Recent Advances in Vision Foundation Models

Zhe Gan

Apple AI/ML

6/19/2023

A little bit history: from VQA to VLP, from pandemic back to normal

2020

1:15 - 1:25	**Opening Remarks** presented by JJ Liu and Xiaodong He (Slides , YouTube , Bilibili)
1:25 - 2:15	**Visual QA and Reasoning** presented by Zhe Gan (Slides , YouTube , Bilibili)
2:15 - 2:30	**Coffee Break**
2:30 - 3:10	**Visual Captioning** presented by Luowei Zhou (Slides , YouTube , Bilibili)
3:10 - 3:40	**Text-to-image Synthesis** presented by Yu Cheng (Slides , YouTube , Bilibili)
3:40 - 4:00	**Coffee Break**
4:00 - 5:00	**Self-supervised Learning** presented by Licheng Yu , Linjie Li and Yen-Chun Chen (Slides

2021

Prerecorded Sessions	
4min	**Opening Remarks** [Video]
50min	**Representations and Training Strategies for VLP** [Video] [Slides]
40min	**Robustness, Efficiency and Extensions for VLP** [Video] [Slides]
40min	**Video-and-Language Pre-training** [Video] [Slides]
42min	**Introduction to VLN** [Video] [Slides]
55min	**Generalizable VLN Methods** [Video] [Slides]
58min	**Forward to Realistic VLN** [Video] [Slides]
15min	**VLN Summary** [Video] [Slides]
Live Session	
16:00-17:00	**Panel Discussion** ((o)) LIVE on Zoom [Video]

2022

Morning Session	
9:00 - 9:15	**Opening Remarks** [Bilibili, YouTube]
9:15 - 10:00	**Overview of Image-Text Pre-training** [Slides] [Bilibili, YouTube]
10:00 - 10:15	**Coffee Break & QA**
10:15 - 11:00	**Unified Image-Text Modeling** [Slides] [Bilibili, YouTube]
11:00 - 11:45	**Advanced Topics in Image-Text Pre-training** [Slides] [Bilibili, YouTube]
11:45 - 12:00	**Q & A**
Afternoon Session	
13:00 - 13: 30	**Overview of Video-Text Pre-training** [Slides] [Bilibili, YouTube]
13:30 - 14:00	**Learning from Multi-channel Videos: Methods and Benchmarks** [Slides] [Bilibili, YouTube]
14:00 - 14: 30	**Advanced Topics in Video-Text Pre-training** [Slides] [Bilibili, YouTube]
14:30 - 14:45	**Coffee Break & QA**
14:45 - 15: 15	**VLP for Image Classification** [Slides] [Bilibili, YouTube]
15:15 - 15:45	**VLP for Object Detection** [Slides] [Bilibili, YouTube]
15:45 - 16:15	**Benchmarks for Computer Vision in the Wild** [Slides] [Bilibili, YouTube]
16:15 - 17:00	**VLP for Text-to-Image Synthesis** [Slides] [Bilibili, YouTube]
17:00 - 17:15	**Q & A**

Check out our survey paper

Vision-Language Pre-training: Basics, Recent Advances, and Future Trends

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao
Microsoft Corporation
{zhgan,linjli,chunyl,lijuanw,zliu,jfgao}@microsoft.com

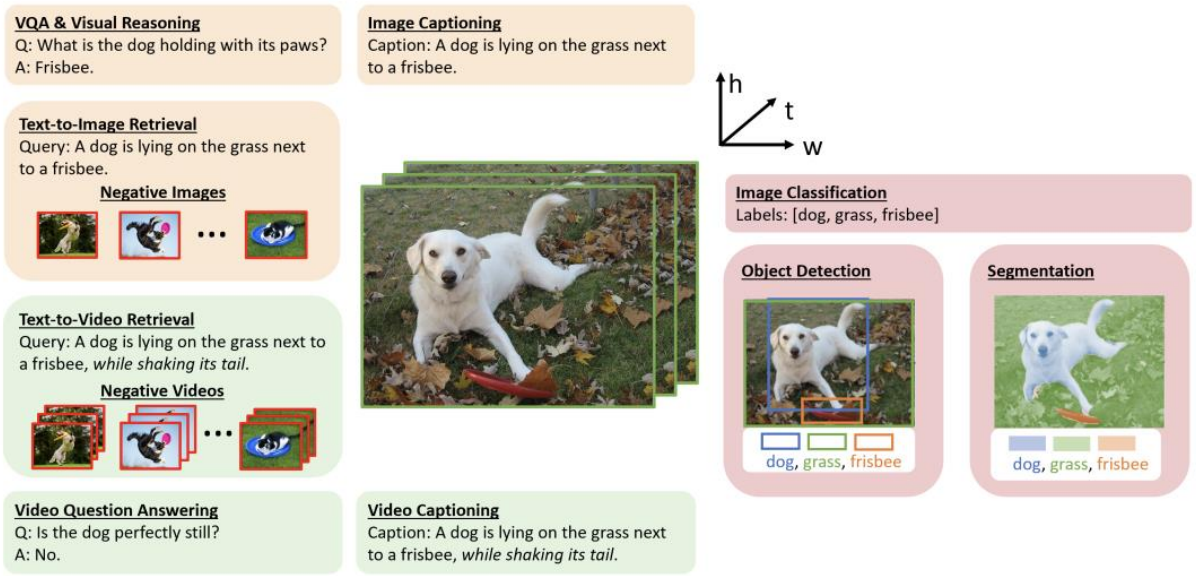
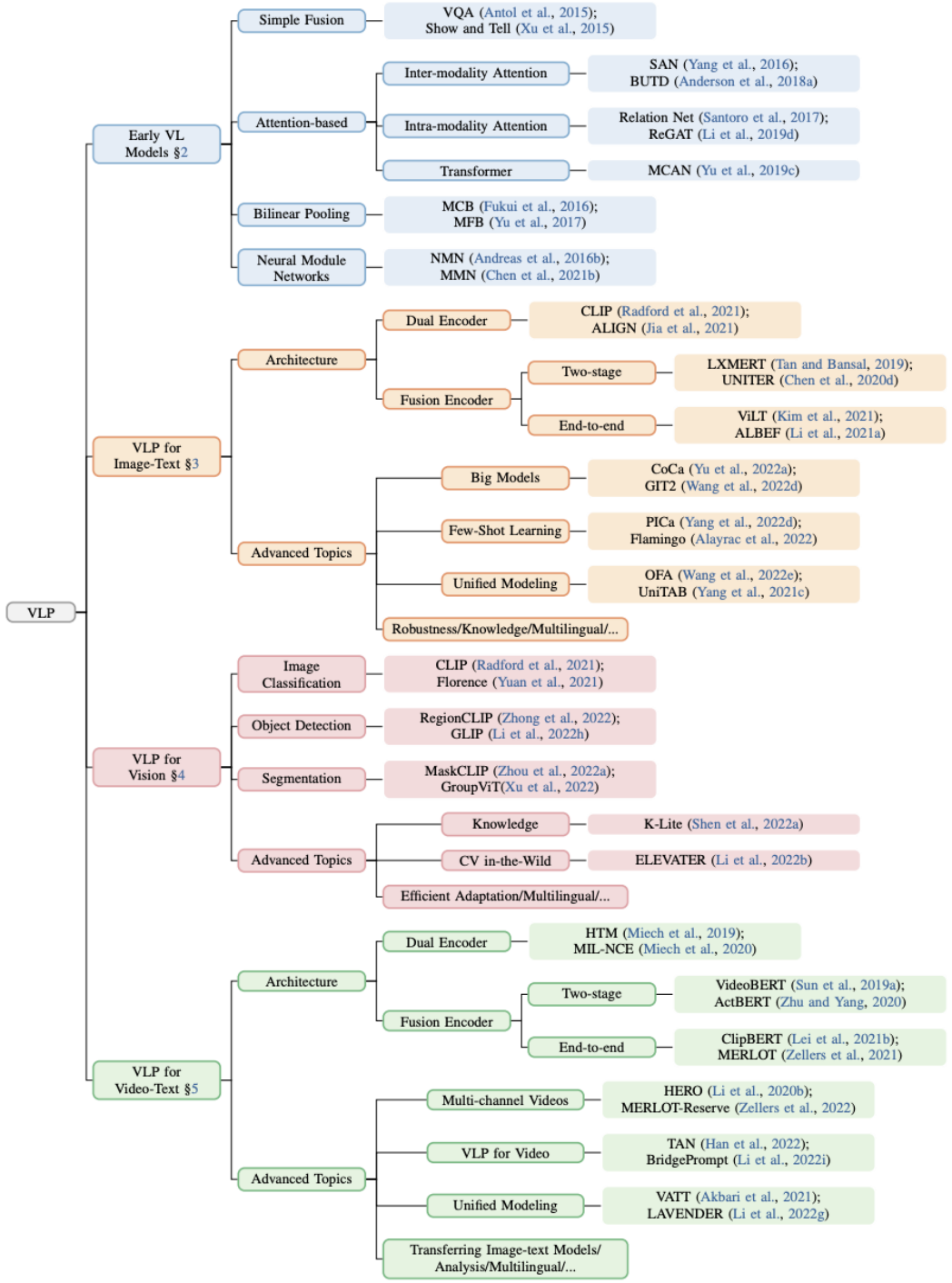


Figure 1.2: Illustration of representative tasks from three categories of VL problems covered in this paper: image-text tasks, vision tasks as VL problems, and video-text tasks.



Brand new design for this year's tutorial

- Things are evolving quickly …

- From the LLM perspective, now we have
  - ChatGPT, GPT-4 from OpenAI, PaLM, Bard from Google, LLaMA from Meta
  - Alpaca, Vicuna, etc. from the open-source community
  - And other LLMs from many startups

- From the computer vision perspective, now we have
  - SAM, DINOv2, Stable Diffusion, Midjourney, etc.
  - LLaVA, MiniGPT-4, etc.
  - Visual ChatGPT, MMReACT, etc.

- So, what's new this year?

So, what's new this year?

( , a dog is running through the grass)

Q3: how to do image generation?

Image Generation

Produce visual data

LLM for language understanding and generation

General-purpose interface

Q4: how to train multimodal LLM?
Q5: how to chain multimodal experts with LLM?

Image Encoder

Consume visual data

Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?

Agenda today

- Part I: Visual and Vision-Language Pre-training
  - To consume visual data, how to learn a strong image backbone
- Part II: Towards Generic Vision Interface
  - How to design vision interface that is interactive and promptable
- Part III: Text-to-Image Generation
  - How to produce visual data that is also aligned with human intent
- Part IV: Multimodal LLM
  - How to make an LLM that can see and chat
- Part V: Multimodal Agents
  - How to chain multimodal experts with LLM to unlock new capabilities

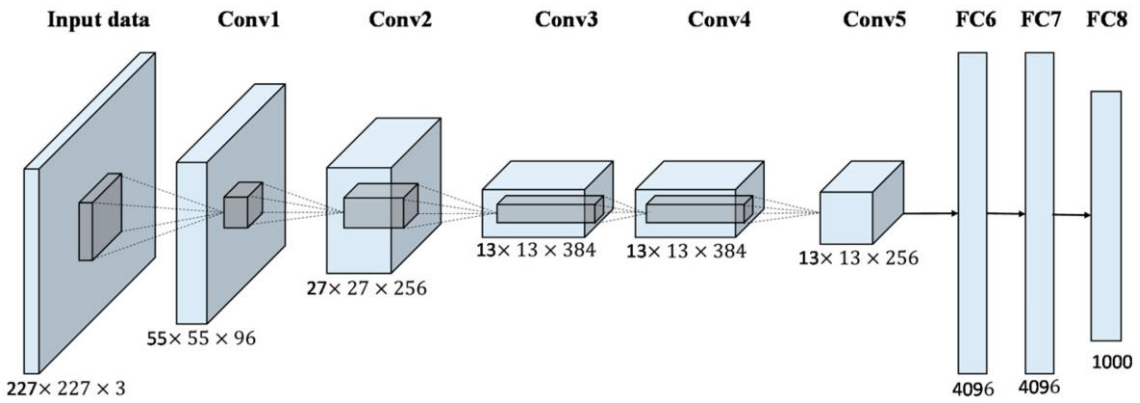




Agenda today

- Part I: Visual and Vision-Language Pre-training
  - To consume visual data, how to learn a strong image backbone
- Part II: Towards Generic Vision Interface
  - How to design vision interface that is interactive and promptable
- Part III: Text-to-Image Generation
  - How to produce visual data that is also aligned with human intent
- Part IV: Multimodal LLM
  - How to make an LLM that can see and chat
- Part V: Multimodal Agents
  - How to chain vision experts with LLM to unlock new capabilities

Supervised Learning



Contrastive Language-Image Pre-training

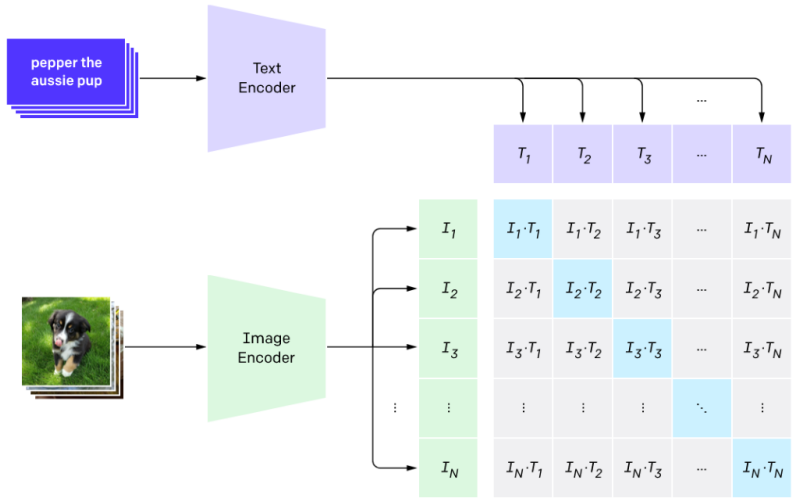
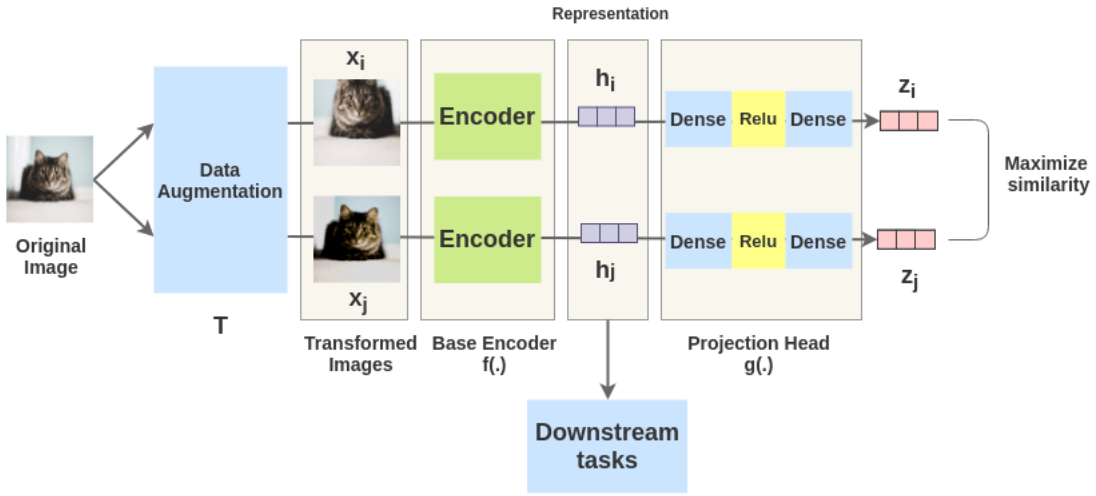
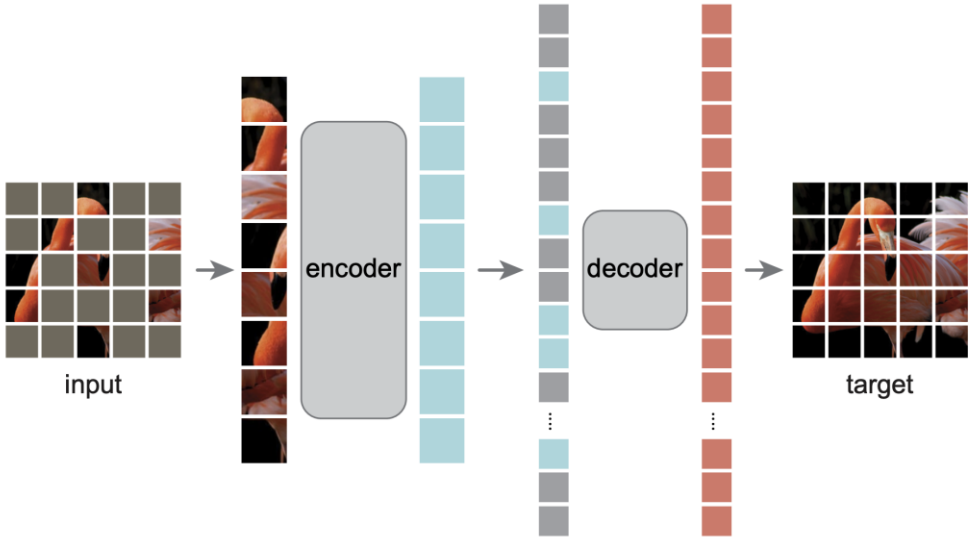


Image-only (Non-)Contrastive Learning



Masked Image Modeling

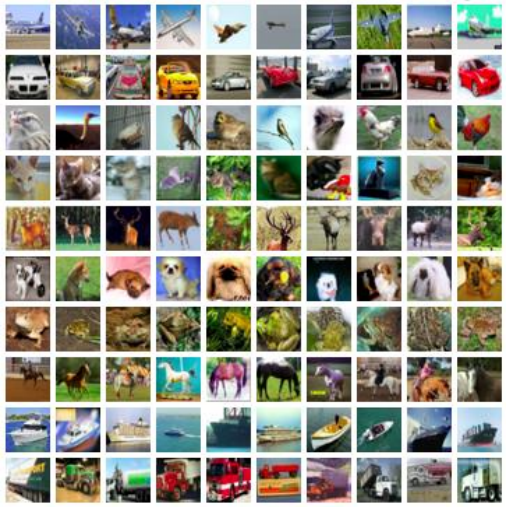


Note that there is a vast amount of literature on this topic. Due to time limit, in this tutorial, we will use CLIP as the anchor point, and only select papers based on our own preference and judgement (and surely with our own bias). ☺

Supervised learning

- Mapping an image to a *discrete label* which is associated to a visual concept
- Human annotation is expensive, and the labels can be limited
- *Private* datasets created by industrial labs:
  - JFT-300M, JFT-3B[1], IG-3.6B[2] (called weakly-supervised pre-training in this case)
  - Noisy weak supervision, can be very powerful for learning universal image embeddings

