VLMs
BEV

D.A. Forsyth,

University of Illinois at Urbana Champaign

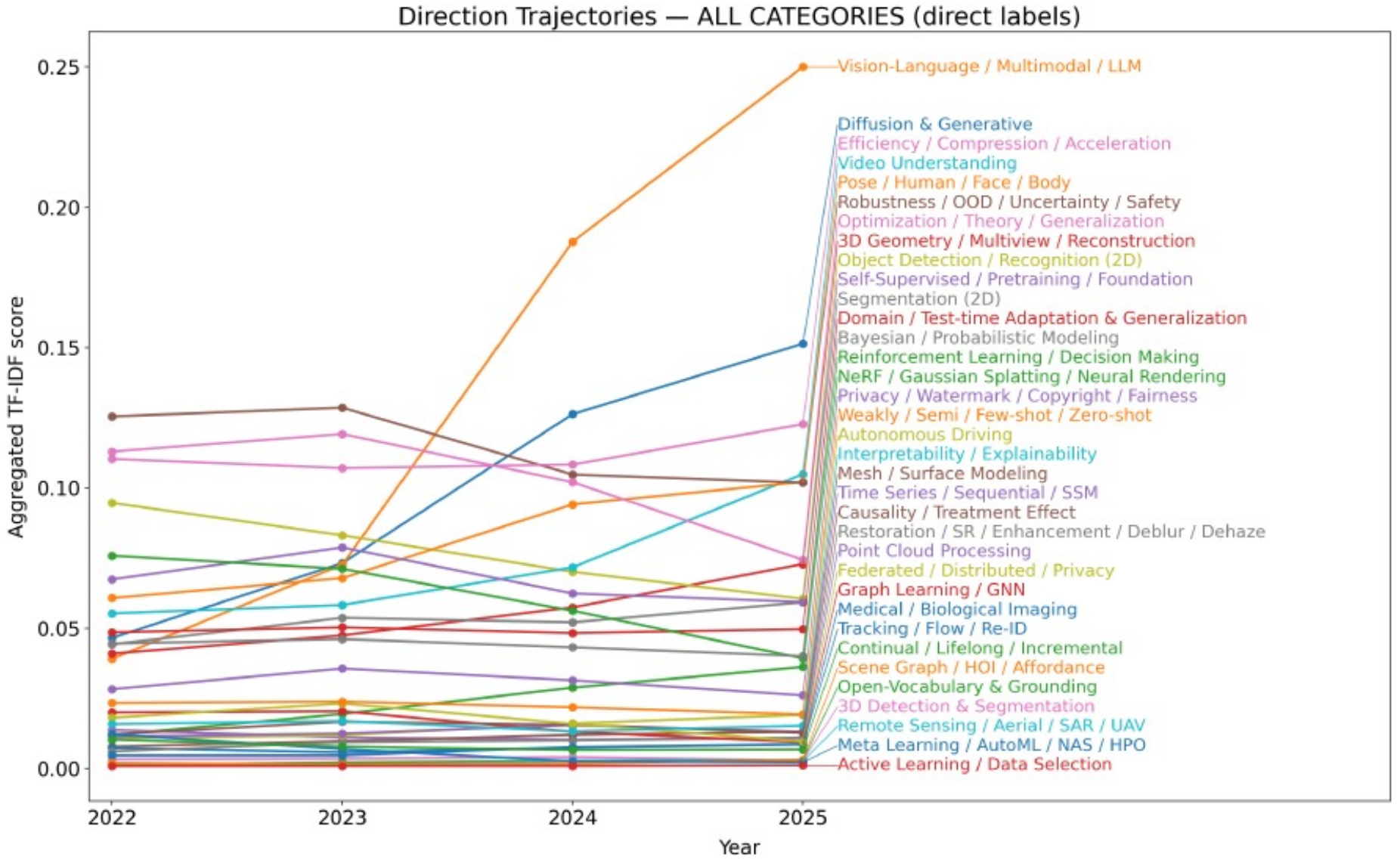


Figure 1: **Direction Trajectories across CVPR+ICLR+NeurIPS — ALL CATEGORIES (direct labels).** Each curve is the yearly aggregated TF–IDF mass for a direction (integer year ticks).

Vision Language Models: A Survey of 26K Papers (CVPR, ICLR, NeurIPS 2023–2025), Lin, 2025

Big Picture

- Train models using both images and text
- Models should
  - accept text/images, produce text/images
  - answer questions about images

- Advantages
  - extremely rich training opportunities
    - for a general vision encoder
  - lever text descriptions of objects to build detectors
    - particularly important for rare or unseen objects
      - a version of distributional semantics

