/32	57.1	75.9	37.5	22.9	169 M	4208	9 G
BiT-M-R50	55.2	77.6	37.3	23.9	134 M	2159	11 G

Tables: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Text Models

	Model	Tok	INet 0shot	I→T	T→I
YFCC-CLIP	ViT	SP	57.2	29.7	16.9
	T5	SP	57.8 (+1.4)	29.4 (+1.6)	17.2 (+1.2)
	mT5	SP	58.1 (+1.2)	28.3 (+0.4)	16.4 (+1.0)
	BERT	WP	**58.8** (+0.7)	**35.2** (+1.1)	**20.0** (+0.7)
	ViT	WP	56.4	28.2	17.3
Ours	ViT	SP	68.8	43.6	28.5
	ViT	WP	68.8	45.4	29.7
	BERT	WP	65.8	43.8	28.6

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Text Models

	Model	Tok	INet 0shot	I→T	T→I
YFCC-CLIP	ViT	SP	57.2	29.7	16.9
	T5	SP	57.8 (+1.4)	29.4 (+1.6)	17.2 (+1.2)
	mT5	SP	58.1 (+1.2)	28.3 (+0.4)	16.4 (+1.0)
	BERT	WP	**58.8** (+0.7)	**35.2** (+1.1)	**20.0** (+0.7)
	ViT	WP	56.4	28.2	17.3
Ours	ViT	SP	68.8	43.6	28.5
	ViT	WP	68.8	45.4	29.7
	BERT	WP	65.8	43.8	28.6

Datasets

Table: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Text Models

	Model	Tok	INet 0shot	I→T	T→I
YFCC-CLIP	ViT	SP	57.2	29.7	16.9
	T5	SP	57.8 (+1.4)	29.4 (+1.6)	17.2 (+1.2)
	mT5	SP	58.1 (+1.2)	28.3 (+0.4)	16.4 (+1.0)
	BERT	WP	**58.8** (+0.7)	**35.2** (+1.1)	**20.0** (+0.7)
	ViT	WP	56.4	28.2	17.3
Ours	ViT	SP	68.8	43.6	28.5
	ViT	WP	68.8	45.4	29.7
	BERT	WP	65.8	43.8	28.6

Tokenization

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Text Models

	Model	Tok	INet 0shot	I→T	T→I
YFCC-CLIP	ViT	SP	57.2	29.7	16.9
	T5	SP	57.8 (+1.4)	29.4 (+1.6)	17.2 (+1.2)
	mT5	SP	58.1 (+1.2)	28.3 (+0.4)	16.4 (+1.0)
	BERT	WP	**58.8** (+0.7)	**35.2** (+1.1)	**20.0** (+0.7)
	ViT	WP	56.4	28.2	17.3
Ours	ViT	SP	68.8	43.6	28.5
	ViT	WP	68.8	45.4	29.7
	BERT	WP	65.8	43.8	28.6

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: De-duplication

Dedup	#tune	#eval	ImgNet	I→T	T→I
-	0	0	70.2	43.6	28.4
test	2.6M	76K	70.2	43.3	28.3
train+test	3.6M	220K	69.9	43.7	28.4

Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Locked Image Benefits

- No gradients for image
- → Reduced memory, faster training
- If no augmentation → features need to be computed only once

Feature precomputation

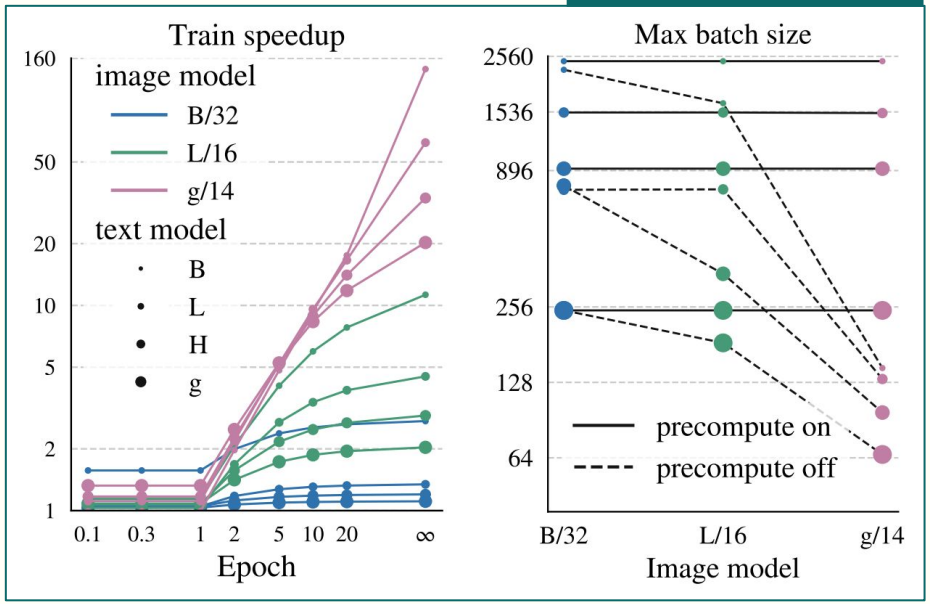
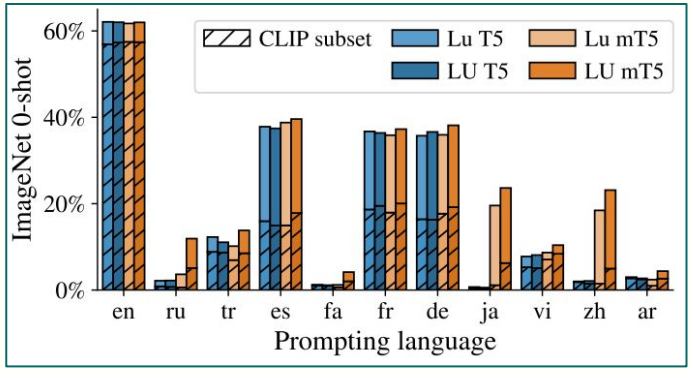