MNIST



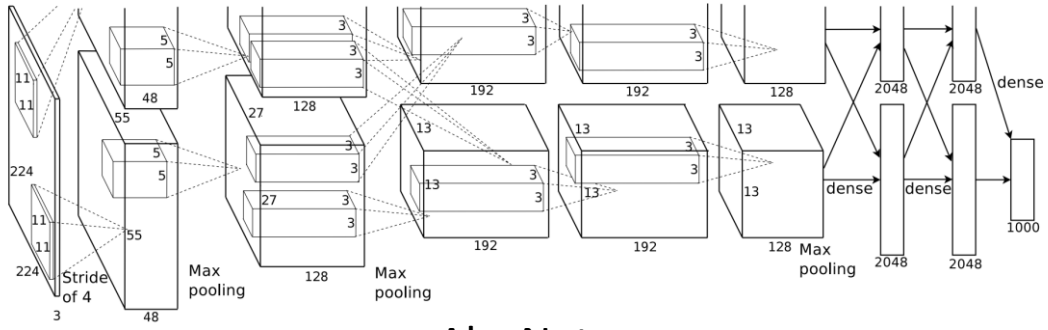
CIFAR-10

ImageNet

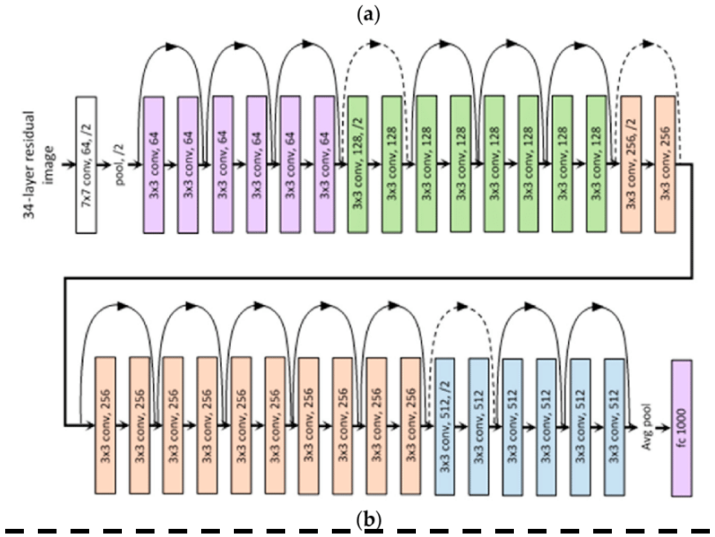
[1] Scaling vision transformers, CVPR 2022
[2] Revisiting weakly supervised pre-training of visual perception models, CVPR 2022

Supervised learning

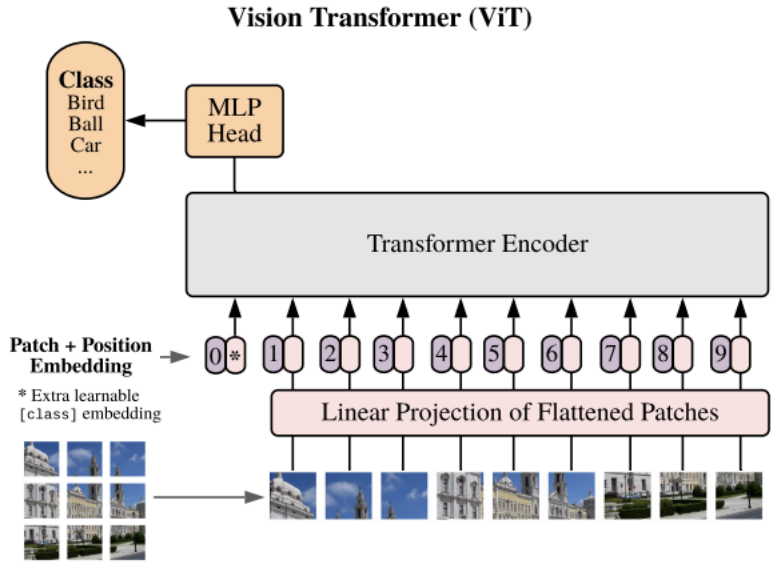
- Powered architectures ranging from AlexNet, ResNet, ViT, to Swin, and all the modern vision backbones



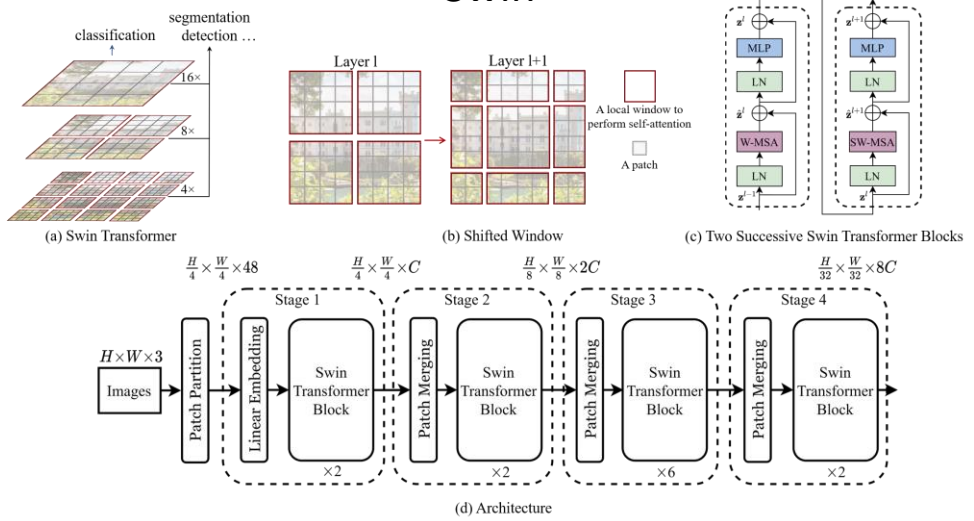
AlexNet



ResNet

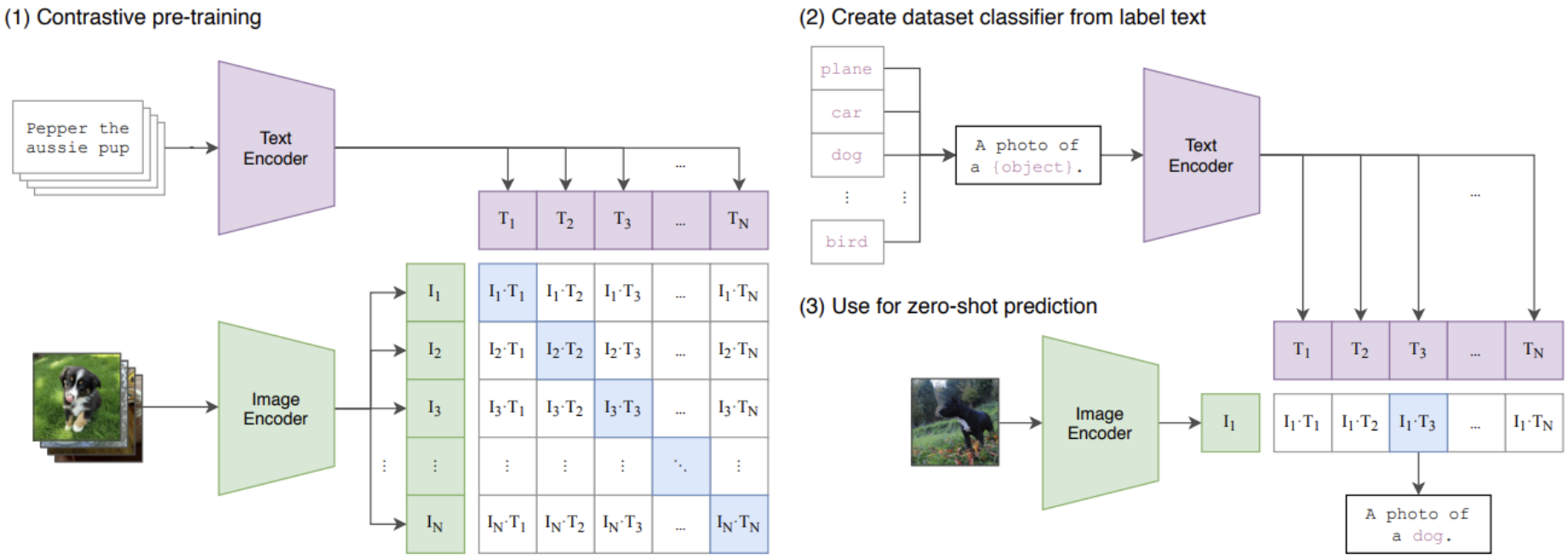


Swin



Contrastive language-image pre-training

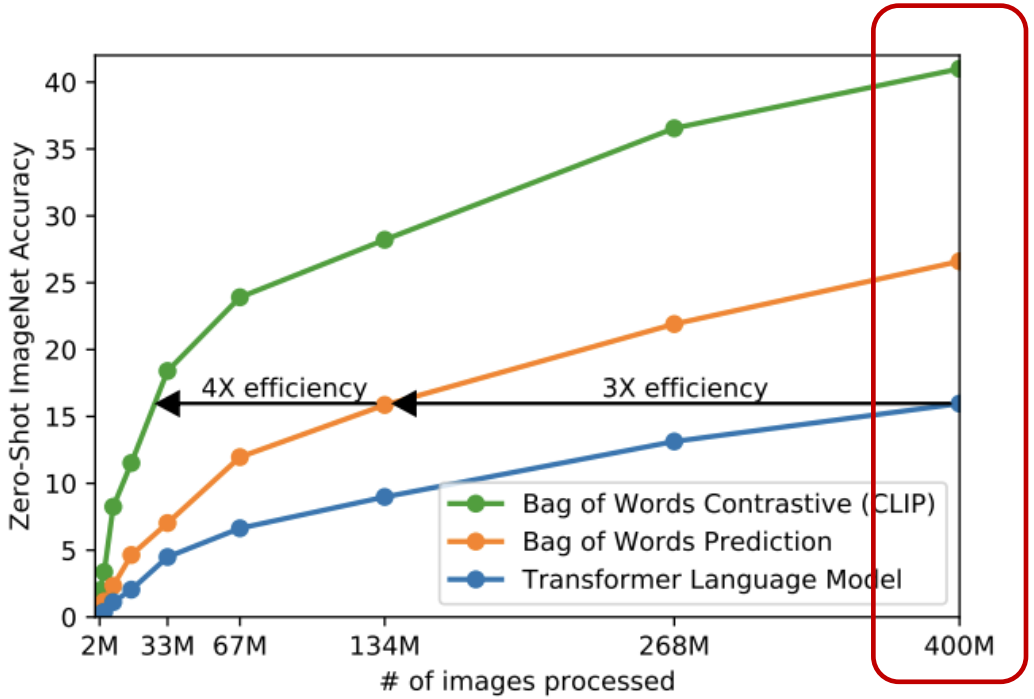
- Learning image representations from web-scale noisy text supervision
  - Training: simple *contrastive* learning, and the beauty lies in large-scale pre-training
  - Downstream: *zero-shot* image classification and image-text retrieval
    - Image classification can be reformatted as a retrieval task via considering the semantics behind label names



[1] Learning transferable visual models from natural language supervision, ICML 2021
[2] Scaling up visual and vision-language representation learning with noisy text supervision, ICML 2021

Contrastive language-image pre-training

- The idea is simple, and can be dated back to a long while ago
  - In the large-scale pre-training era: CLIP[1] and ALIGN[2]
  - *Data scale* matters: Models are frequently trained with billions of image-text pairs
  - *Batch size* matters: 32k by default; *Model size* matters



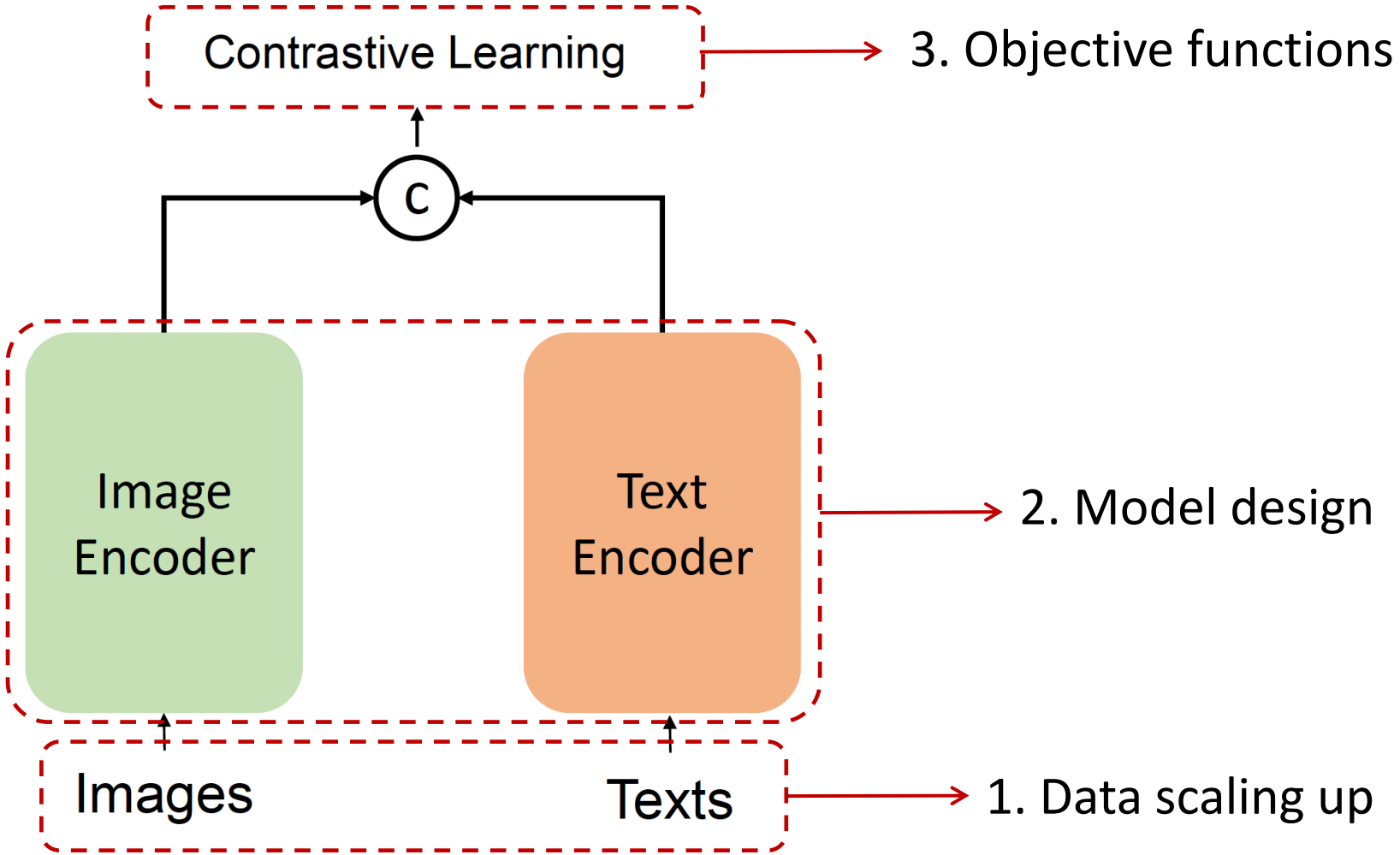
> Language is a stronger form of supervision than classical closed-set labels. Language provides rich information for supervision. Therefore, *scaling*, which can involve increasing capacity (model scaling) and increasing information (data scaling), is essential for attaining good results in language-supervised training.
>
> CLIP [52] is an outstanding example of "*simple algorithms that scale well*". The simple design of CLIP allows it to be relatively easily executed at substantially larger scales and achieve big leaps compared to preceding methods. Our method largely maintains the simplicity of CLIP

Quote from the FLIP paper

[1] Learning transferable visual models from natural language supervision, ICML 2021
[2] Scaling up visual and vision-language representation learning with noisy text supervision, ICML 2021

How to improve CLIP

- Since the birth of CLIP, tons of follow-up works and applications

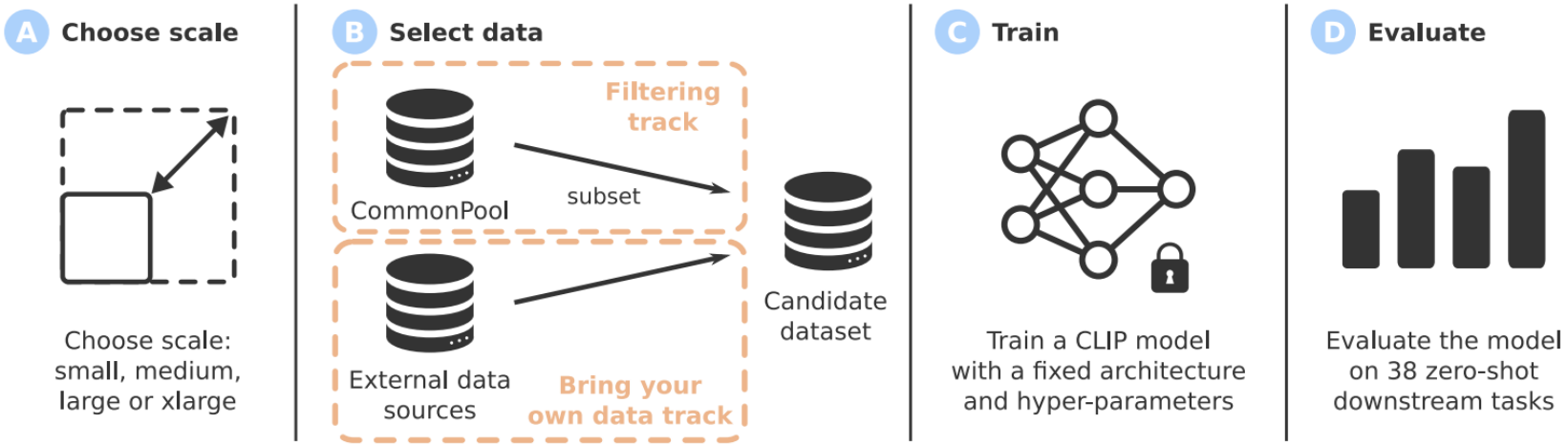


Data scaling up

- Reproducible scaling laws for CLIP training
  - Open large-scale LAION-2B dataset
  - Pre-training OpenCLIP across various scales

	Data	Arch.	ImageNet	VTAB+	COCO
CLIP [55]	WIT-400M	L/14	75.5	55.8	61.1
Ours	LAION-2B	L/14	75.2	54.6	71.1
Ours	LAION-2B	H/14	78.0	56.4	73.4

- DataComp: We know scale matters, how to further scale it up
  - In search of the next-generation image-text datasets
  - Instead of fixing the dataset, and designing different algorithms, the authors propose to fix the CLIP training method, but select the datasets instead



[1] Reproducible scaling laws for contrastive language-image learning, CVPR 2023
[2] Datacomp: In search of the next generation of multimodal datasets, 2023

Model design: from the image side

- FLIP: Scaling CLIP training via masking
  - Training: still use CLIP loss, without incorporating the MIM loss
  - Trick: randomly masking out image patches with a high masking ratio, and only encoding the visible patches
  - Results: turns out this does not hurt performance, but improves training efficiency
    - Training is done in 256 TPU-v3 cores, with LAION-400M for 6.4, 12.8, or 32 epochs

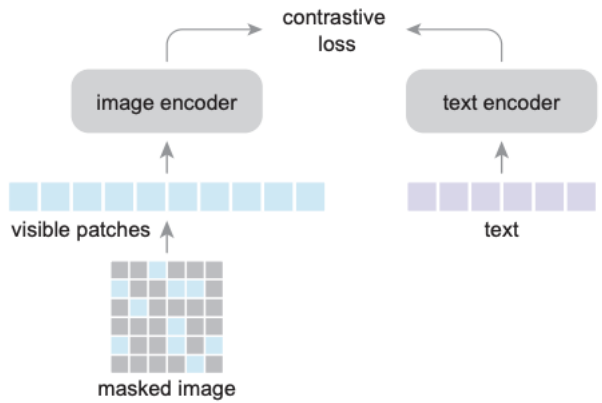
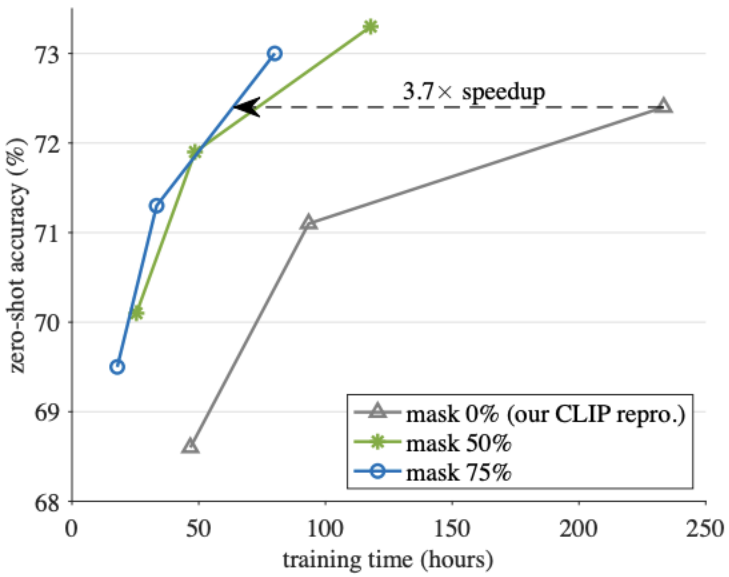


Figure 2. **Our FLIP architecture**. Following CLIP [52], we per-
form contrastive learning on pairs of image and text samples. We
randomly mask out image patches with a high masking ratio and
encode only the visible patches. We do not perform reconstruction

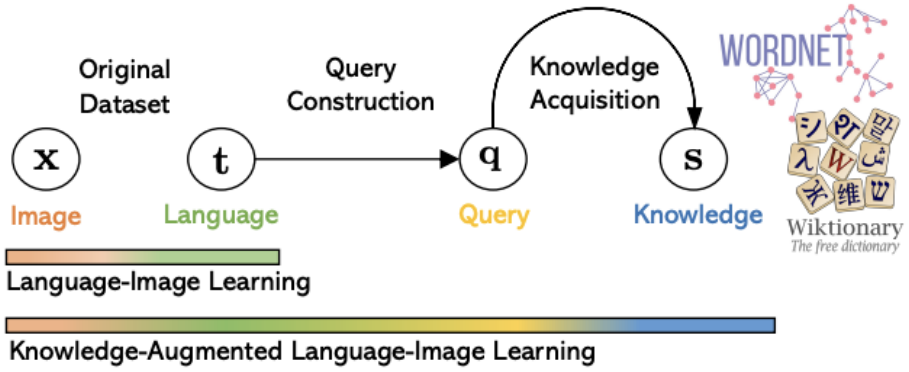


[1] Scaling language-image pre-training via masking, CVPR 2023

Model design: from the language side

- K-Lite: External knowledge
  - The Wiki definition of entities (or, the so-called knowledge) can be naturally used together with the original alt-text for contrastive pre-training

Figure 1: Motivating examples: knowledge explains the content of the rare dish concepts.