An Introduction to Vision-Language Modeling
Bordes et al, 2024

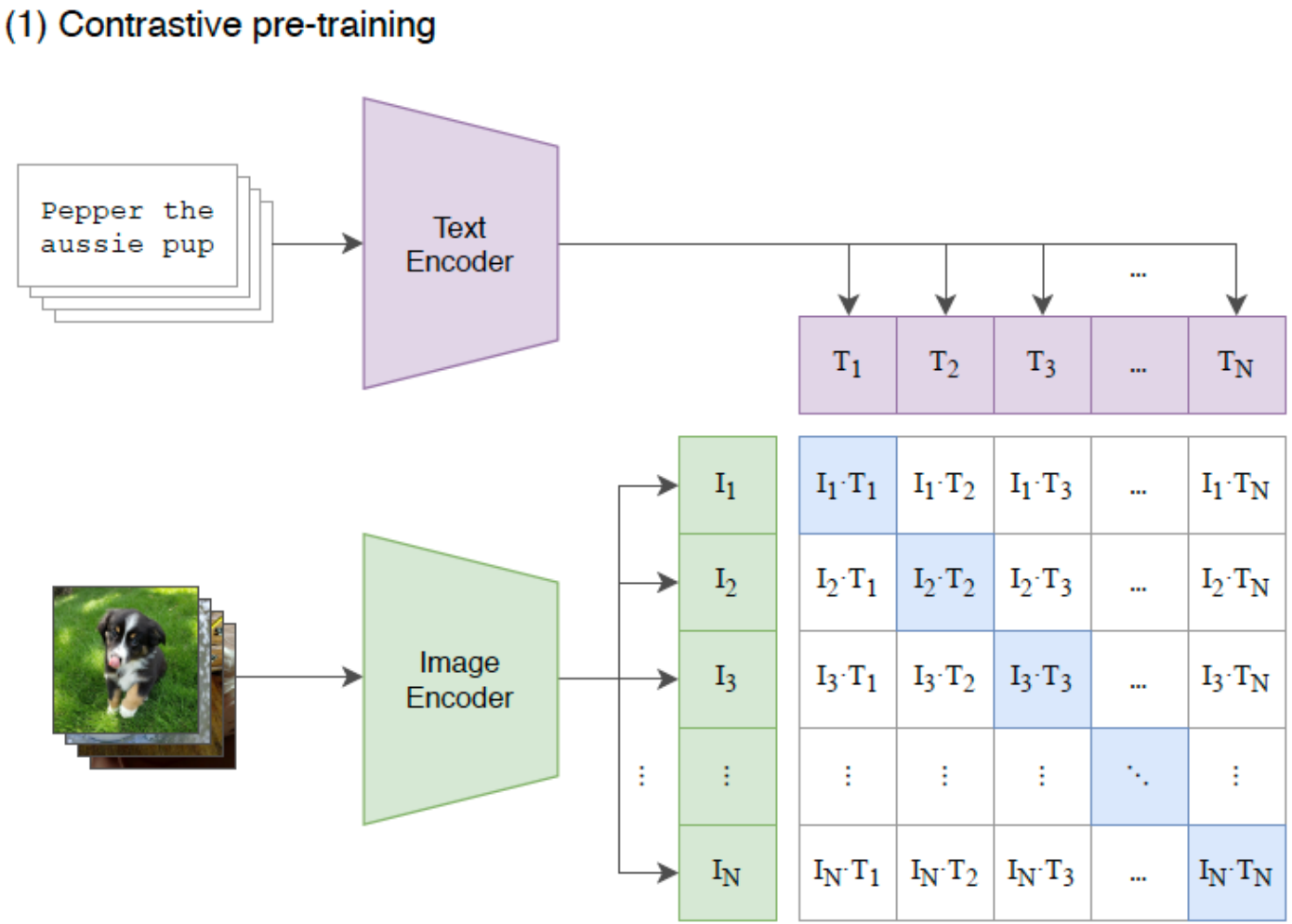
Vision – language strategies

- Shared embedding space
  - Embed image and language in shared space
    - corresponding picture-caption pairs are close
    - non corresponding pairs are far apart
  - CLIP

Association (CLIP)

- 4e8 pairs (image, text) from internet queries
  - queries:
    - all words appearing 100 times or more in English wikipedia
      - also, some bigrams
- Train ViT, Text encoder so that
  - matches are close
  - embeddings spread out
- Larger models
  - ResNet: 18 days on 592 V100’s
  - VIT: 12 days on 256 V100’s

CLIP: joint vision language training



Learning Transferable Visual Models From Natural Language Supervision, Radford et al 21

Evaluating encoder

- Take some dataset
- Train a linear probe
  - multiclass linear classifier
  - applied to embedding
  - trained on some training split
- Look at score, averaged over datasets
  - Better at a given compute;
  - improves with more compute

Better than best ImageNet model

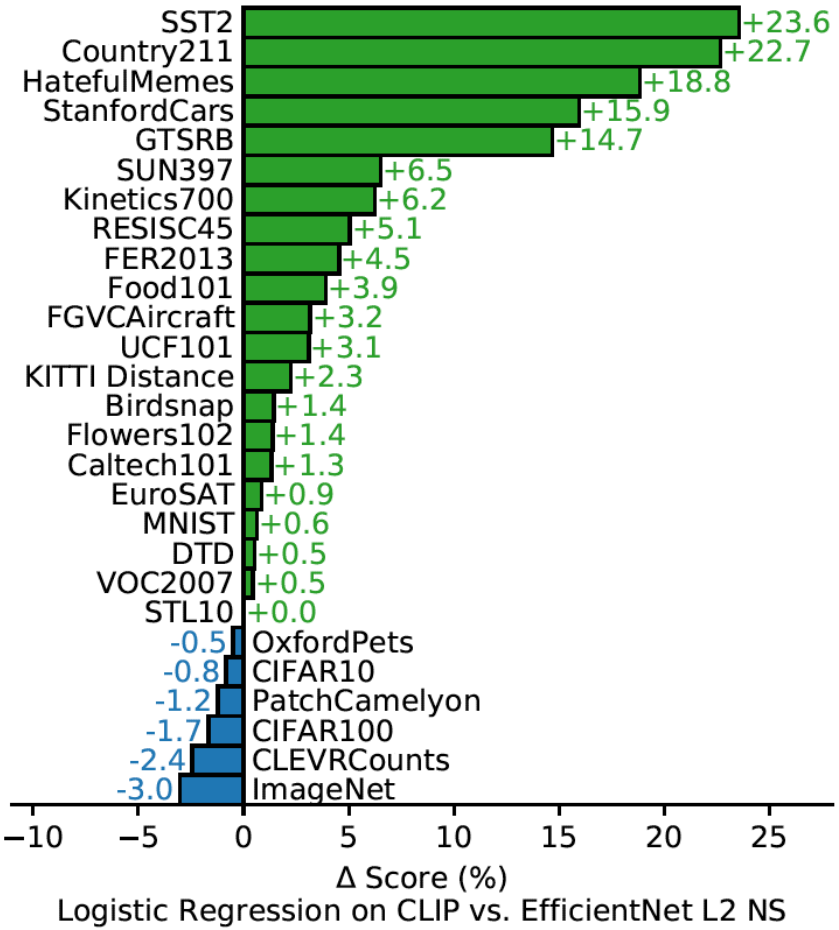
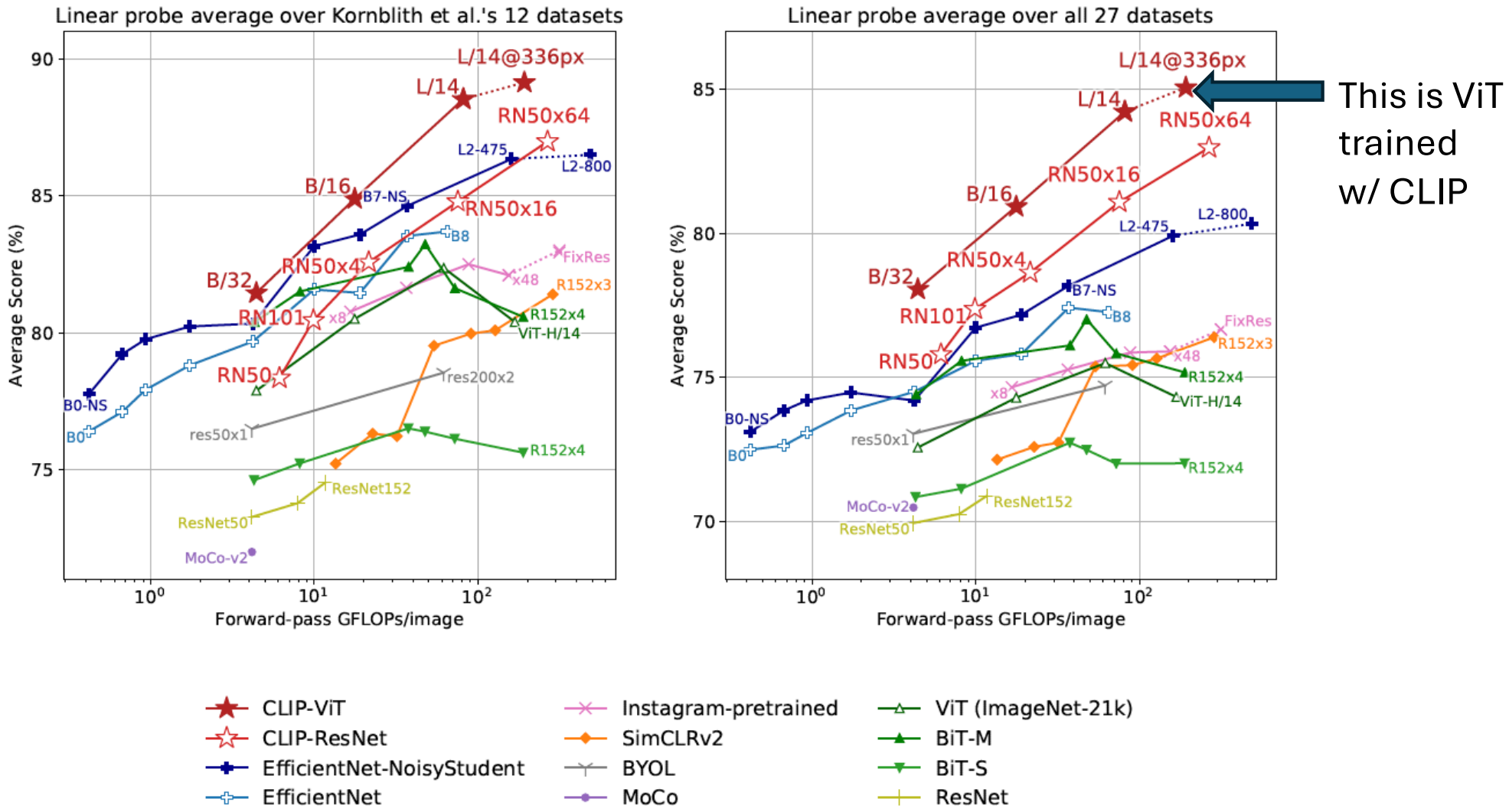