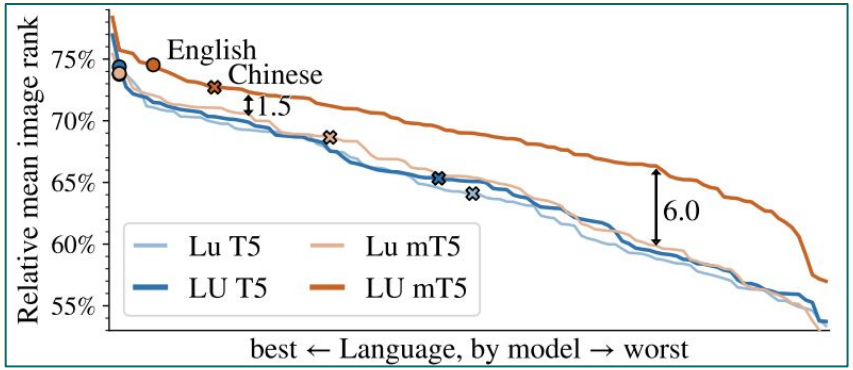
Table: Zhai, Xiaohua, et al. "LiT🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Experiments: Multilinguality

- Translates ImageNet prompts into most popular languages
- Perform zero-shot evaluation



- Perform T→I retrieval on 100+ languages using Wikipedia-based Image Text (WIT; Srinivasan ‘21)



Images: Zhai, Xiaohua, et al. “LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning.” *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Strengths

- Stabler training
  - Updates only text encoder instead of two at once
- Efficiency
  - Pretrained weights vs. trained from scratch
- Flexibility
  - Can easily switch out pretrained encoders with minimal overhead
- More challenging pretraining datasets
  - CC12M, YFCC100m, noisy LiT🔥 dataset
- Potentially stronger latents than CLIP

Weaknesses

- Failed to address some of CLIP's weaknesses, *e.g.*,
  - Image regions,
  - Reasoning tasks (*e.g.*, VQA), or
  - Abstract tasks (*e.g.*, counting), as reported in paper

VTAB zero-shot performance (slide)

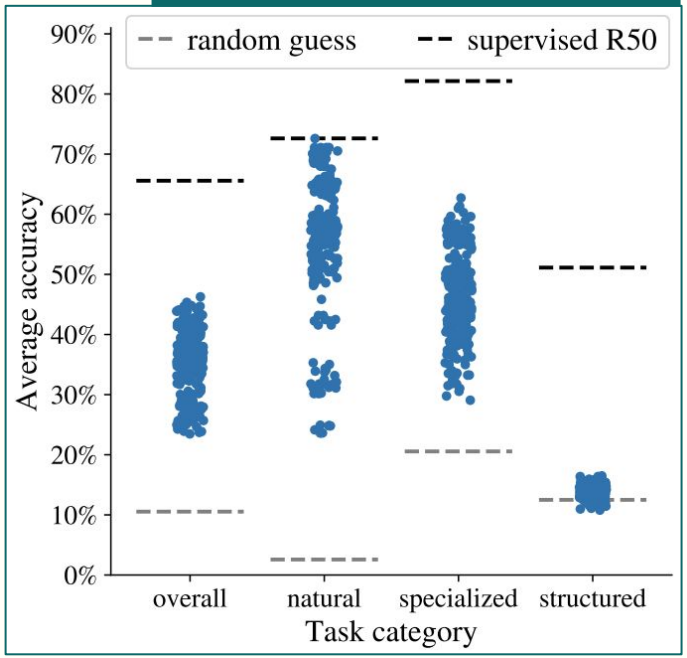


Image: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Weaknesses

- Unfair evaluation of zero-shot performance

LiT🔥 performance report (Zhai '22)

Dataset	Method	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
Private	CLIP [45]	76.2	70.1	88.9	77.2	72.3	-	-
	ALIGN [30]	76.4	70.1	92.2	75.8	-	-	-
	LiT	**84.5**	**78.7**	**93.9**	**79.4**	**81.1**	88.0	72.6
Public	CLIP [45]	31.3	-	-	-	-	-	-
	OpenCLIP [28]	34.8	30.0	-	-	-	-	-
	LiT	**75.7**	**66.6**	60.4	37.8	54.5	82.1	63.1
*	ResNet50 [25]	75.8	63.8	36.1	0.5	26.5	82.5	72.6