Enriching alt-text with entity descriptions enhances performance.

Training Data		Method	ImageNet-1K	ICinW (20 datasets)		
Dataset	# Samples		Zero-shot	Zero-shot	Linear Probing	Fine-tuning
ImageNet-21K	13M (full)	UniCL	28.16	27.15	53.07 ± 4.15	55.96 ± 2.50
	13M (full)	K-LITE	**30.23**	**33.44**	**53.92** ± 1.05	**57.81** ± 1.48
YFCC-14M + ImageNet-21K	14M (half)	UniCL	34.43	34.30	53.50 ± 2.22	56.45 ± 2.48
	14M (half)	K-LITE	36.67	36.50	49.48 ± 2.23	55.88 ± 1.64
	14M (half)	K-LITE◇	42.36	36.50	54.28 ± 3.66	52.11 ± 4.90
	27M (full)	UniCL	43.06	35.99	55.96 ± 3.38	58.25 ± 2.98
	27M (full)	K-LITE	**45.67**	**38.89**	**57.06** ± 1.48	58.24 ± 2.36
GCC-15M + ImageNet-21K	15M (half)	UniCL	41.64	36.31	53.86 ± 2.73	59.04 ± 3.13
	15M (half)	K-LITE	44.26	39.53	55.91 ± 2.53	58.20 ± 3.39
	15M (half)	K-LITE◇	47.30	40.32	57.38 ± 2.70	60.72 ± 2.29
	28M (full)	UniCL	46.83	38.90	57.92 ± 3.31	60.99 ± 2.74
	28M (full)	K-LITE	**48.76**	**41.34**	**58.56** ± 3.12	**63.39** ± 1.74

[1] K-lite: Learning transferable visual models with external knowledge, NeurIPS 2022

Model design: improved interpretability

- STAIR: Learning Sparse Text and Image Representation in Grounded Tokens
  - Mapping images and text to a high-dim sparse embedding space
  - Each dimension in the sparse embedding is a (sub-)word in a large dictionary in which the predicted non-negative scalar corresponds to the weight associated with the token
  - Better performance than CLIP with improved interpretability

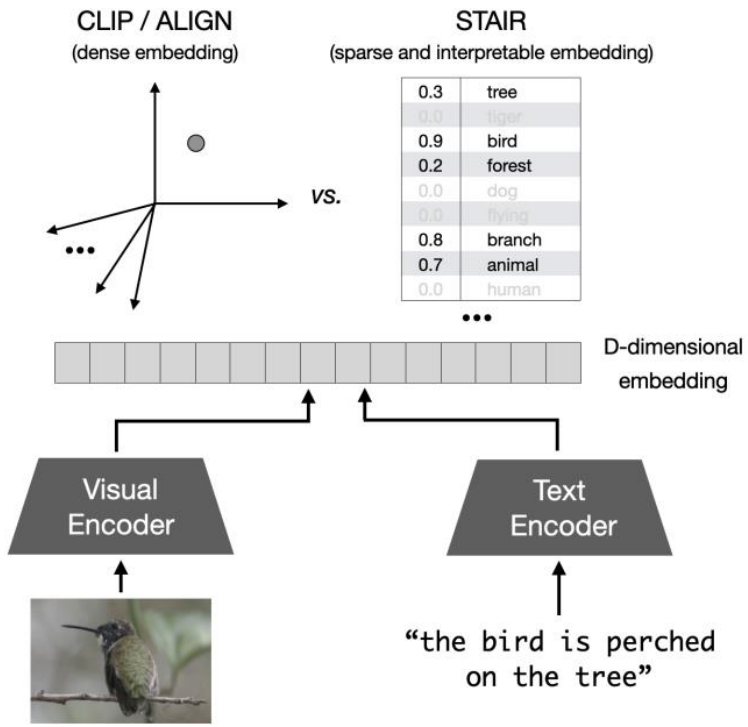


Table 1. **Zero-shot text/image retrieval.** Reporting recall@K on Flickr30K and COCO.

	COCO 5K						Flickr30K					
	text → image			image → text			text → image			image → text		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	36.2	62.2	72.2	53.4	78.3	85.6	63.0	86.7	92.5	79.6	95.5	98.1
STAIR	**41.1**	**65.4**	**75.0**	**57.7**	**80.5**	**87.3**	**66.6**	**88.7**	**93.5**	**81.2**	**96.1**	**98.4**

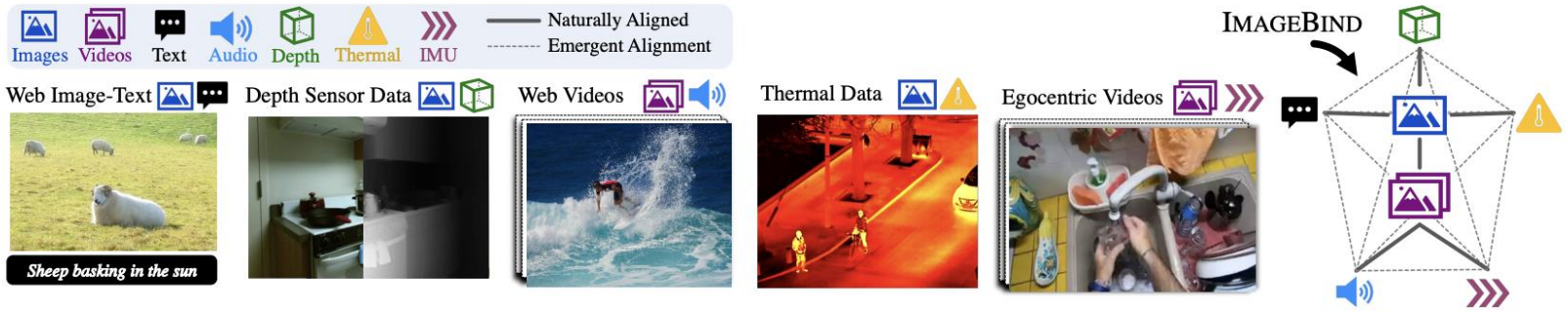
Original Caption: A seagull standing on the sand of a beach.

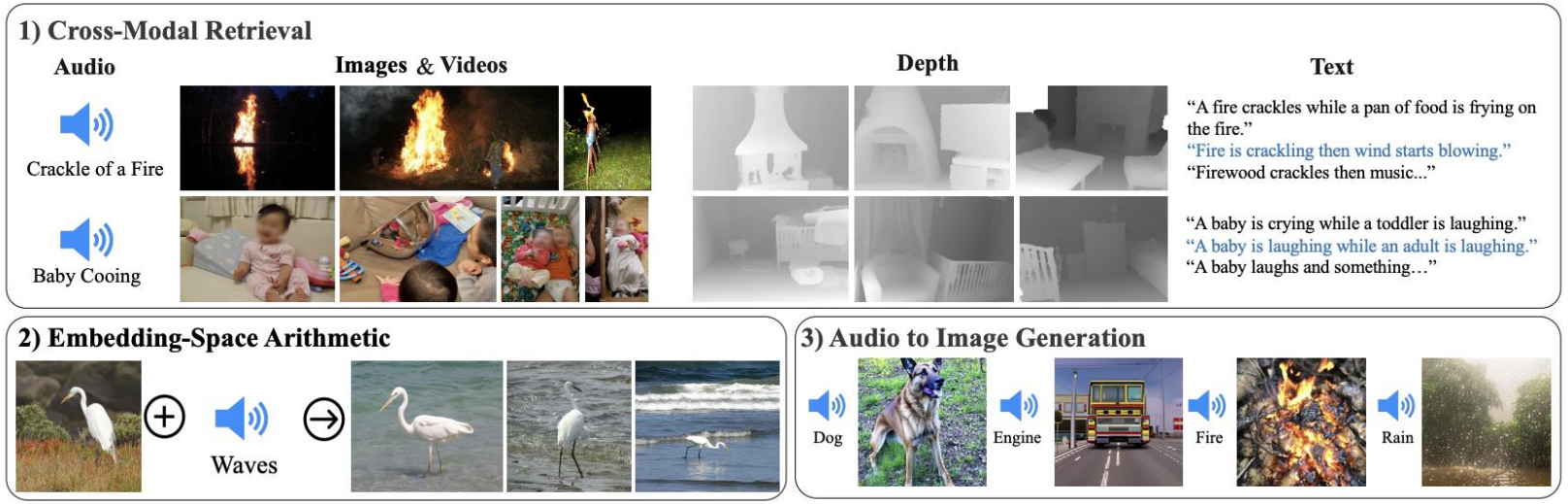
Bride and grooms arms cutting the wedding cake with fruit on top.

[1] STAIR: Learning Sparse Text and Image Representation in Grounded Tokens, 2023

Model design: more modalities

- ImageBind: One embedding space to bind them all
  - Linking all modalities (7 in this paper) into a common space
  - A pre-trained CLIP is used and kept frozen, i.e., learning other modality encoders to align the CLIP embedding space

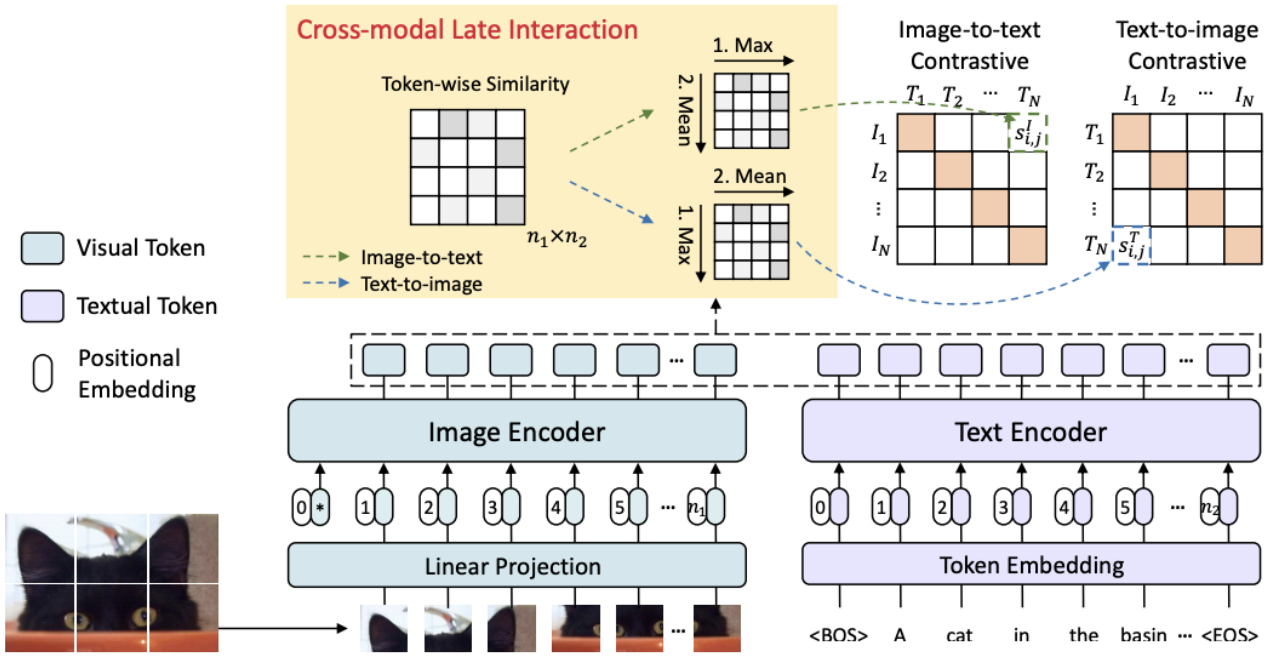


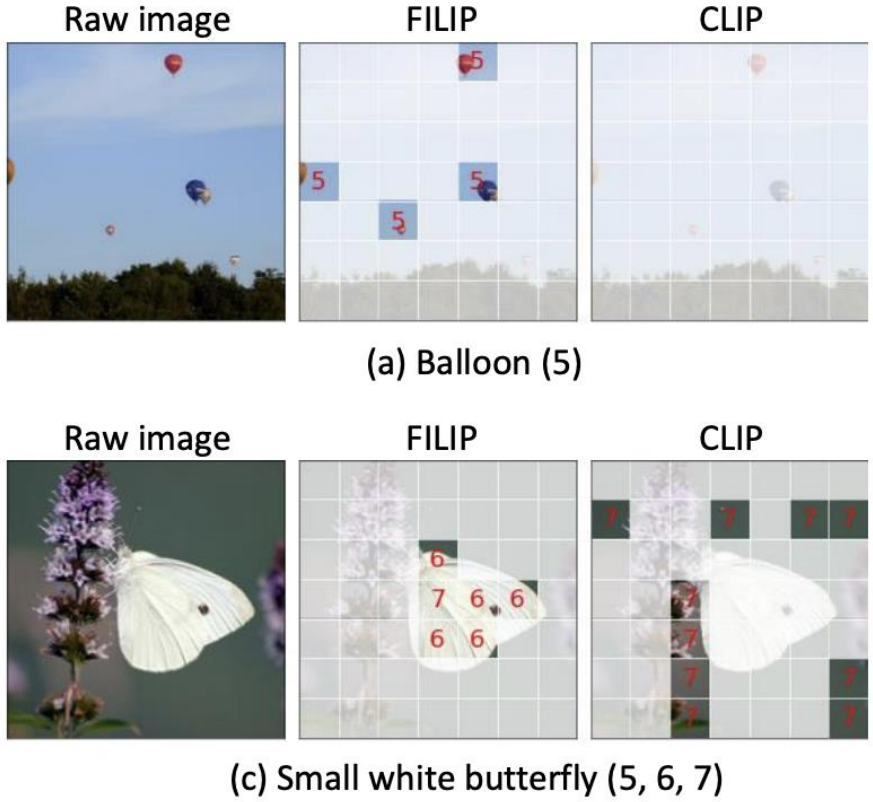


[1] IMAGEBIND: One Embedding Space To Bind Them All, 2023

Objective function: fine-grained supervision

- FILIP: Fine-grained supervision
  - Still dual encoder, not a fusion encoder
  - But compute the loss by first computing the token-wise similarity, and then aggregating the matrix by max pooling
  - Learns word-patch alignment that is good for visualization

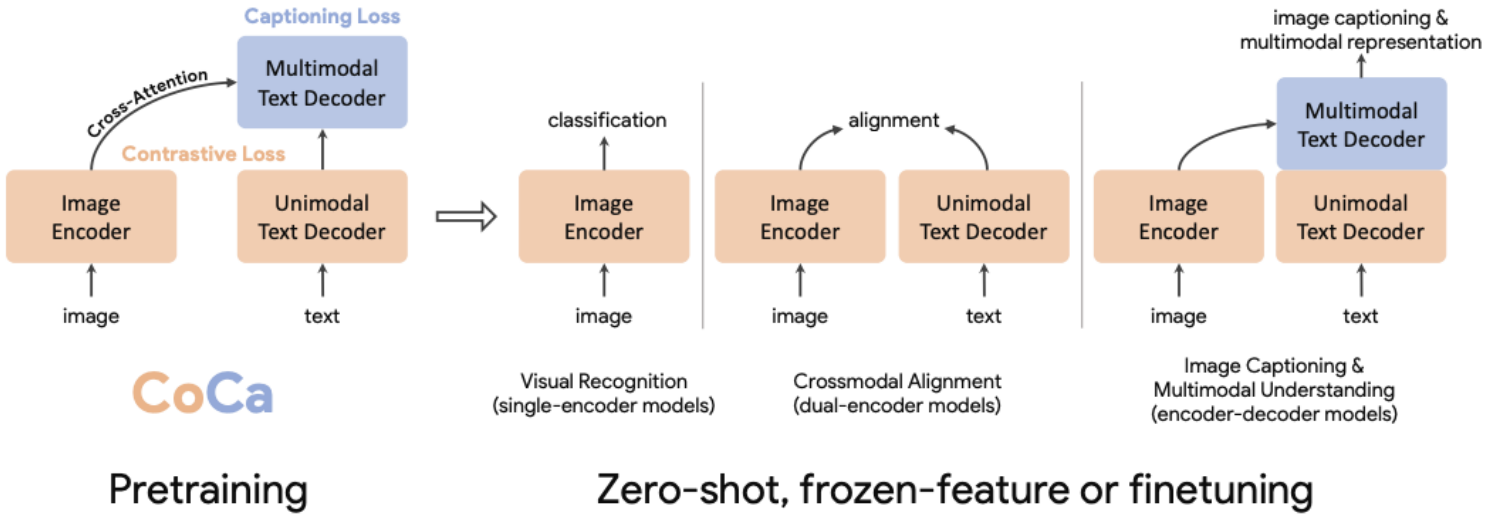




[1] FILIP: Fine-grained Interactive Language-Image Pre-Training, ICLR 2022

Objective function: adding a generative branch

- CoCa: Contrastive Captioner
  - Use mixed image-text and image-label (JFT-3B) data for pre-training
  - But adding an additional generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)
  - Similar to many vision-language models such as ALBEF, with the key difference that CoCa aims to learn a better image encoder from scratch



[1] Coca: Contrastive captioners are image-text foundation models, 2022

Objective function: adding a generative branch

- How about using the captioning loss alone?
  - VirTex was proposed to learn image encoders via an image captioning loss, but the scale is very small (COCO images)
  - In CLIP, it was also shown contrastive pre-training is a much better choice
  - In SimVLM, the authors also found that the learned image encoder was not competitive than CLIP, that's also why they later proposed CoCa
  - In Cap/CapPa, the authors argue that *image captioners are scalable vision learners, too.* Captioning exhibits the same or better scaling behavior.