Figure 11. **CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.** Fitting a linear classifier on CLIP's features outperforms using the Noisy Student EfficientNet-L2 on 21 out of 27 datasets.

Better at a given compute;
improves with more compute



This is ViT trained w/ CLIP

Figure 10. **Linear probe performance of CLIP models in comparison with state-of-the-art computer vision models**, including EfficientNet (Tan & Le, 2019; Xie et al., 2020), MoCo (Chen et al., 2020d), Instagram-pretrained ResNeXt models (Mahajan et al., 2018; Touvron et al., 2019), BiT (Kolesnikov et al., 2019), ViT (Dosovitskiy et al., 2020), SimCLRv2 (Chen et al., 2020c), BYOL (Grill et al., 2020), and the original ResNet models (He et al., 2016b). (Left) Scores are averaged over 12 datasets studied by Kornblith et al. (2019) (Right) Scores are averaged over 27 datasets that contain a wider variety of distributions. Dotted lines indicate models fine-tuned or evaluated on images at a higher-resolution than pre-training. See Table 10 for individual scores and Figure 20 for plots for each dataset.

Learning Transferable Visual Models From Natural Language Supervision, Radford et al 21

Zero-shot classification in CLIP

- build a list of classes
- classify image by
  - compute embedding of “class” on language side
    - e-class
  - compute (image, e-class) distance for each class
  - choose the closest

Works

- Terminology:
  - linear probe = fit linear classifier to features produced by encoder