ViT-L/16

Pretrained on YFCC100m$_{CLIP}$ + CC12M

CLIP performance report (Radford '21)

Dataset	*Linear Classifier* YFCC	WIT	Δ	*Zero Shot* YFCC	WIT	Δ
Birdsnap	47.4	35.3	+12.1	19.9	4.5	+15.4
Country211	23.1	17.3	+5.8	5.2	5.3	+0.1
Flowers102	94.4	89.8	+4.6	48.6	21.7	+26.9
GTSRB	66.8	72.5	−5.7	6.9	7.0	−0.1
UCF101	69.2	74.9	−5.7	22.9	32.0	−9.1
Stanford Cars	31.4	50.3	−18.9	3.8	10.9	−7.1
ImageNet	**62.0**	60.8	+1.2	**31.3**	27.6	+3.7
Dataset Average	65.5	**66.6**	−1.1	29.6	**30.0**	−0.4
Dataset "Wins"	10	**15**	−5	**19**	18	+1

Table 12. **CLIP performs similarly when trained on only YFCC100M.** Comparing a ResNet-50 trained on only YFCC100M with a same sized subset of WIT shows similar average performance and number of wins on zero shot and linear classifier evals.

Tables: From linked articles.

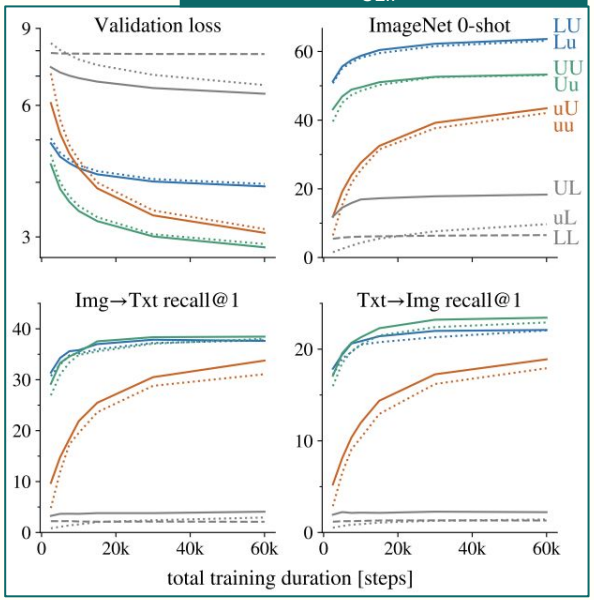
Weaknesses

- Insufficient evaluation on `LU` setting
  - Better performance on YFCC100m$_{CLIP}$, but no evaluation on full dataset

Full dataset (3.6B pairs)

Method	ImgNet	ImgNet-v2	Cifar100	Pets
Lu	70.1	61.7	70.9	88.1
Uu	57.2	50.2	62.1	74.8
uu	50.6	43.3	47.9	70.3

YFCC100m$_{CLIP}$ (15M pairs)



Table, Image: Zhai, Xiaohua, et al. "LiT 🔥: Zero-Shot Transfer with Locked-Image Text Tuning." *Proceedings of the IEEE/CVF Conference on CVPR*. 2022.

Future Work

- Use LiT🔥latents for image generation
- Add intra-modal contrastive objectives
- Apply same locked tuning idea to video domain

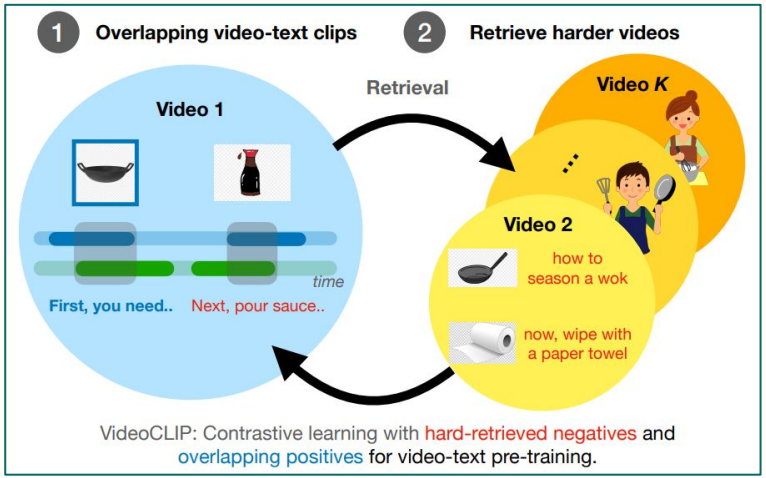


Image: Xu, Hu, et al. "VideoCLIP: Contrastive Pre-Training for Zero-Shot Video-Text Understanding." *arXiv preprint arXiv:2109.14084* (2021).

Thank you!

Discussion section follows.

Discussion

- LU setting on full dataset
- Tackling abstract tasks (*e.g.*, counting, depth, distance)
- Tackling reasoning tasks (*e.g.*, VQA)
- How to approach supervised classification SotA?
  - Scale?