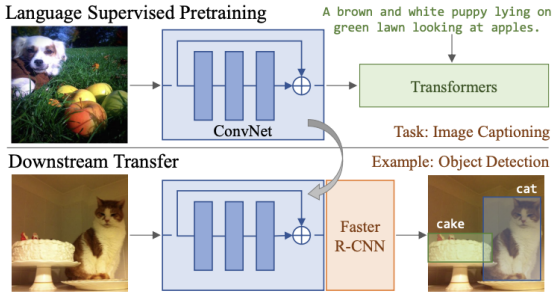
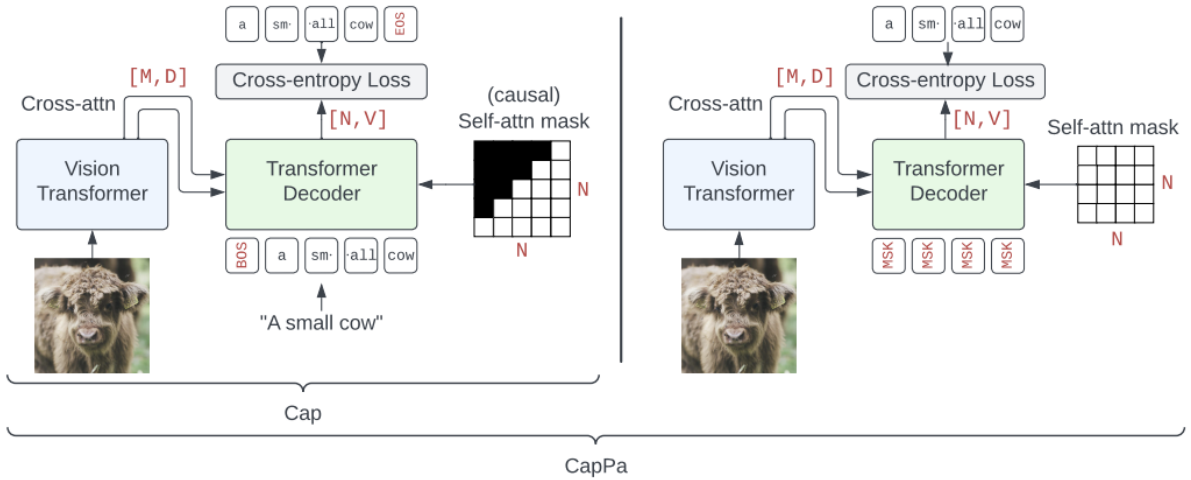


Figure 1: **Learning visual features from language:** First, we jointly train a ConvNet and Transformers using image-caption pairs, for the task of image captioning (top). Then, we transfer the learned ConvNet to several downstream vision tasks, for example object detection (bottom).

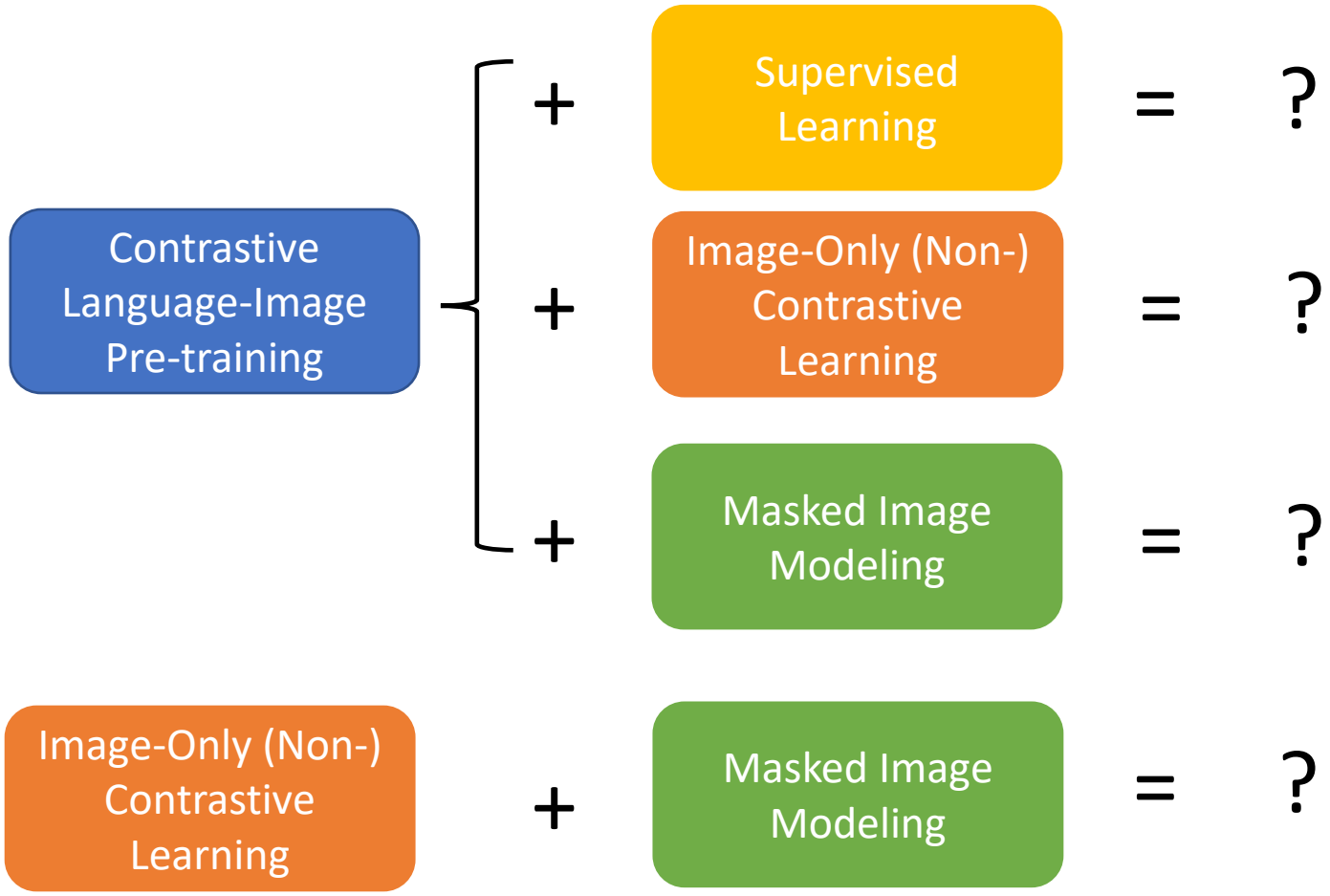
Method	Acc@1
SimCLRv2 (Chen et al., 2020a)	79.8
DINO (Caron et al., 2021)	80.1
CLIP (Radford et al., 2021)	85.4
ALIGN (Jia et al., 2021)	**85.5**
$\text{SimVLM}_{\text{base}}$	80.6
$\text{SimVLM}_{\text{large}}$	82.3
$\text{SimVLM}_{\text{huge}}$	83.6

Table 5: Linear evaluation on ImageNet classification, compared to state-of-the-art representation learning methods.

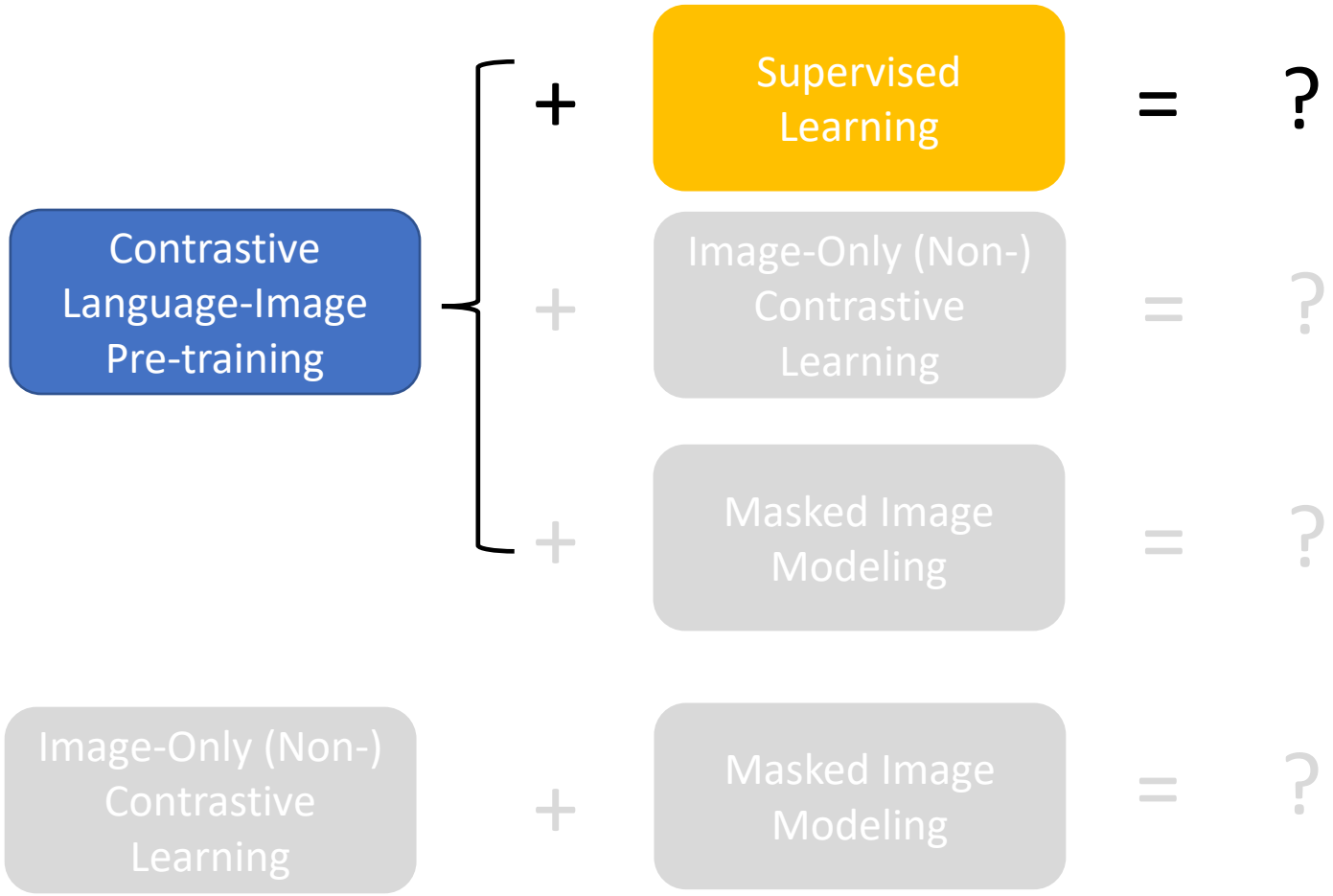


[1] Image Captioners Are Scalable Vision Learners Too, 2023
[2] VirTex: Learning Visual Representations from Textual Annotations, CVPR 2021
[3] SimVLM: Simple Visual Language Model Pretraining with Weak Supervision, ICLR 2022

Can CLIP be combined with other learning approaches?

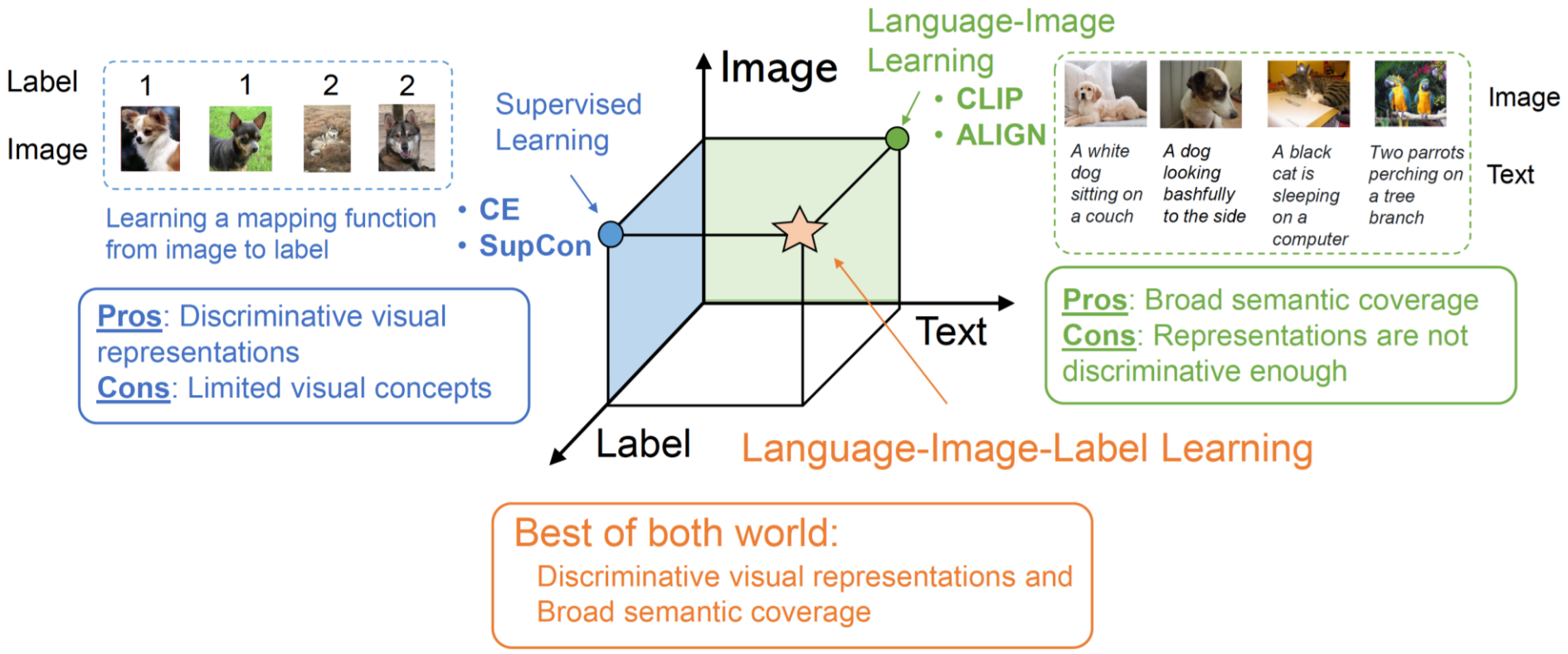


Can CLIP be combined with other learning approaches?



Noisy label+text supervision

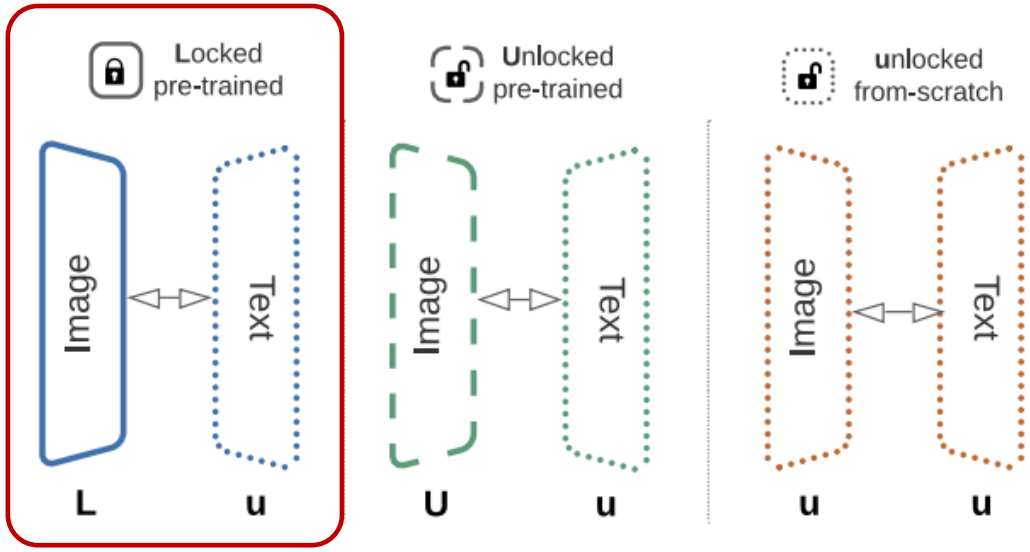
- UniCL: Image-text-label space
  - A principled way to use image-label and image-text data together
  - A scaled-up version is the Florence model

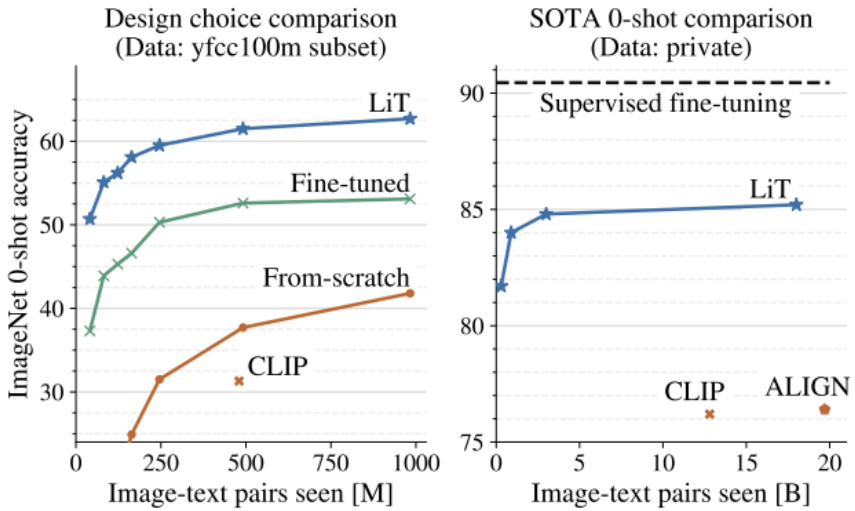


[1] Unified contrastive learning in image-text-label space, CVPR 2022
[2] Florence: A new foundation model for computer vision, 2021

Noisy label+text supervision

- LiT: Locking the image encoder
  - Use a pre-trained ViT-g/14 image encoder learned from JFT-3B (*image-label data*)
  - Then make it open-vocabulary by learning an additional text tower (*image-text data*)
  - Just teaches a text model to read out good representations from a pre-trained image model for new tasks





[1] Lit: Zero-shot transfer with locked-image text tuning, CVPR 2022

Noisy label+text supervision

- MOFI: Learn image representations from noisy entity annotated images
  - I2E data: The largest dataset of its kind in terms of the number of entities, 1 billion images with 2 million entities, 66 times more than JFT-3B and IG-3.6B

Model	GPR1200 mAP@1k (%)	ImageNet-ZS Acc@1 (%)
CLIP-L/14$_{\text{OpenAI}}$ [37]	72.19	75.27
MOFI-L/14	86.15	77.17
MOFI-H/14	86.66	78.46

(a) Comparison of MOFI and CLIP on GPR1200 image retrieval and ImageNet zero-shot classification tasks.

Dataset	# Images	# Classes
ImageNet-1K [41]	1.2M	1K
ImageNet-21K [40]	14M	21K
JFT-300M [48]	300M	18K
JFT-3B [62]	3B	30K
IG-3.6B [46]	3.6B	27K
I2E (Ours)	1.1B	2M

(b) I2E dataset and existing large-scale image classification datasets.