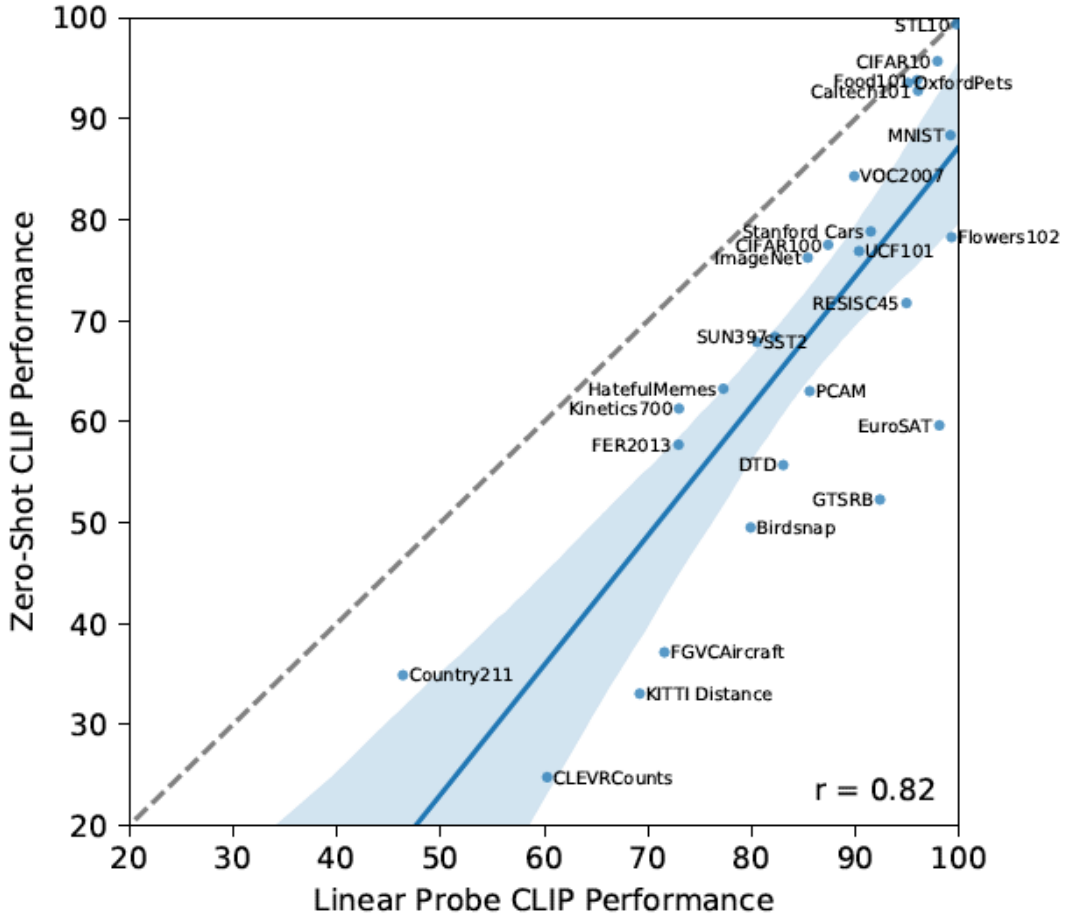


Figure 8. **Zero-shot performance is correlated with linear probe performance but still mostly sub-optimal.** Comparing zero-shot and linear probe performance across datasets shows a strong correlation with zero-shot performance mostly shifted 10 to 25 points lower. On only 5 datasets does zero-shot performance approach linear probe performance ($\leq$3 point difference).

Examples are seldom better

- For most datasets, you need more than one example to fit a linear classifier that is better than zero-shot

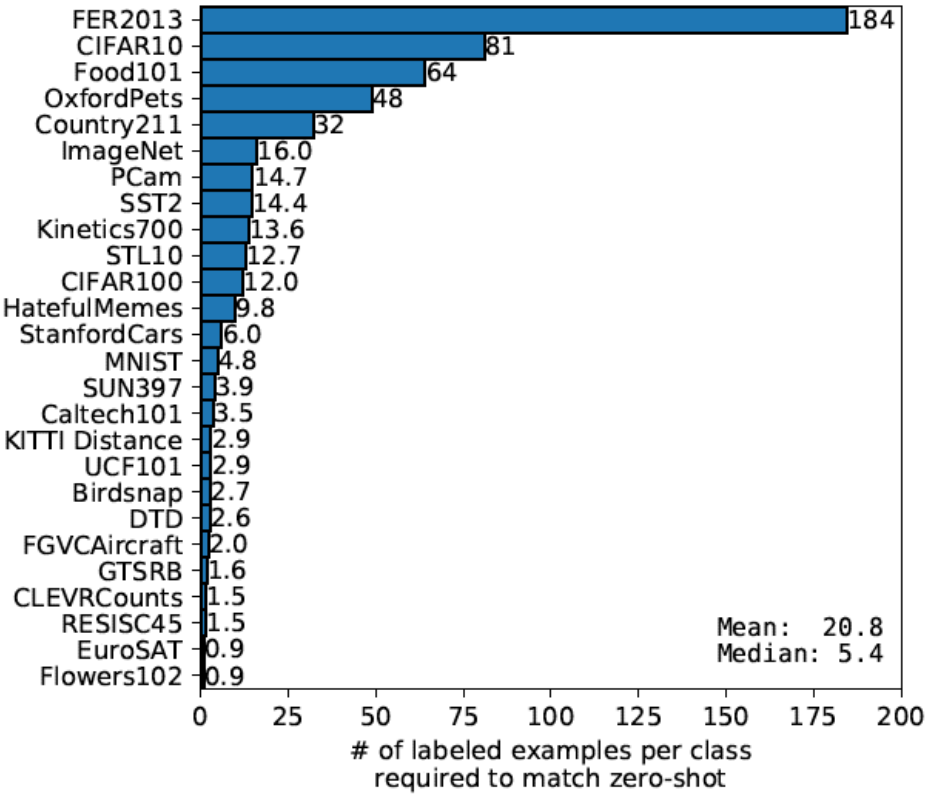


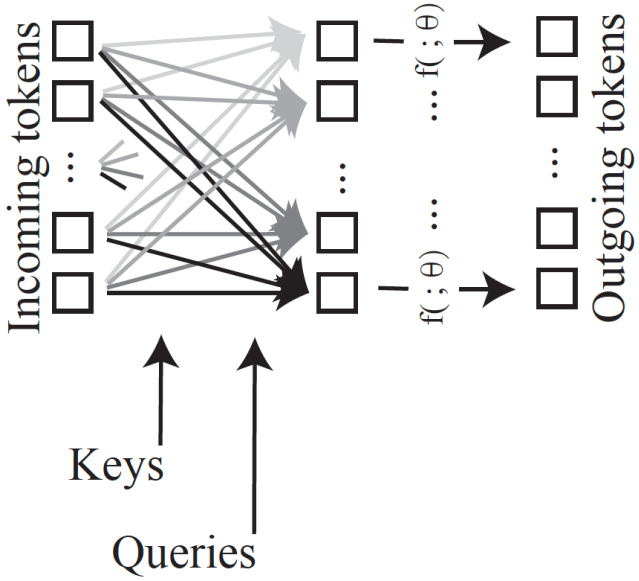
Figure 7. **The data efficiency of zero-shot transfer varies widely.** Calculating the number of labeled examples per class a linear classifier on the same CLIP feature space requires to match the performance of the zero-shot classifier contextualizes the effectiveness of zero-shot transfer. Values are estimated based on log-linear interpolation of 1, 2, 4, 8, 16-shot and fully supervised results. Performance varies widely from still underperforming a one-shot classifier on two datasets to matching an estimated 184 labeled examples per class.

Vision – language strategies

- Cross-attention
  - Both models are transformers
    - pass tokens across models

Transformers again

These weights are determined by attention – depend on keys and queries.