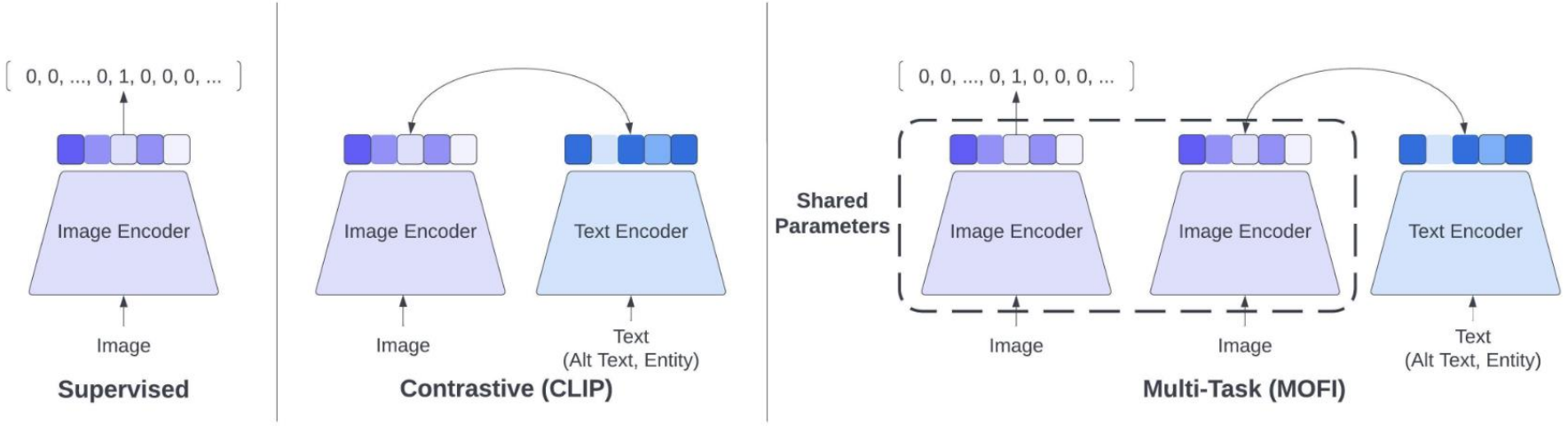
Table 1. MOFI is trained on the newly constructed Image-to-Entities (I2E) dataset, which has 66x more classes than the previous datasets. MOFI achieves significantly better performance on the image retrieval tasks when compared with CLIP.

Figure 1. **Examples of the I2E dataset.** Each caption is formatted as Entity_id (Entity_name).[4]

[1] MOFI: Learning Image Representations from Noisy Entity Annotated Images, 2023

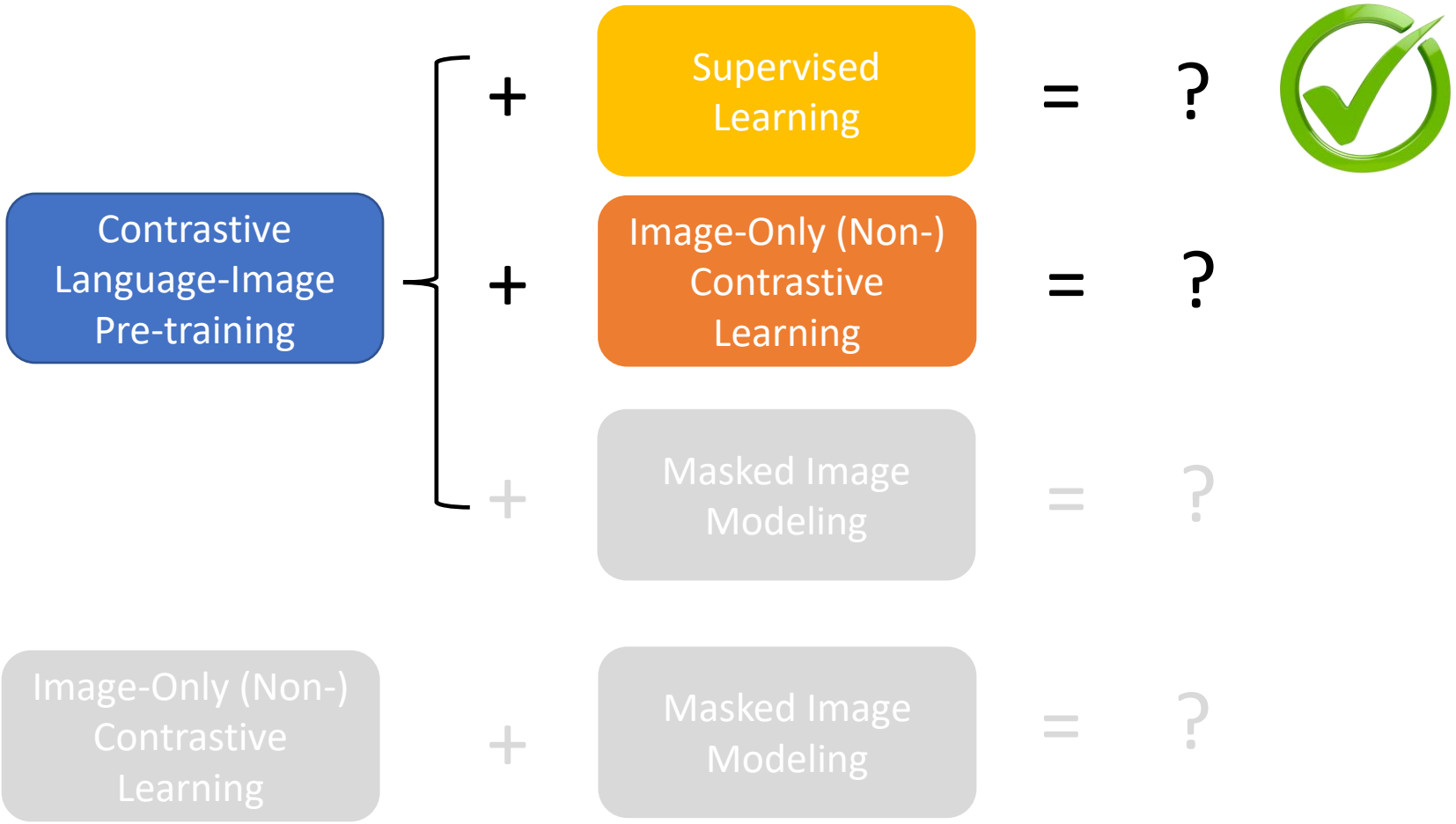
Noisy label+text supervision

- MOFI: Learn image representations from noisy entity annotated images
  - For supervised learning, treat entities as labels (2M labels)
    - Though classical, it's very effective for image-to-image retrieval tasks
  - For CLIP, treat entity names as free-form text, and enrich them with entity descriptions
    - Similar to K-Lite, but in a much larger scale (28M vs. 1B)
  - Combine supervised pre-training and CLIP for multi-tasking
  - Strong performance when compared with DINOv2



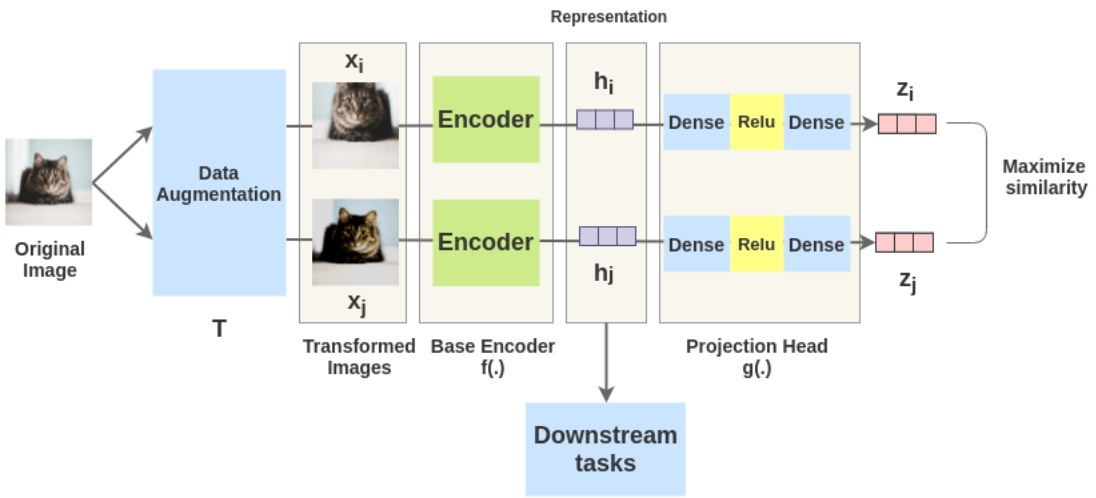
[1] MOFI: Learning Image Representations from Noisy Entity Annotated Images, 2023
[2] Dinov2: Learning robust visual features without supervision, 2023

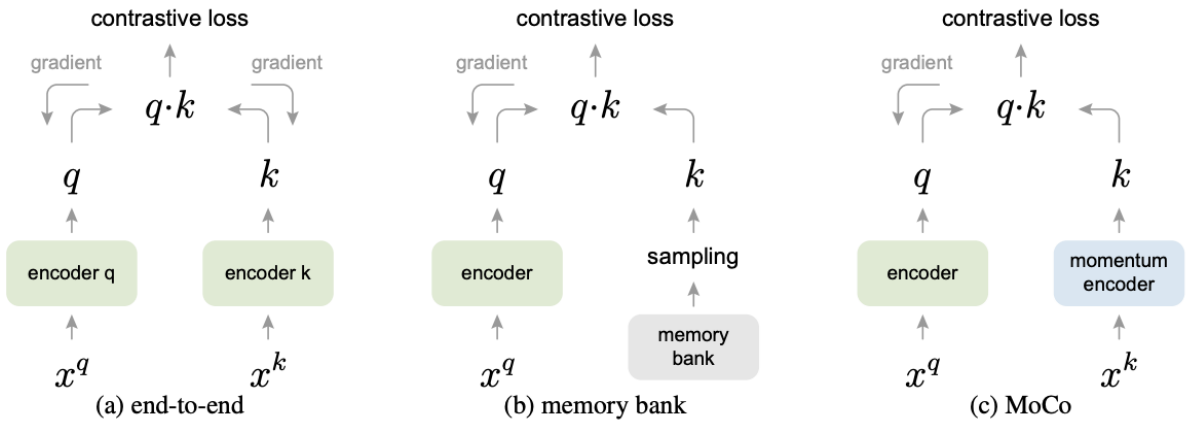
Can CLIP be combined with other learning approaches?



A high-level recap of image-only (non-)contrastive learning

- SimCLR: A Simple Framework of Contrastive Learning of Visual Representations
  - Given one image, two separate data augmentations are applied
  - A base encoder is followed by a project head, which is trained to maximize agreement using a contrastive loss (i.e., they are from the same image or not)
  - The project head is thrown away for downstream tasks
  - Nicely connected to mutual information maximization
  - A caveat of these line of methods is the requirement of large batch size or memory bank





[1] A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020
[2] Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

Image-only (Non-)Contrastive Learning

- Recent SSL methods relieve the dependency on negative samples
  - The use of negatives can be replaced by asymmetric architectures (BYOL, SimSiam), dimension de-correlation (Barlow twins), and clustering (SWaV, DINO), etc.

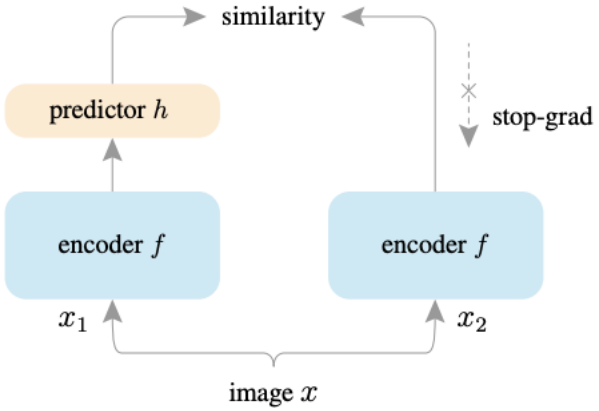


Figure 1. **SimSiam architecture**. Two augmented views of one image are processed by the same encoder network f (a backbone plus a projection MLP). Then a prediction MLP h is applied on one side, and a stop-gradient operation is applied on the other side. The model maximizes the similarity between both sides. It uses neither negative pairs nor a momentum encoder.

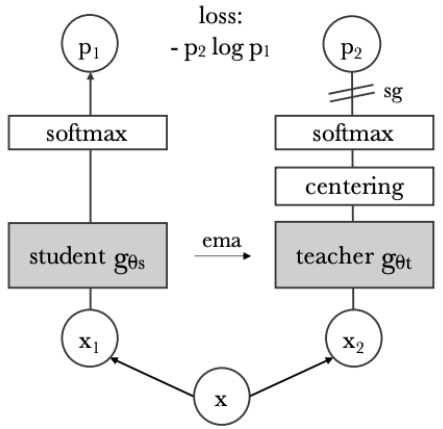
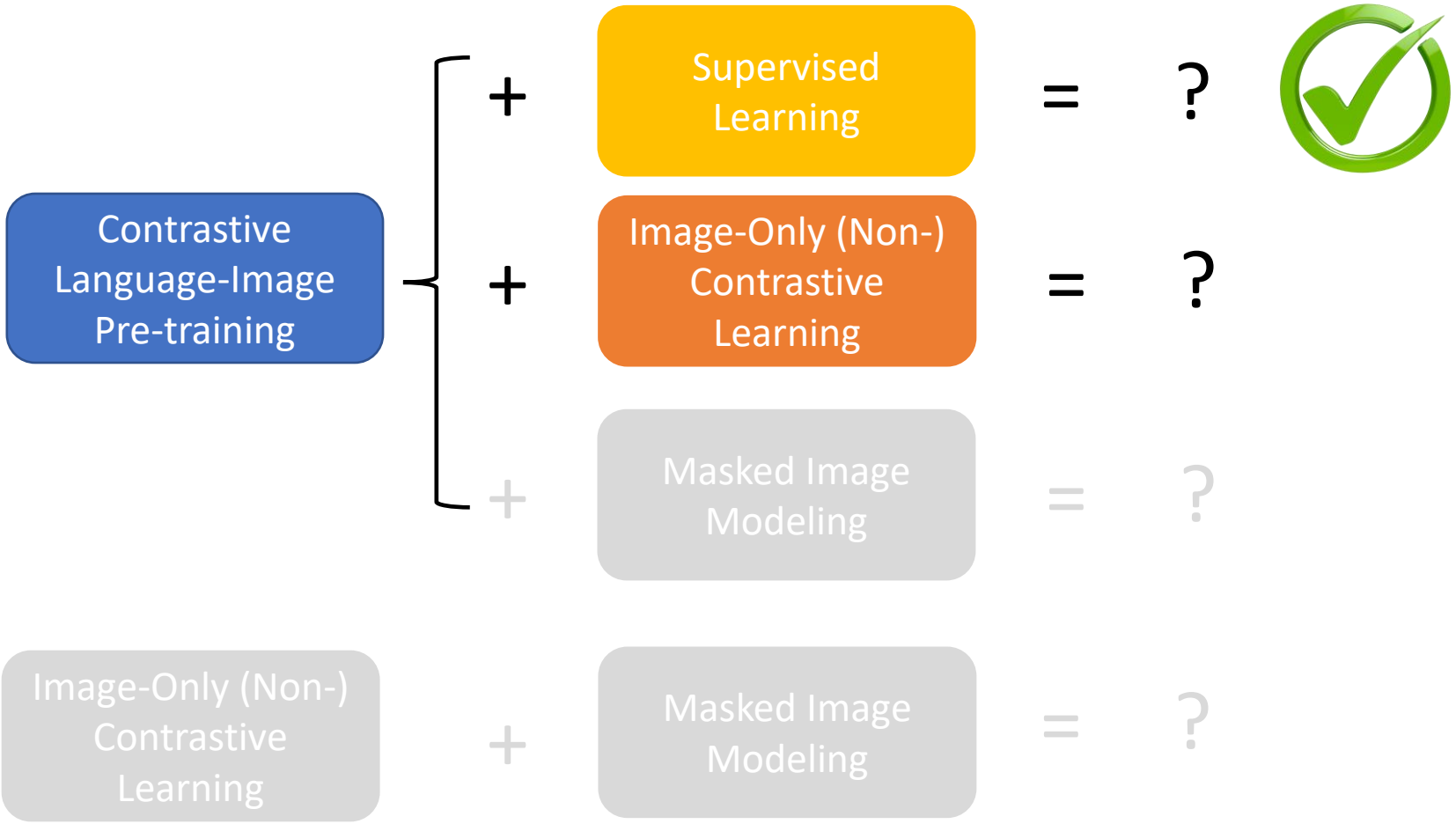


Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views (x_1, x_2) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a K dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

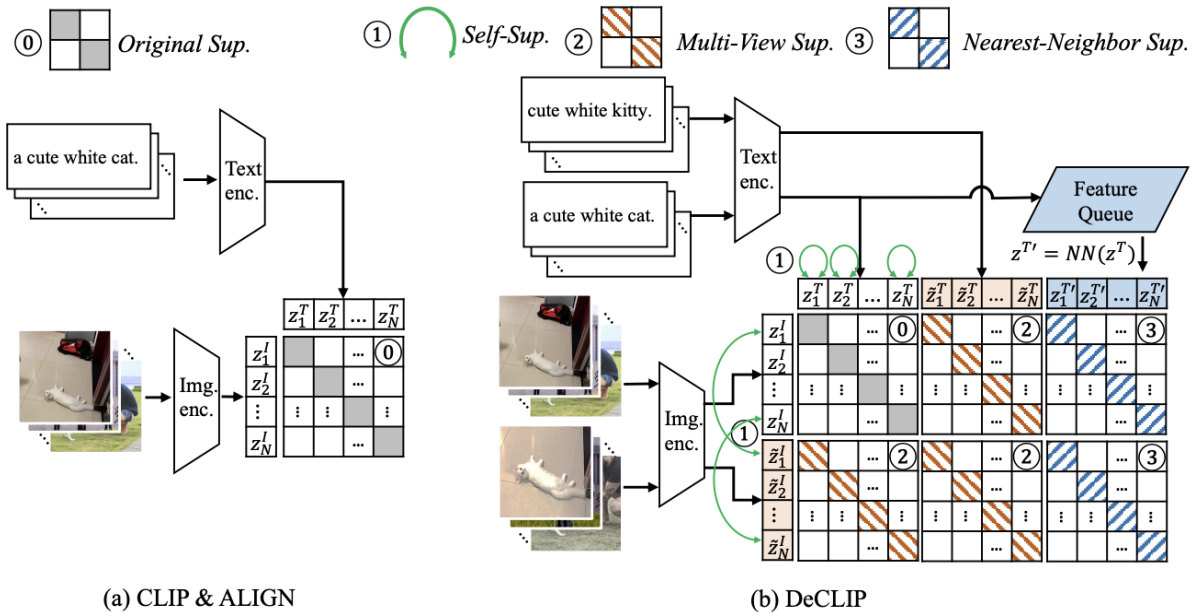
[1] Bootstrap your own latent-a new approach to self-supervised learning, NeurIPS 2020
[2] Exploring simple siamese representation learning, CVPR 2021
[3] Variance-invariance-covariance regularization for self-supervised learning, ICLR 2022
[4] Barlow twins: Self-supervised learning via ´ redundancy reduction, ICML 2021
[5] Unsupervised learning of visual features by contrasting cluster assignments, NeurIPS 2020
[6] Emerging properties in self-supervised vision transformers, ICCV 2021

Can CLIP be combined with other learning approaches?