- Details now matter

Softmax

Definition: 25.1 *The softmax function*

The function that maps a vector $\mathbf{g}$ to a normalized vector $\mathbf{n}$ by

$$n_i = \frac{e^{g_i}}{\sum_u e^{g_u}}$$

is often referred to as the *softmax* function. Notice that all components of $\mathbf{n}$ are non-negative, and $\mathbf{1}^T\mathbf{n} = 1$.

Definition: 25.2 *Softmax for matrices*

Write softmax($\mathcal{G}$) for the function that maps a matrix $\mathcal{G}$ to a matrix $\mathcal{W}$ where

$$w_{ij} = \frac{e^{g_{ij}}}{\sum_u e^{g_{iu}}}.$$

In this case, $\mathcal{W}\mathbf{1} = \mathbf{1}$. This means each row represents a set of *convex weights* (all non-negative, and sum to one; $\mathbf{1}$ is a vector of all ones).

Tokens, keys and queries

- Tokens
  - e_t = e_l dimensional vectors
  - stack into matrix

$$\mathcal{V} = \begin{bmatrix} \mathbf{v}_0^T \\ \mathbf{v}_1^T \\ \ldots \mathbf{v}_{N_t}^T \end{bmatrix}$$

- Keys and Queries
  - d_k dimensional vectors
    - There are N_t+1 of them
      - there are a variety of places they could come from
  - Stack into matrices

$$\mathcal{Q} = \begin{bmatrix} \mathbf{q}_0^T \\ \mathbf{q}_1^T \\ \ldots \mathbf{q}_{N_t}^T \end{bmatrix} \text{ and } \mathcal{K} = \begin{bmatrix} \mathbf{k}_0^T \\ \mathbf{k}_1^T \\ \ldots \mathbf{k}_{N_t}^T \end{bmatrix}$$

Attention

Procedure: 25.1 *Attention*

Given an $(N_t + 1) \times e_t$ matrix of tokens $\mathcal{V}$, an $(N_t + 1) \times d_k$ matrix of queries $\mathcal{Q}$ corresponding to those tokens, and a $(N_t + 1) \times d_k$ matrix of keys $\mathcal{K}$ corresponding to those tokens, attention computes a new set of tokens

$$\text{Attention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V}$$

This is an N_t+1 x N_t+1 non-negative matrix.
Rows sum to one.

Attention

Procedure: 25.1 *Attention*

Given an $(N_t + 1) \times e_t$ matrix of tokens $\mathcal{V}$, an $(N_t + 1) \times d_k$ matrix of queries $\mathcal{Q}$ corresponding to those tokens, and a $(N_t + 1) \times d_k$ matrix of keys $\mathcal{K}$ corresponding to those tokens, attention computes a new set of tokens

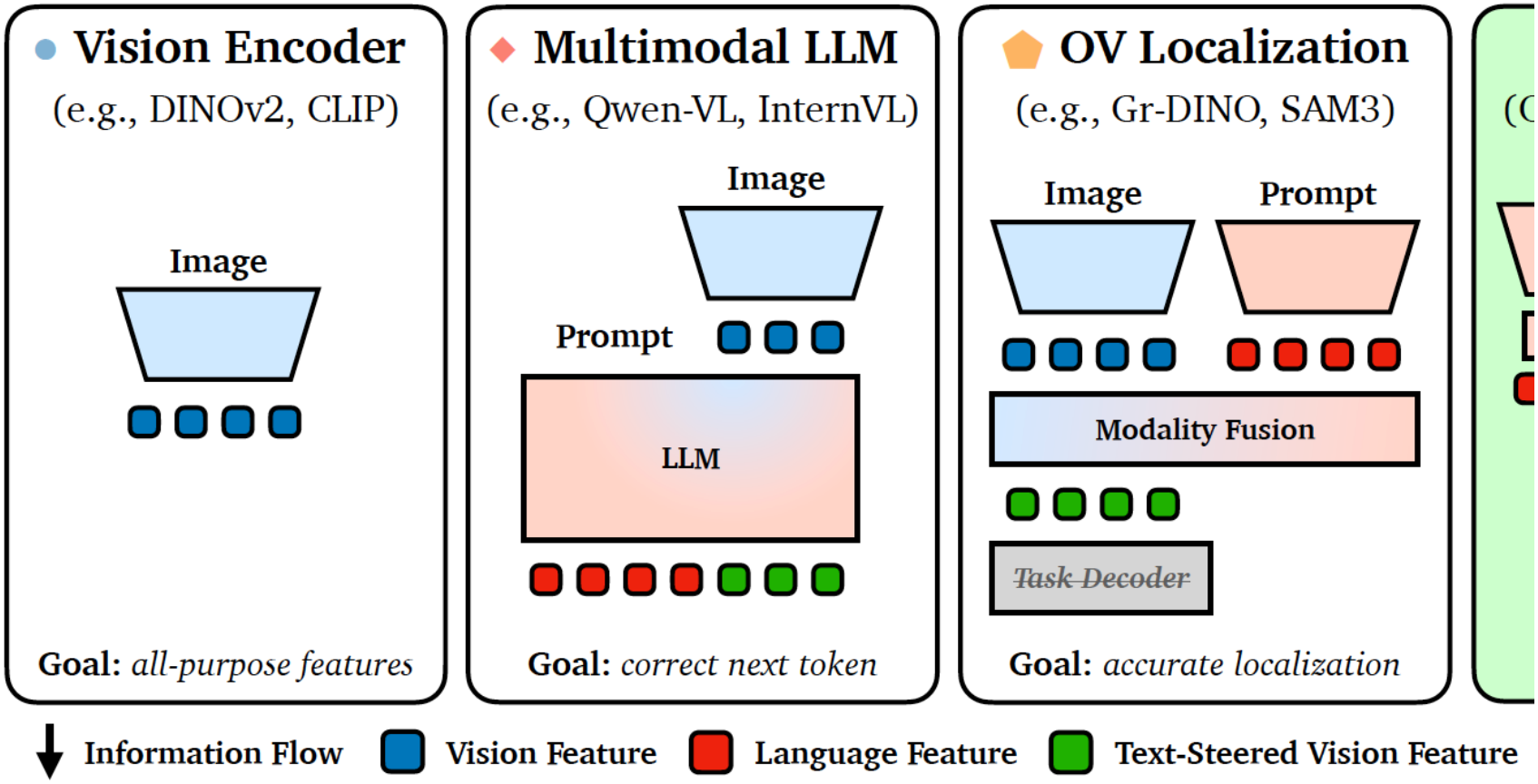
$$\text{Attention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V}$$

Result is an N_t+1 x e_t matrix - another matrix of tokens.