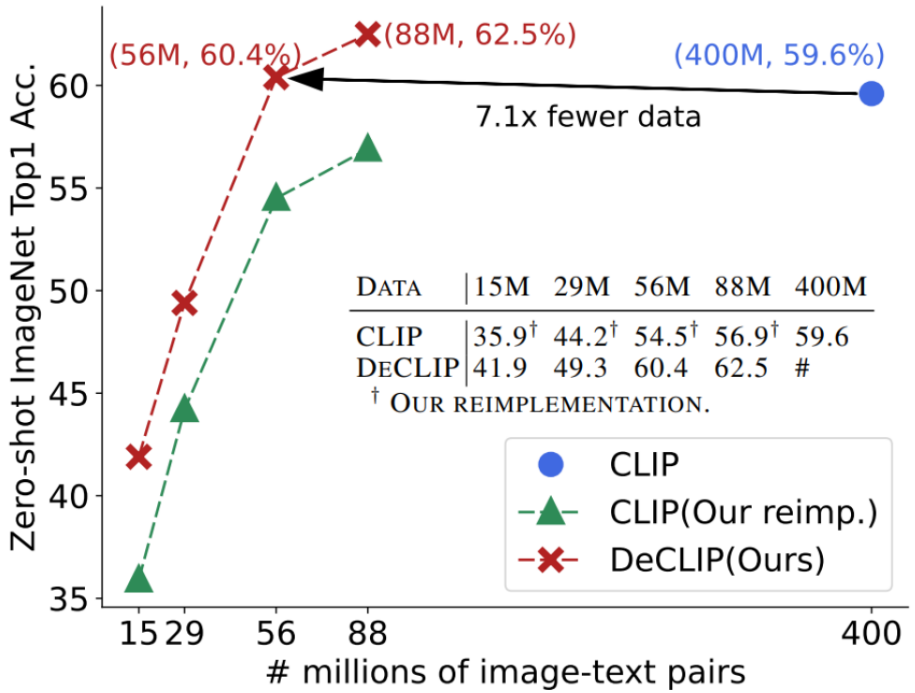
How to combine CLIP with image-only SSL?

- DeCLIP: supervision exists everywhere
  - Self-supervised learning on each modality: Image (SimSam), Text (MLM)
  - Multi-view supervision and Nearest-neighbor supervision



(a) CLIP & ALIGN

(b) DeCLIP



Data	15M	29M	56M	88M	400M
CLIP	35.9†	44.2†	54.5†	56.9†	59.6
DeCLIP	41.9	49.3	60.4	62.5	#

† Our reimplementation.

Combining vision-language and self-supervised learning improves data efficiency significantly

[1] Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR 2022

How to combine CLIP with image-only SSL?

- SLIP: Self-supervision meets language-image pretraining
  - Simply combine SimCLR and CLIP for model training
  - SLIP outperforms CLIP on both zero-shot transfer and linear probe settings

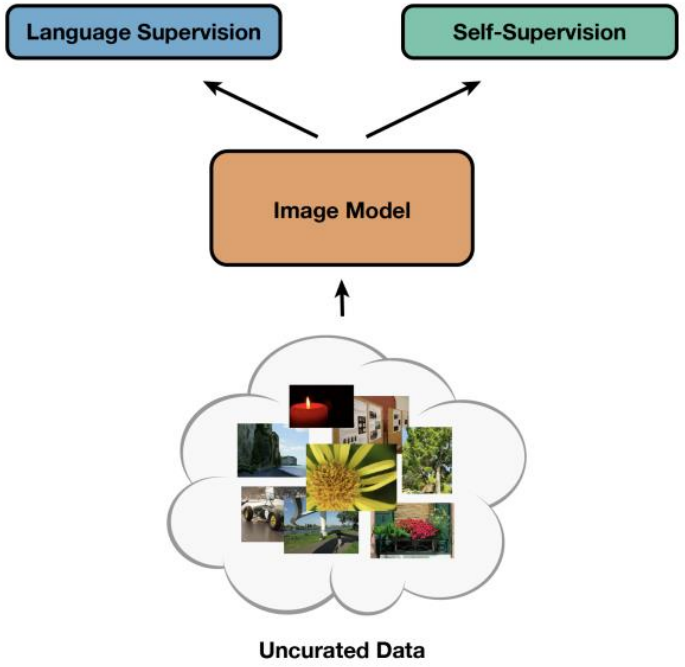


Figure 2. Illustration of SLIP, our multi-task framework. An image model has access to and can be trained with both language supervision from captions and self-supervision on images.

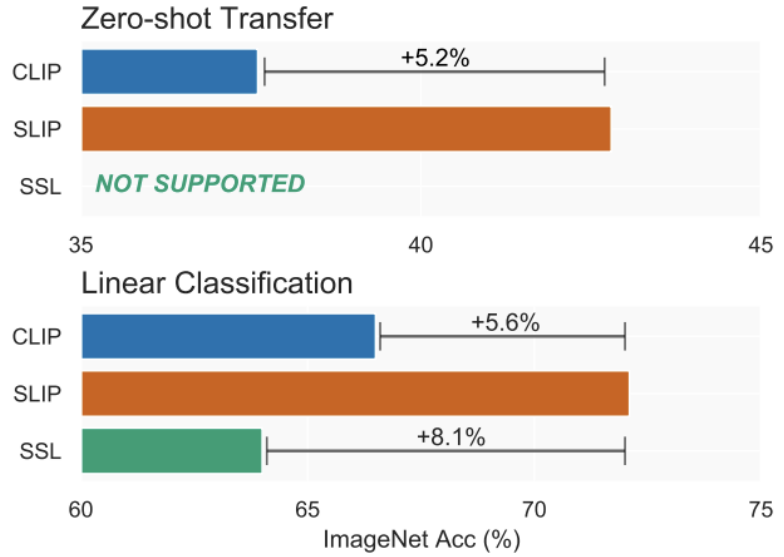
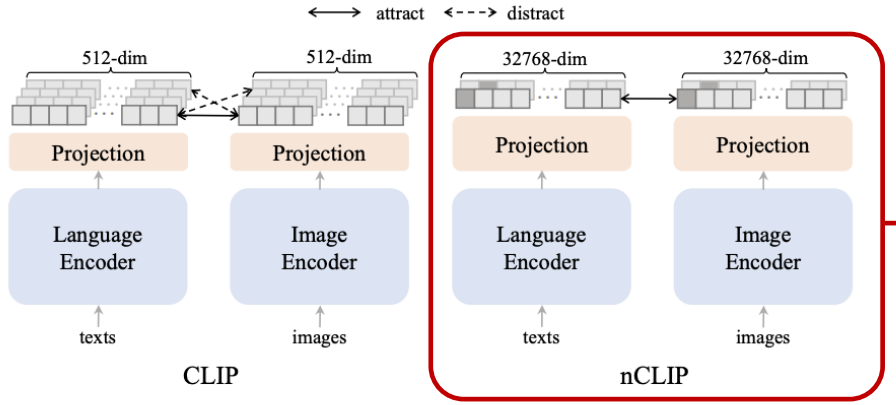


Figure 1. **SLIP pre-training on YFCC15M.** Combining image-only self-superivsion and image-text supervision simultaneously improves zero-shot transfer and linear classification on ImageNet.

[1] Slip: Self-supervision meets language-image pretraining, ECCV 2022

How to combine CLIP with image-only SSL

- xCLIP: Make CLIP non-contrastive using techniques from image-only SSL
  - Without using negatives, ensure non-trivial solutions via sharpness and smoothness regularization
  - CLIP works under 512-dim, but nCLIP needs project each modality to 32k-dim



$$\text{Let } \boldsymbol{p} = \text{softmax}(\boldsymbol{g}) \quad \boldsymbol{q} = \text{softmax}(\boldsymbol{h})$$

$$\mathcal{L}_{\text{CE}} = -\boldsymbol{p}^{\text{T}} \log(\boldsymbol{q}).$$

$$\mathcal{L}_{\text{EH}} = -\boldsymbol{p}^{\text{T}} \log(\boldsymbol{p}), \; -\mathcal{L}_{\text{HE}} = \overline{\boldsymbol{p}}^{\text{T}} \log(\overline{\boldsymbol{p}}),$$

$$\mathcal{L}_{\text{nCLIP}} = \mathcal{L}_{\text{CE}} + \lambda_1 \cdot \mathcal{L}_{\text{EH}} - \lambda_2 \cdot \mathcal{L}_{\text{HE}},$$

Figure 1. **Architecture comparison between CLIP and nCLIP.**

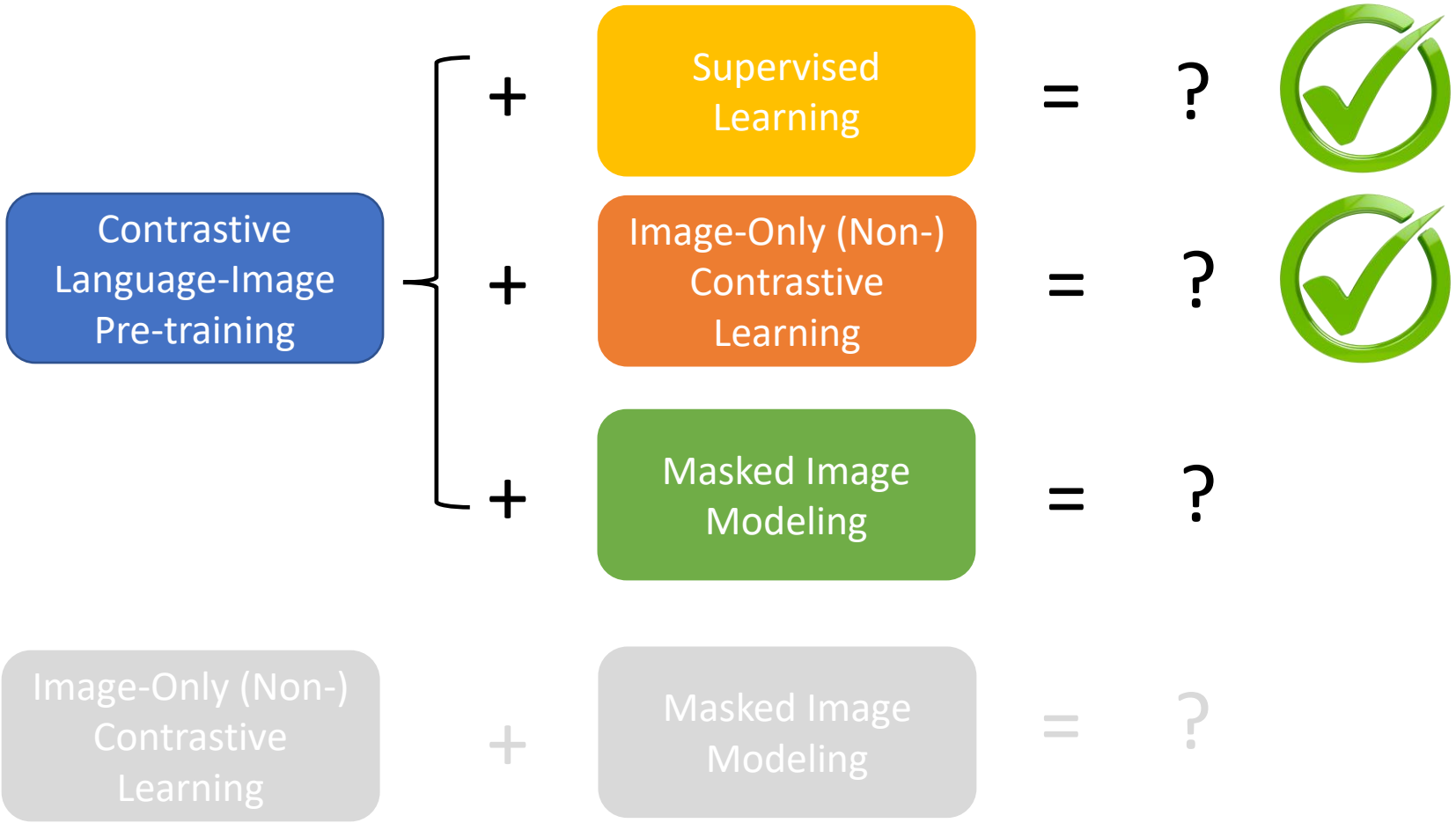
Model	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	FER2013	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST	ImageNet	Average
CLIP	61.2	79.7	50.6	23.7	56.5	15.9	5.8	46.4	27.6	54.7	71.3	48.9	10.5	37.3	91.3	24.5	39.7	13.1	31.6	9.1	50.0	45.0	32.4	12.8	**53.0**	49.1	45.7	40.3
nCLIP	28.4	79.5	49.1	11.3	57.0	5.9	4.5	51.4	22.9	14.6	65.0	23.1	9.9	13.5	**94.8**	15.1	21.2	2.7	35.4	5.8	**51.2**	42.0	28.4	12.4	52.7	50.0	37.0	32.7
xCLIP	**65.8**	**83.4**	**54.5**	**25.1**	**59.9**	**18.0**	**5.8**	**52.2**	**33.2**	**57.1**	**73.9**	**50.0**	**12.3**	**39.0**	92.8	**40.0**	**43.6**	**16.3**	**39.8**	**9.3**	51.1	**49.8**	**35.4**	**18.4**	52.5	**50.2**	**48.8**	**43.6**

Table 1. **Zero-shot classification.** We report on a variety of classification benchmarks with ViT-B/16 pre-trained on IT35M. Detailed protocols for each dataset strictly follow CLIP [78]. xCLIP achieves a consistent performance gain compared to CLIP in a wide range of classification datasets. Best results of each column are **bolded**.

Only nCLIP is not enough for strong performance, need to be used together with CLIP, i.e., xCLIP = CLIP + nCLIP.

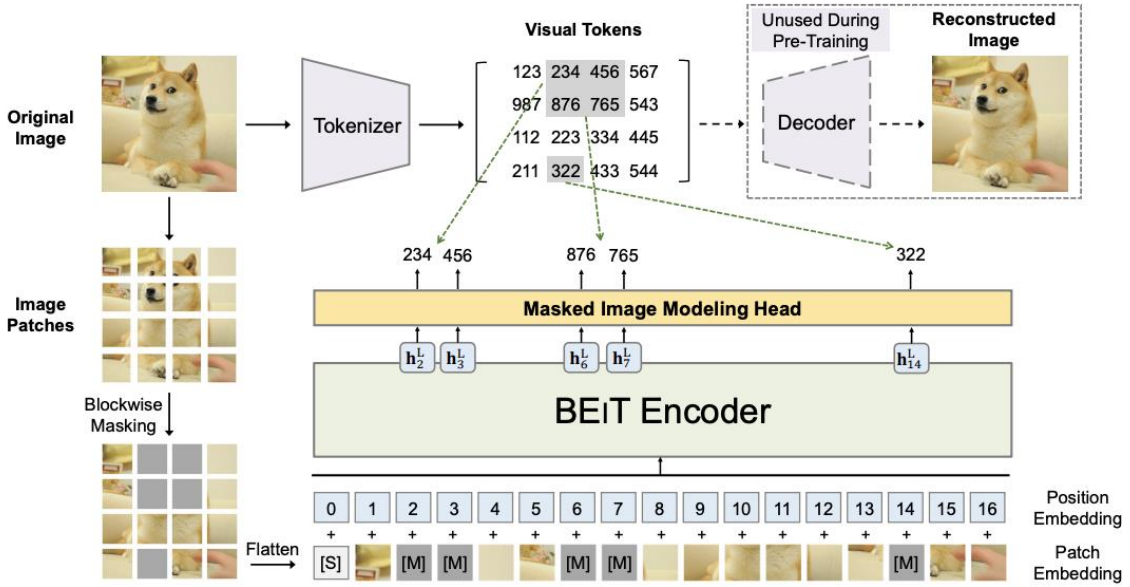
[1] Non-Contrastive Learning Meets Language-Image Pre-Training, CVPR 2023

Can CLIP be combined with other learning approaches?

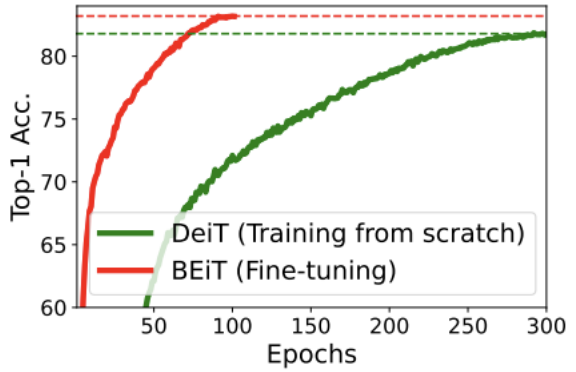


A high-level recap of masked image modeling

- BEiT: BERT Pre-Training of Image Transformers
  - Before pre-training, learn an "image tokenizer" via VQ-VAE/GAN, where an image is tokenized into discrete visual tokens
    - Similar approaches have been used for image generation, such as DALLE, Parti.
  - Randomly masking image patches, pre-train the model to predict masked visual tokens
  - Can be understood as knowledge distillation between the image tokenizer and the BEiT encoder, but the latter only sees partial of the image



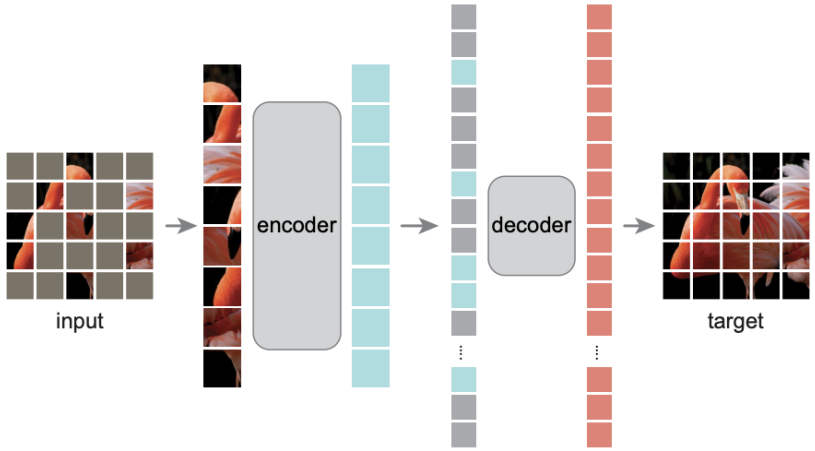
Strong model finetuning performance



[1] BEiT: BERT Pre-Training of Image Transformers, ICLR 2022
[2] iBOT: Image BERT Pre-Training with Online Tokenizer, ICLR 2022

Masked autoencoders and masked feature prediction

- MAE: Using pixel values as targets also works great
  - A large random subset of images (75%) is masked out
  - The encoder is applied to visible patches, mask tokens are introduced after the encoder
  - MAE pre-training is especially helpful for object detection and segmentation tasks
- MaskFeat: Other image features can be used as targets as well



method	pre-train data	AP^{box} ViT-B	AP^{box} ViT-L	AP^{mask} ViT-B	AP^{mask} ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT

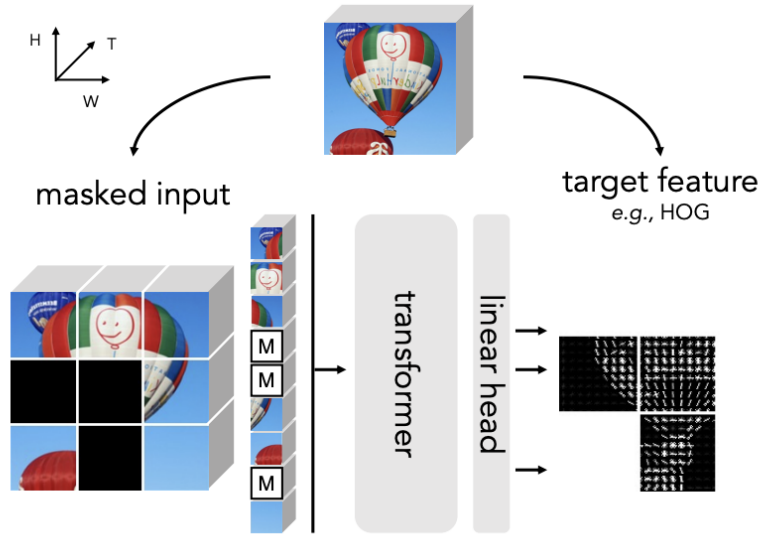


Figure 2. **MaskFeat pre-training.** We randomly replace the input space-time cubes of a video with a `[MASK]` token and directly regress features (*e.g.* HOG) of the masked regions. After pre-training, the Transformer is fine-tuned on end tasks.