Each new token has received a contribution from all others.

Contributions are larger when key and query are similar, smaller when they are different.

Architectures

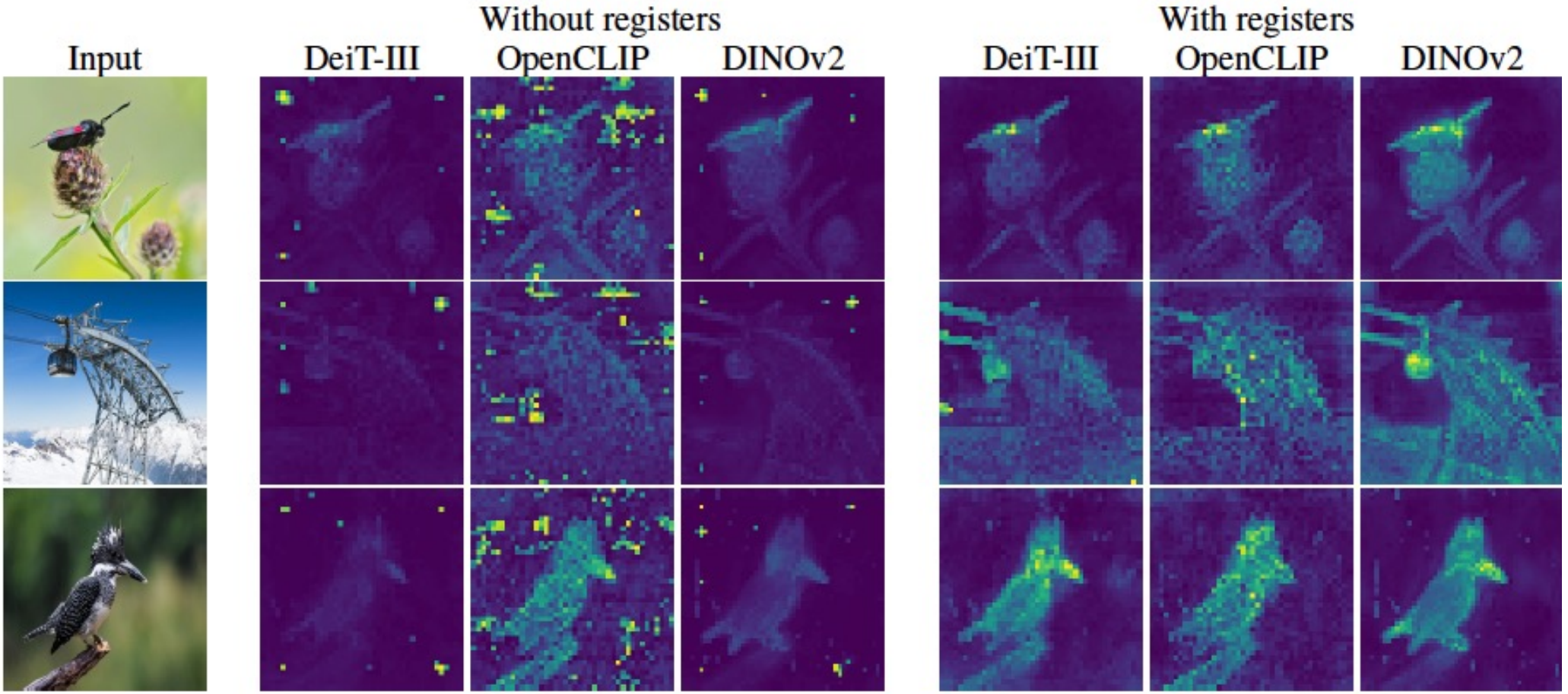


Steerable Visual Representations, Ruthard et al 2026

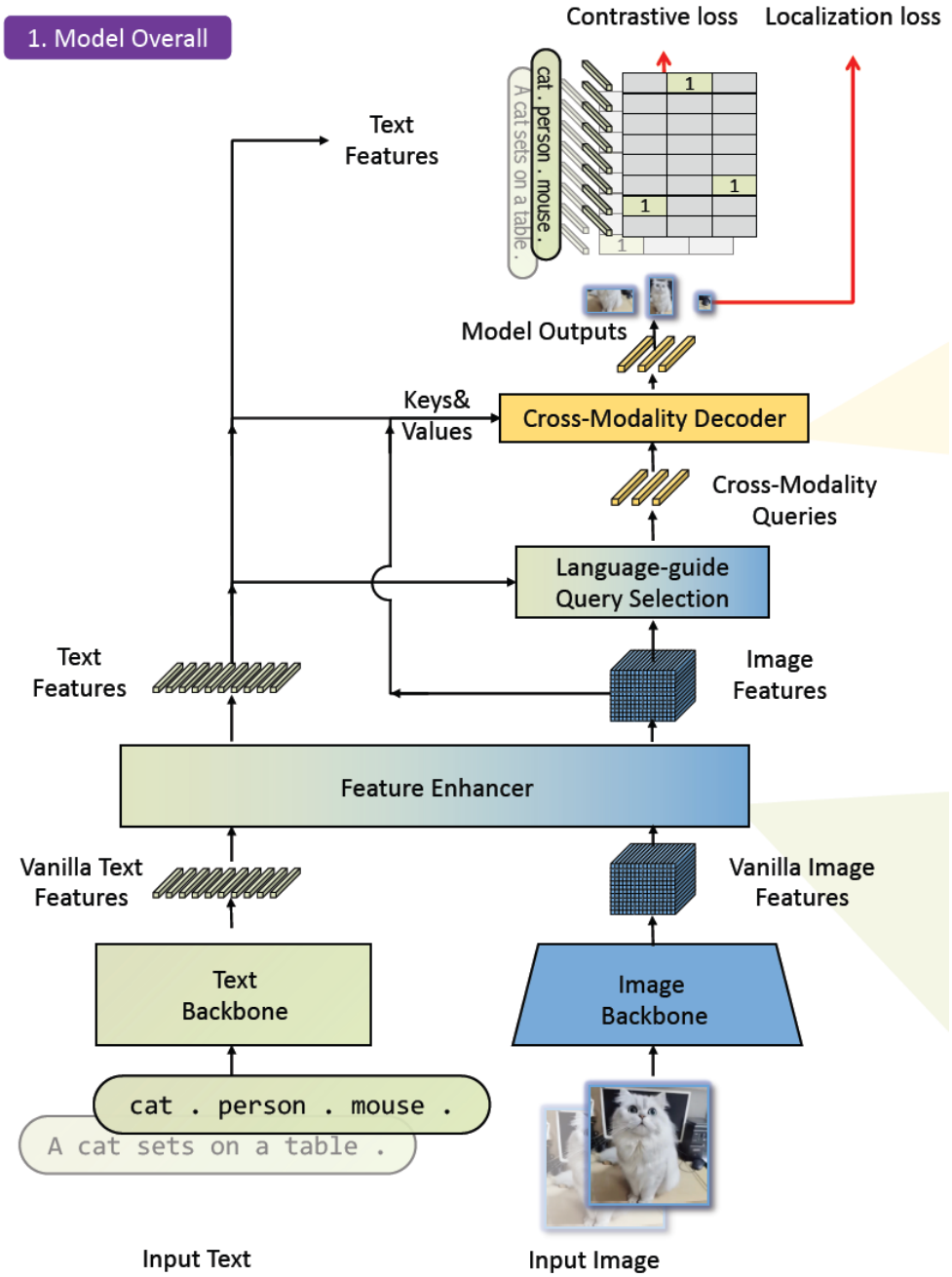
Attention to class token

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE Dosovitskiy et al, 2021

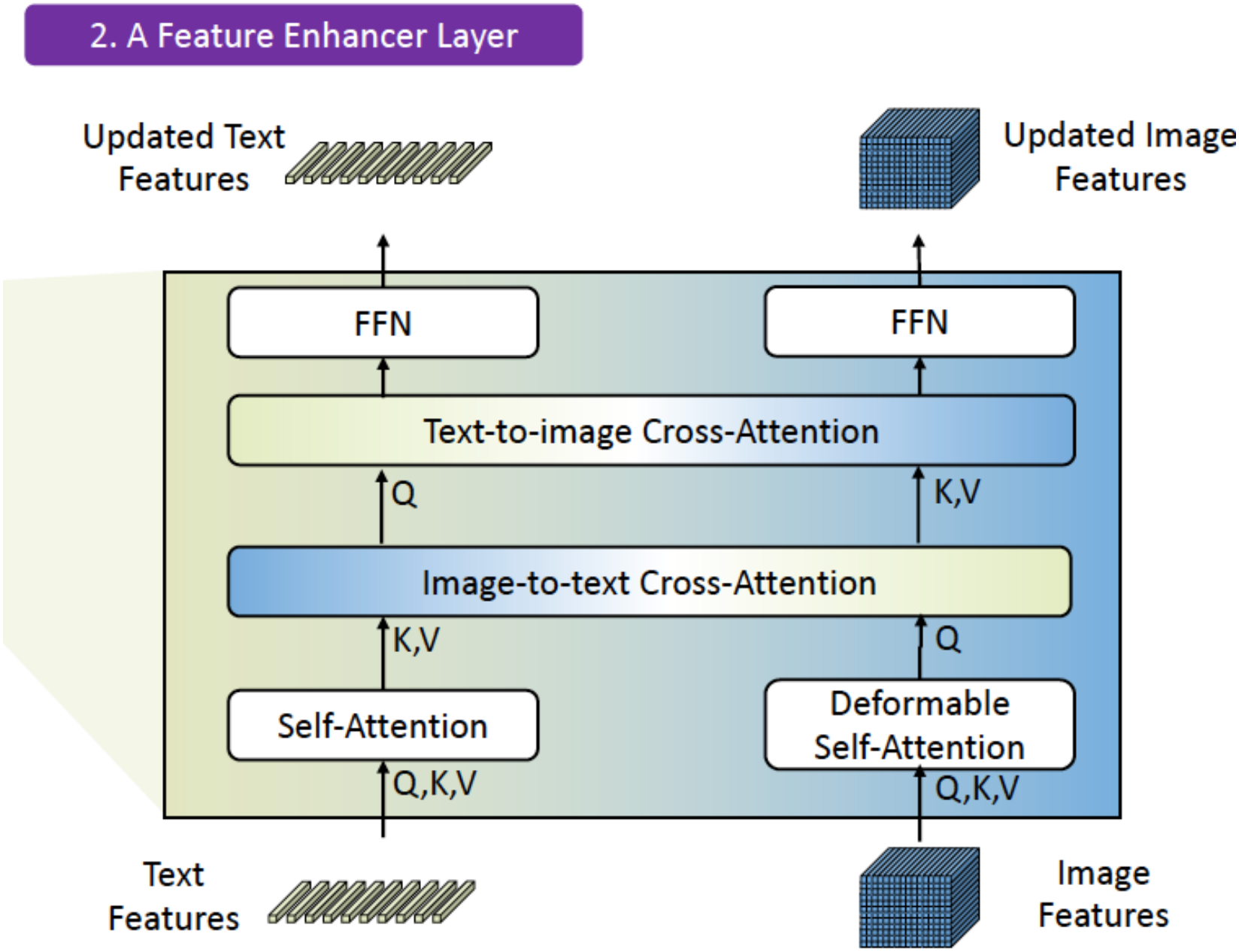
Attention to class token



VISION TRANSFORMERS NEED REGISTERS, Darcet et al, 2024



Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. Liu et al 2024



Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. Liu et al 2024

Evaluation strategies - I

1) *Answer Matching*

Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

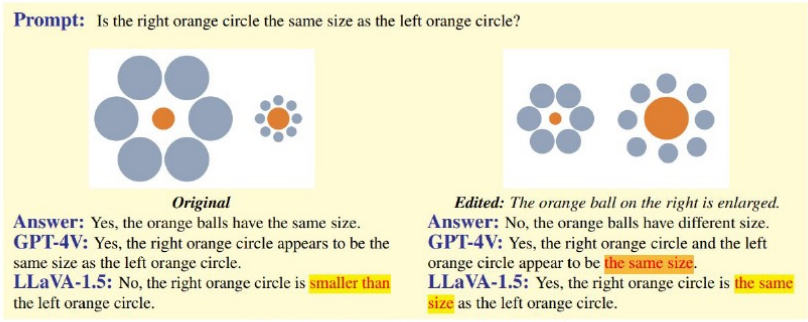
ST-VQA [20]

Evaluation: Accuracy
Metric: Exact Match
Format: Specific short-form answers, such as objects…

Q: Can you explain this meme?
GT: This meme is a humorous take on procrastination and the tendency to delay tasks until a specific time ...