[1] Masked Autoencoders Are Scalable Vision Learners, CVPR 2022
[2] Masked feature prediction for self-supervised visual pre-training., CVPR 2022

Potential problems with MIM

- Scaling properties of MIM is still not that clear yet
  - MIM is scalable in terms of model size
  - However, how about scaling data size like CLIP using billions of image-text pairs?
  - There are some studies, but not much ([1] [2] [3] shown in the footnote)
- MIM is good for model finetuning, but does not learn a global image representation
  - Roughly iBOT = DINO + BEiT, and then we also have DINOv2

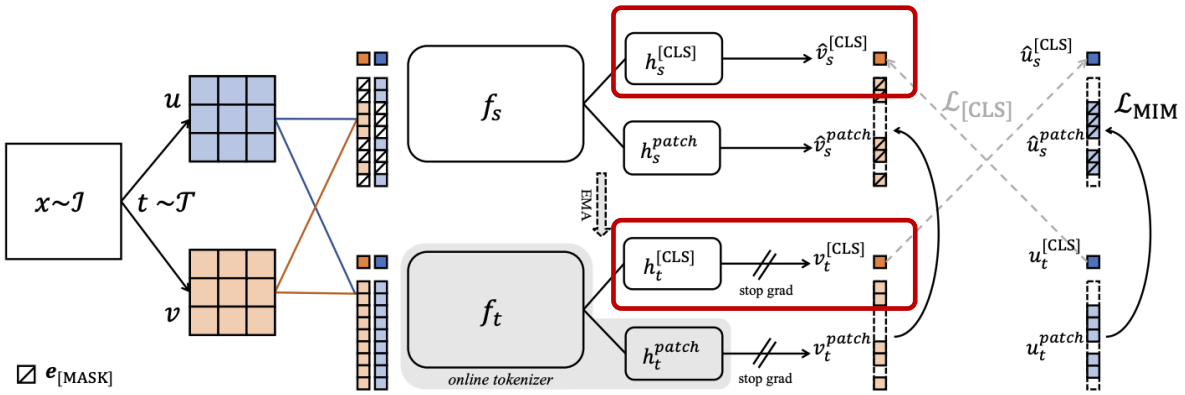


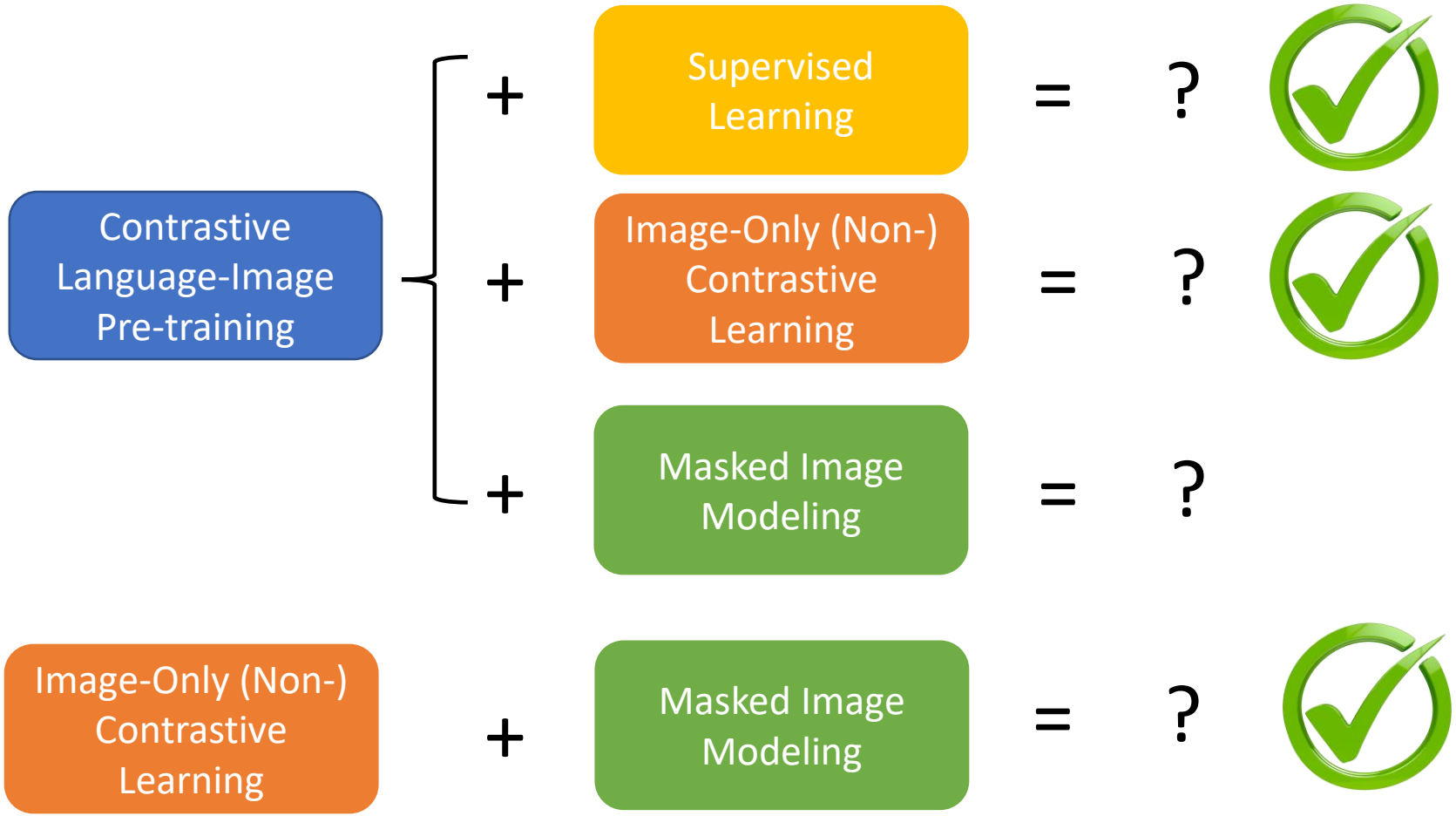
Table 9: **Effect of design choices of semantically meaningful tokenization.**

Method	$\mathcal{L}_{MIM}$	$\mathcal{L}_{[CLS]}$	SH	k-NN	Lin.	Fin.
iBOT	✓	✓	✓	69.1	74.2	81.5
	✓	✓	✗	69.0	73.8	81.5
	✓	✗	-	9.5	29.8	79.4
	○	✗	-	44.3	60.0	81.7
BEiT	△	✗	-	6.9	23.5	81.4
DINO	✗	✓	-	67.9	72.5	80.6
BEiT + DINO	△	✓	-	48.0	62.7	81.2

○: standalone DINO (w/o mcrop, 300-epoch)
△: pre-trained DALL-E encoder

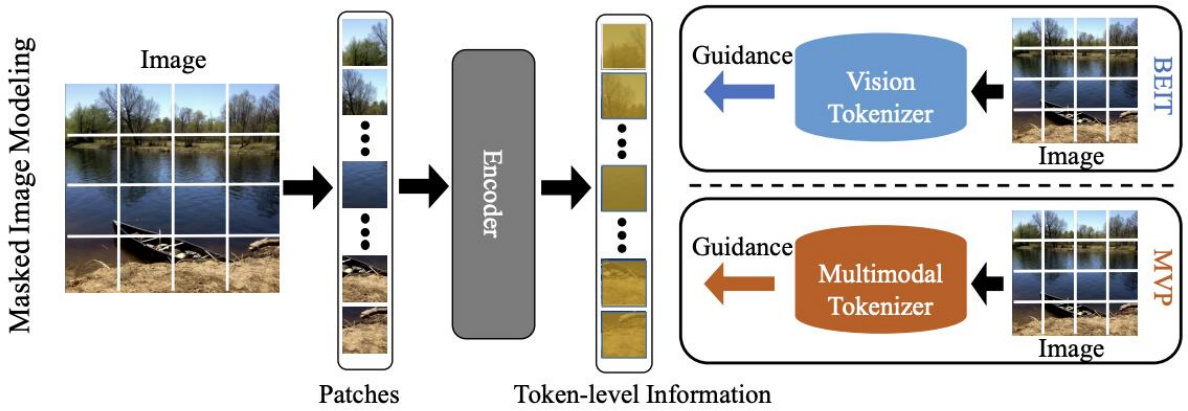
[1] On Data Scaling in Masked Image Modeling, 2022
[2] Delving Deeper into Data Scaling in Masked Image Modeling, 2023
[3] The effectiveness of MAE pre-pretraining for billion-scale pretraining, 2023
[4] iBOT: Image BERT Pre-Training with Online Tokenizer, ICLR 2022
[5] DINOv2: Learning Robust Visual Features without Supervision, 2023

Can CLIP be combined with other learning approaches?

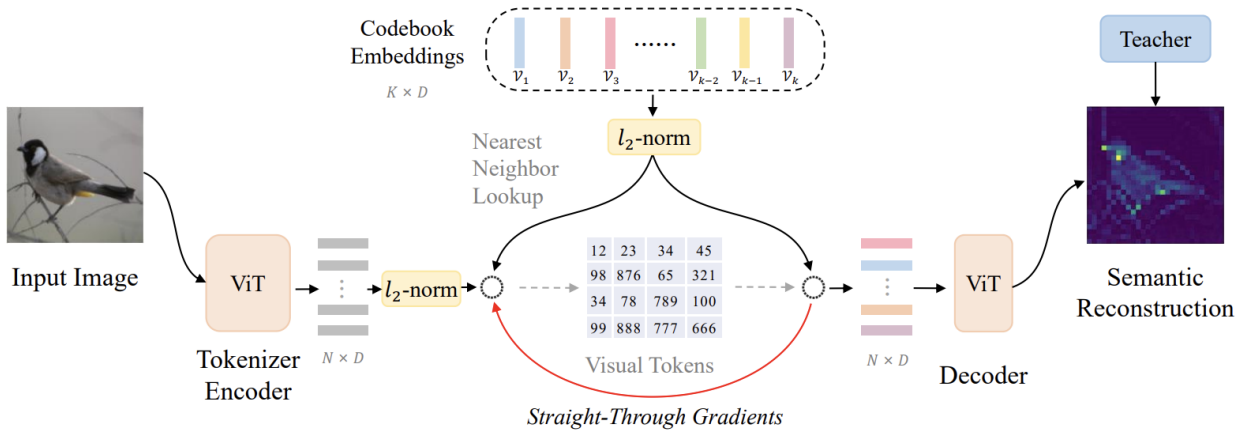


Shallow interaction of CLIP and MIM

- Turns out image features extracted from CLIP is a good target for MIM training
  - Captures the semantics that is missing in MIM training



Approach 1 (MVP): regress CLIP features

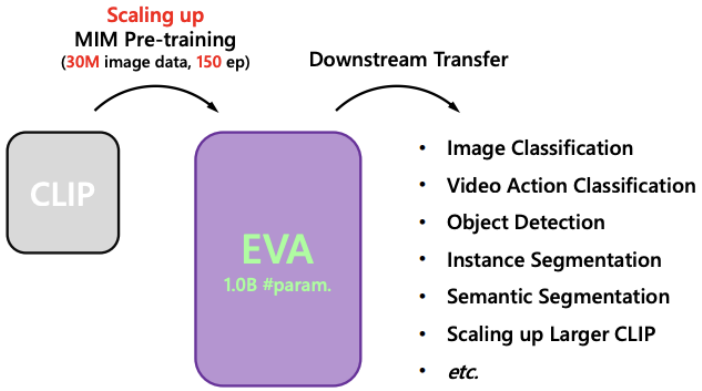


Approach 2 (BEiT v2): compress the information inside CLIP features into the visual tokens, then perform regular BEiT training

[1] MVP: Multimodality-guided Visual Pre-training, ECCV 2022
[2] BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers, 2022

Shallow interaction of CLIP and MIM

- This approach is further popularized by the EVA series of work



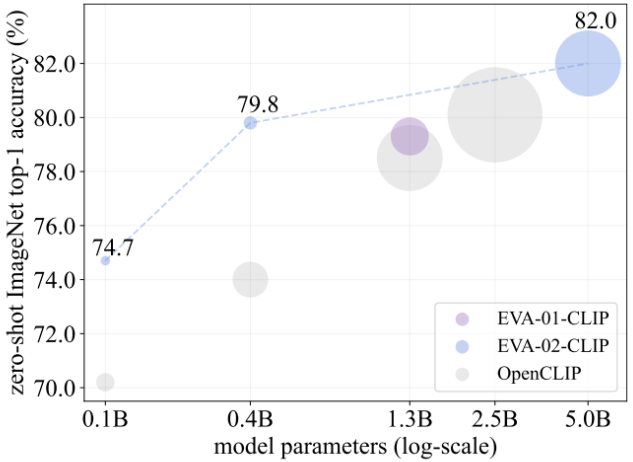


Figure 1: **Summary of CLIP models' ImageNet-1K zero-shot classification performance.** The diameter of each circle corresponds to forward GFLOPs x the number of training samples.

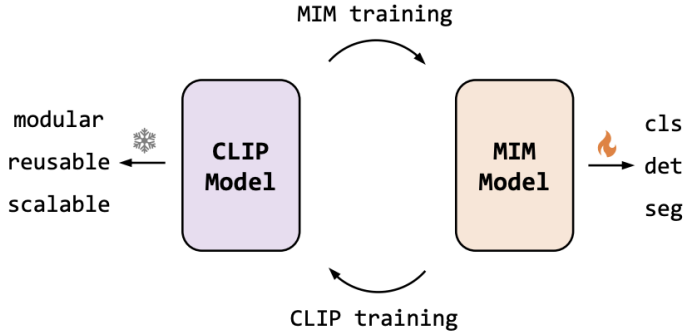
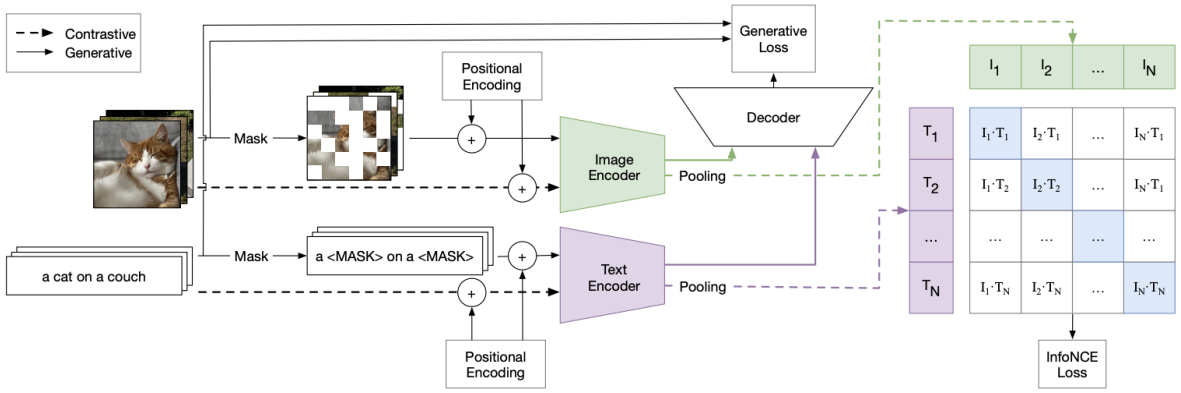


Figure 3: **Alternate learning of MIM and CLIP representations.** Starting with a off-the-shelf CLIP(*e.g.*, OpenAI CLIP [95]), alternate training of the pure MIM visual representations as well as vision-language CLIP representations can improve both MIM and CLIP performances in a bootstrapped manner. The MIM representations can be used to fine-tune various downstream tasks while the (frozen) CLIP representations enable modular, reusable and scalable next-gen model design.

[1] EVA: Exploring the Limits of Masked Visual Representation Learning at Scale, CVPR 2023
[2] EVA-CLIP: Improved Training Techniques for CLIP at Scale, 2023
[3] EVA-02: A Visual Representation for Neon Genesis, 2023.

Is a deeper integration of CLIP and MIM possible?

- Masked autoencoding does not help natural language supervision at scale



Models	Zero-shot	Linear Probing
MAE	–	33.9
M3AE	–	52.5
CLIP	29.7	52.6
MAE-CLIP	33.8	58.9

Table 1: ImageNet classification with zero-shot transfer or linear probing after pretraining on the *CC* dataset (11.3M images). MAE-CLIP significantly improves the classification performance of CLIP in the small scale regime.

Models	Zero-shot	Linear Probing
M3AE*	–	69.3
CLIP_{GAP}	61.8	75.9
CLIP_{MAX}	63.7	77.5
$\text{MAE-CLIP}_{\text{GAP}}$	57.4	75.7
$\text{MAE-CLIP}_{\text{MAX}}$	60.9	76.6

Table 6: ImageNet classification after pretraining on *web-crawled* dataset (1.4B images). In the large scale regime, self-supervision does not complement natural language supervision.

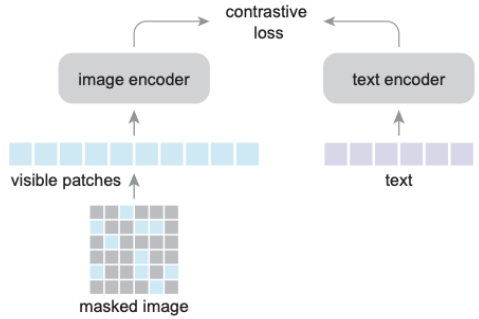


Figure 2. **Our FLIP architecture**. Following CLIP [52], we perform contrastive learning on pairs of image and text samples. We randomly mask out image patches with a high masking ratio and encode only the visible patches. We do not perform reconstruction of masked image content.

	mask 50%	mask 75%
baseline	**69.6**	**68.2**
+ MAE	69.4	67.9

(f) **Reconstruction**. Adding the MAE reconstruction loss has no gain.