MM-Vet [245]

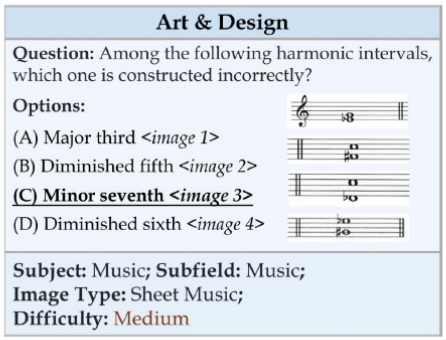
Evaluation: Average Score
Metric: ROUGE, LLM Eval
Format: Long-form open-ended answers



HallusionBench [67]

Evaluation: Accuracy / Precision / Recall
Metric: Exact Match
Format: Yes/No question

2) *Multiple Choice*



MMMU-Pro [248]

Evaluation: Accuracy
Metric: Exact Choice Match
Format: Multiple choice questions

3) *Image-Caption Similarity*

a brown bear? → 0.9925
a blue boat? → 0.9878

Score: **0.9804**

"A brown bear and a blue boat"

T2I-CompBench [63]

Evaluation: Average Similarity
Metric: CLIPScore, GenEval
Format: text to image generation

A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges, Li et al, 2025

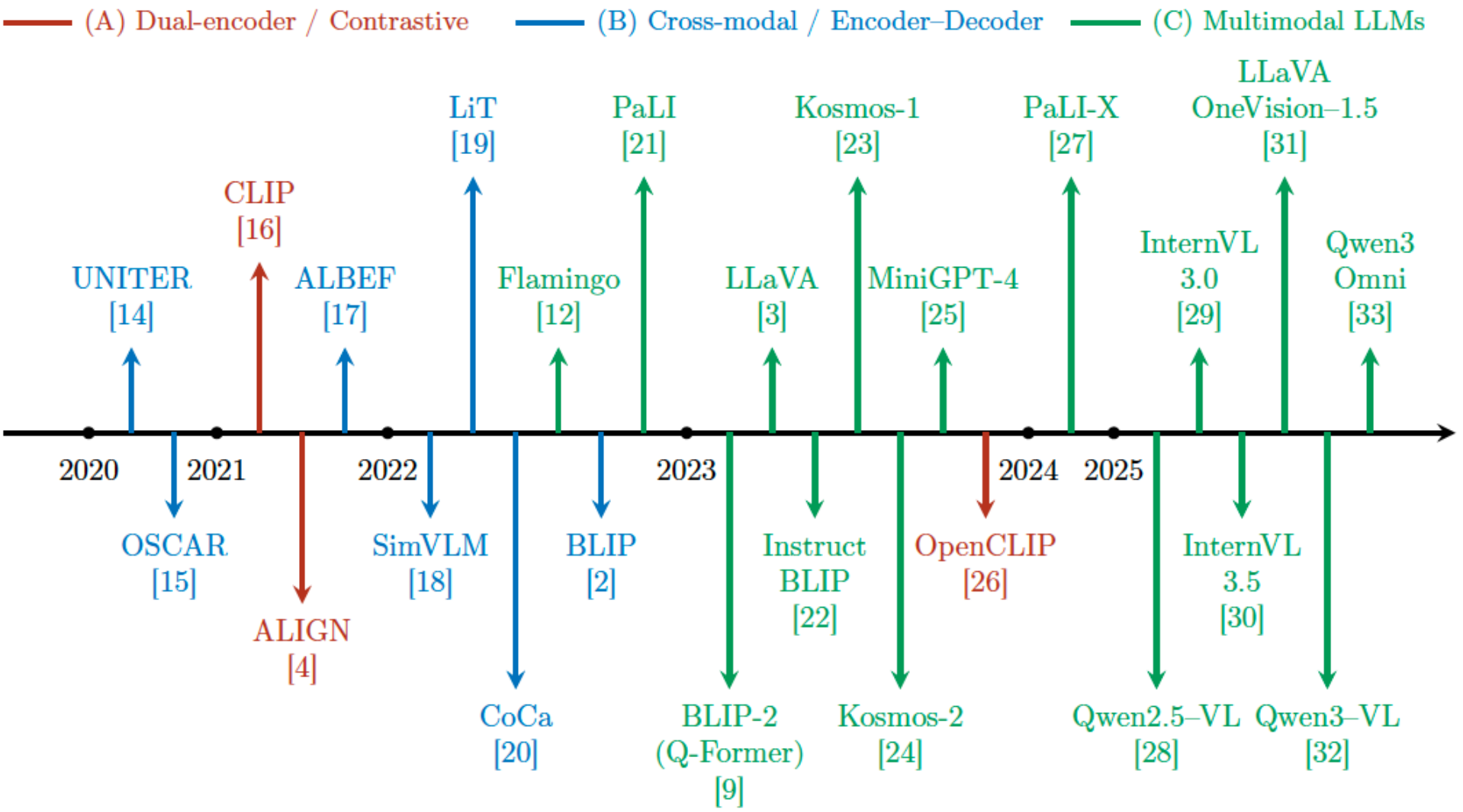


Figure 4: **Chronological overview of representative VLM / Multimodal LLM milestones (2020–2025).** Four color-coded categories: (A) Dual-encoder / Contrastive, (B) Cross-modal / Encoder–Decoder, (C) Multimodal LLMs. Each arrow is vertically offset within its year and labeled above/below with the model name and citation.

Vision Language Models: A Survey of 26K Papers (CVPR, ICLR, NeurIPS 2023–2025), Lin, 2025

Resources

- everything I cite is on arxiv
  - search by name
- Open VLM with pretrained weights, all code, open data:
  - https://allenai.org/molmo
- Survey in: *Vision-Language Models for Vision Tasks:A Survey, Zhang et al, 2024*
- Detailed models in:

QWEN TECHNICAL REPORT, Bai et al, 2023

Flamingo: a Visual Language Model for Few-Shot Learning, Alayrac, 2022

Molmo and PixMo: OpenWeights and Open Data for State-of-the-Art Vision-Language Models, Deitke et al, 2024