[1] Masked Autoencoding Does Not Help Natural Language Supervision at Scale, CVPR 2023
[2] Scaling language-image pre-training via masking, CVPR 2023

From masked image modeling to masked multimodal modeling

- BEiT-3: BERT and BEiT can be incorporated
  - Performing masked data modeling on both image/text and image-text data with a Multiway transformer
  - Shared self-attention layer with 3 FFN modality experts

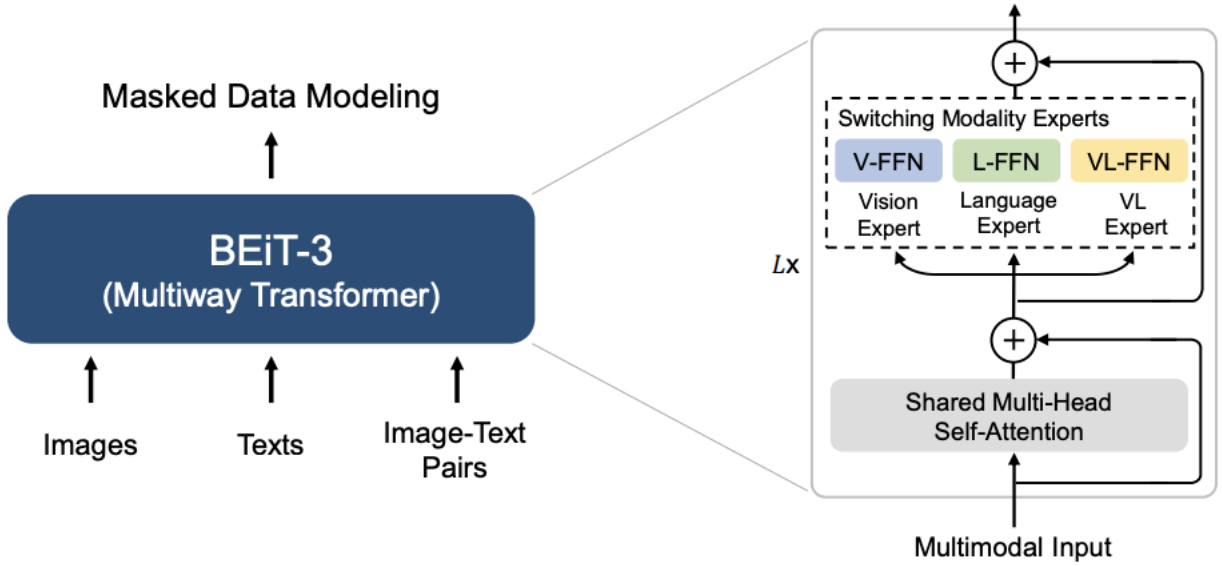
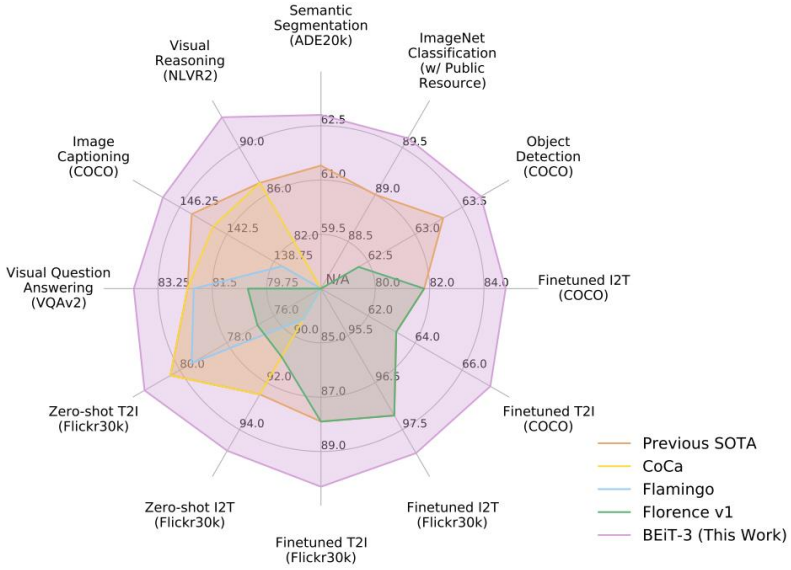
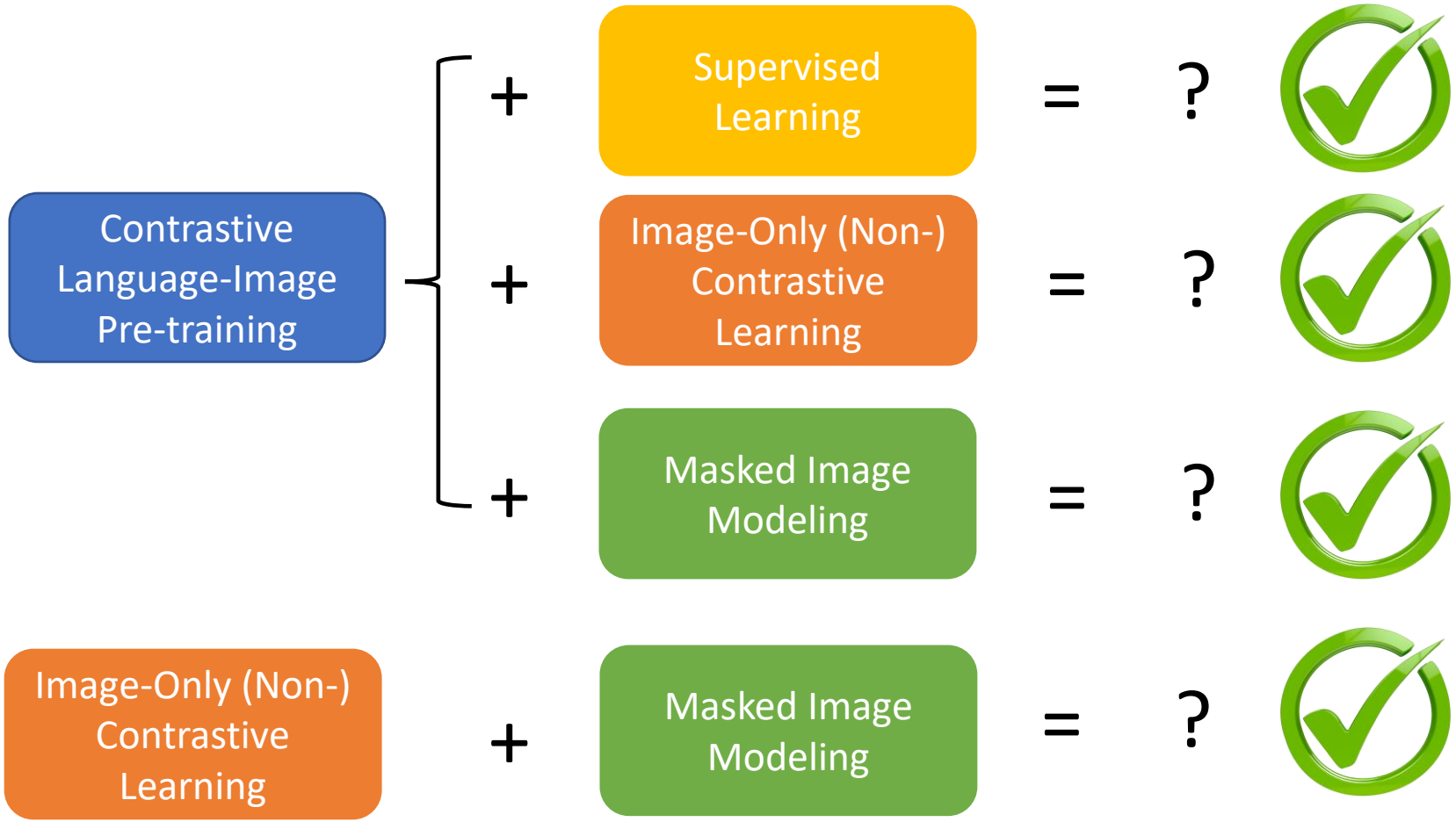


Figure 2: Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.



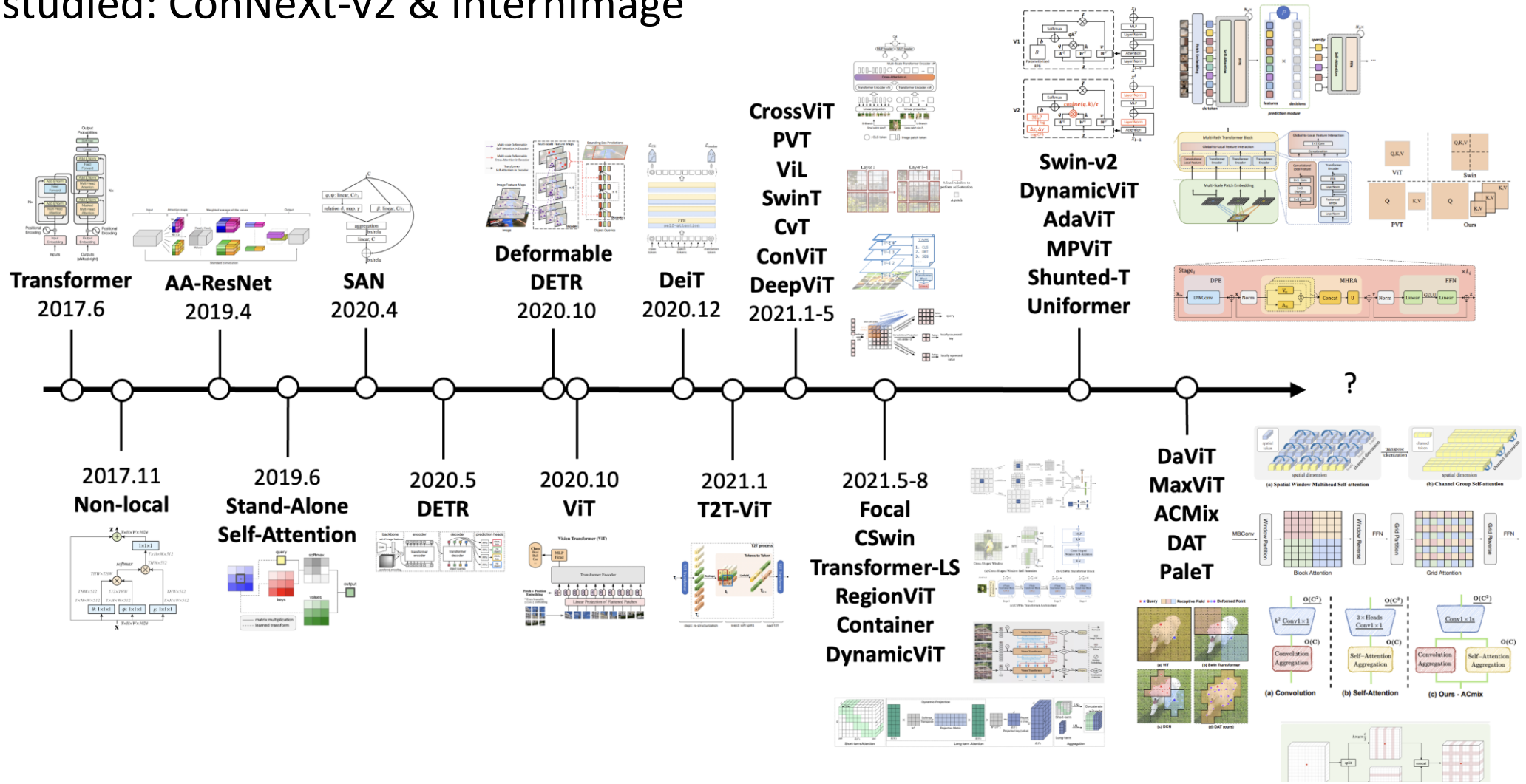
[1] Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks

Can CLIP be combined with other learning approaches?



Backbones other than ViT

- Besides ViT that has been scaled up to 22B, other backbones can also be studied: ConNeXt-v2 & InternImage



[1] InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions, CVPR 2023
[2] ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders, 2023

A high-level summary

- We center around CLIP, and discussed how to train a strong image backbone as the foundation

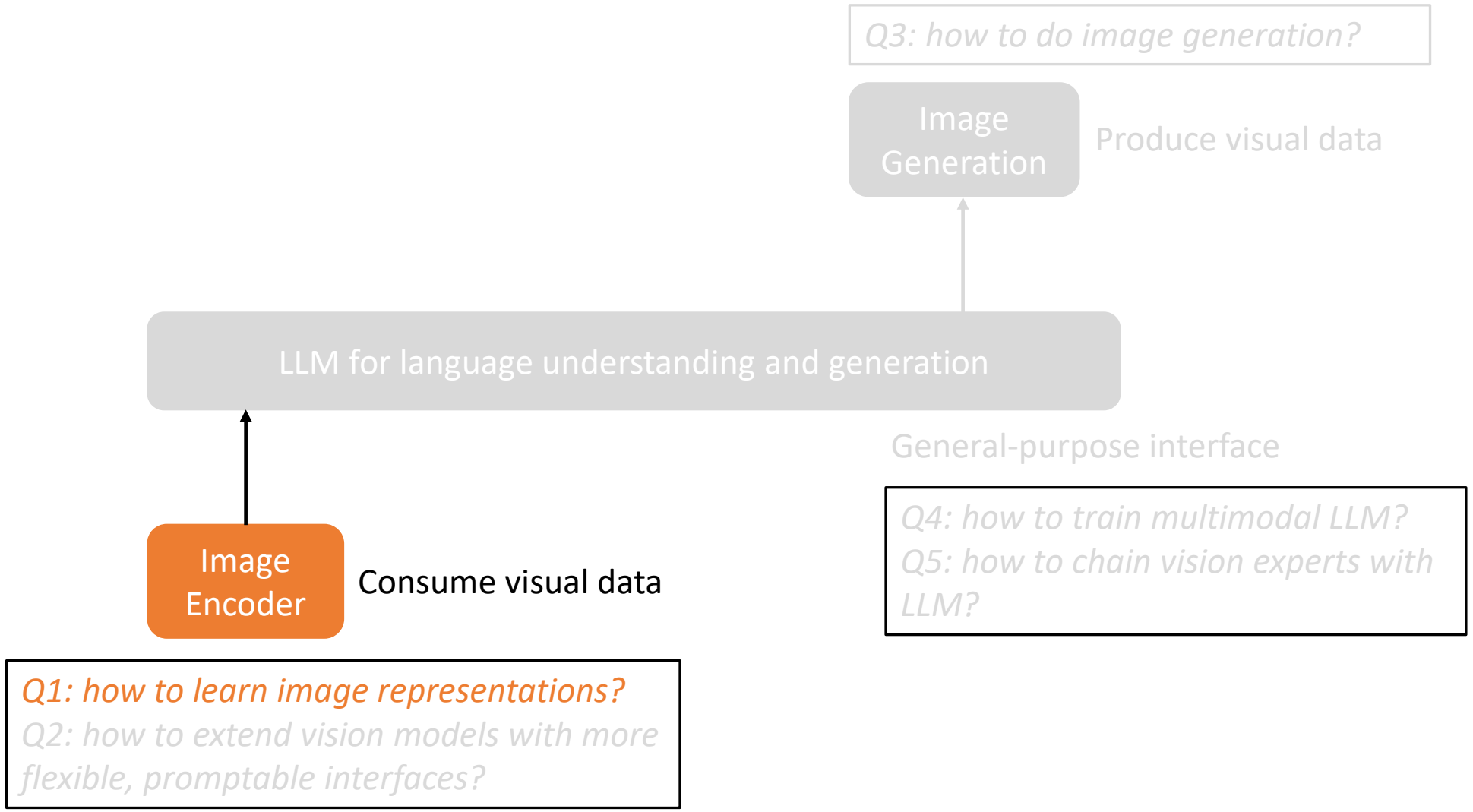
- We focus on image-level pre-training, not further into region/pixel-level pre-training (e.g., GLIP, SAM, X-Decoder), as they typically will use a pre-trained image encoder at first hand

- We observe three high-level principles from the current literature
  - Scaling: a good algorithm should be simple but also scale well
  - Contrasting: From SimCLR to CLIP
  - Masking: From BERT to BEiT

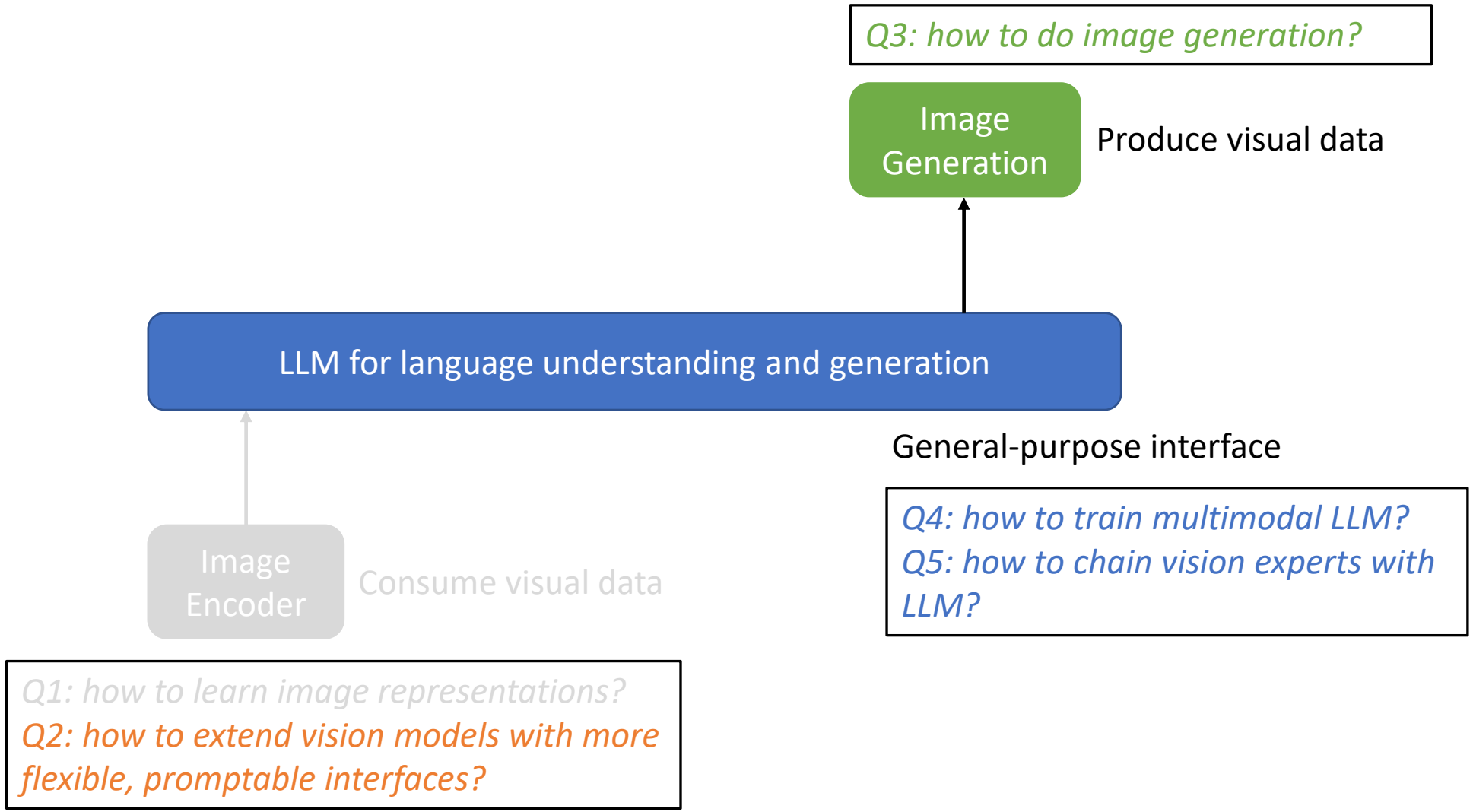
Future challenges

- How to further scale up?
  - In terms of both data scale and model scale
- New model training paradigm?
  - Simple algorithm that scales well and goes beyond CLIP and MIM
- How to perform unified image-/region-/pixel-level pre-training?
  - So that the model can have a holistic view of the image at different granularities
- How to extend vision models with more flexible, promptable interfaces?
  - How NLP concepts like in-context learning, chain-of-thoughts, prompting, emerging properties can be exhibited in the CV context
- How to train vision backbones with more innovative data?
  - So that to unblock new model capabilities such as the ones demonstrated by GPT-4
  - E.g., read a whole scanned paper and then summarize the paper in a few bullet points

Till now, we have discussed how to learn a strong image backbone



More to come …


Q3: how to do image generation?
Image Generation
Produce visual data
LLM for language understanding and generation
General-purpose interface
Q4: how to train multimodal LLM?
Q5: how to chain vision experts with LLM?
Image Encoder
Consume visual data
Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?

Thank you!
Any Questions?