Deep Learning (COSE474)
2024 Fall

Final Project Topics

Hyunwoo J. Kim
(hyunwoojkim@korea.ac.kr)

Agenda

- Rules/Evaluation
- Possible project topics?

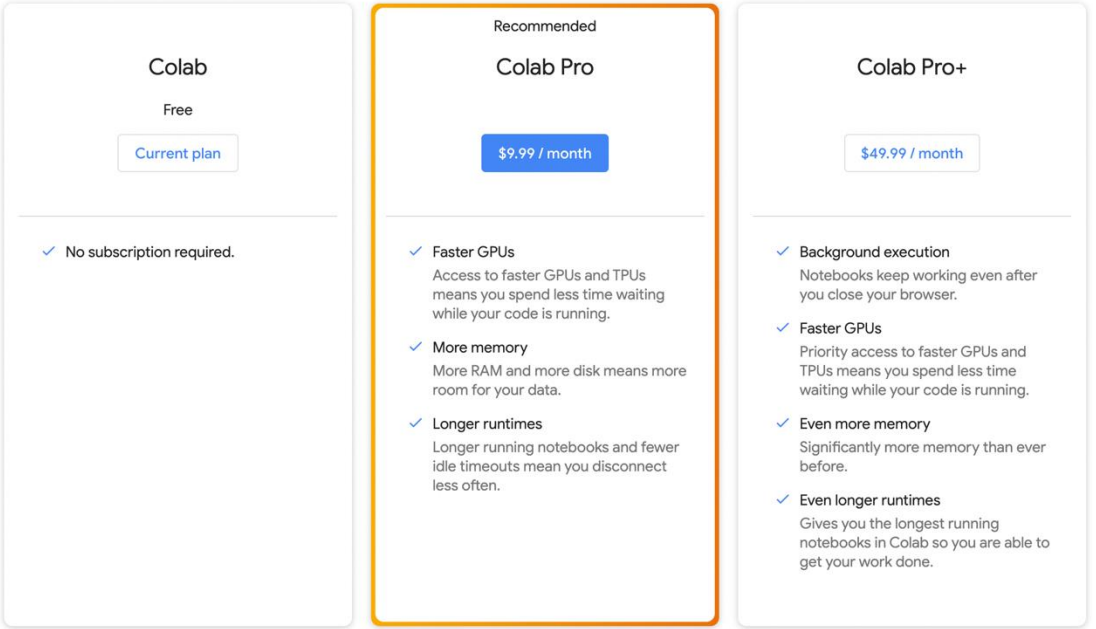
Rule and Evaluation

- Rule? Resources?
  - Computing resource: Colab/AWS/local GPU Servers (default: colab pro)
- Evaluation
  - Absolute Grading
  - We do NOT expect the state-of-the-art performance
  - We expect you to learn how to design an AI project and develop an AI algorithm.
  - It includes motivations, related works, methods (formulation, architecture), experiments (data preparation, hyperparameter tuning, quantitative/qualitative experimental results), discussion & future direction.

Computing resources

- Colab/Colab Pro/Colab Pro+/your own GPUs/AWS
- Describe your computing resources and reformulate your problem managable

- Colab Pro $10/month
- Colab Pro+ $50/month

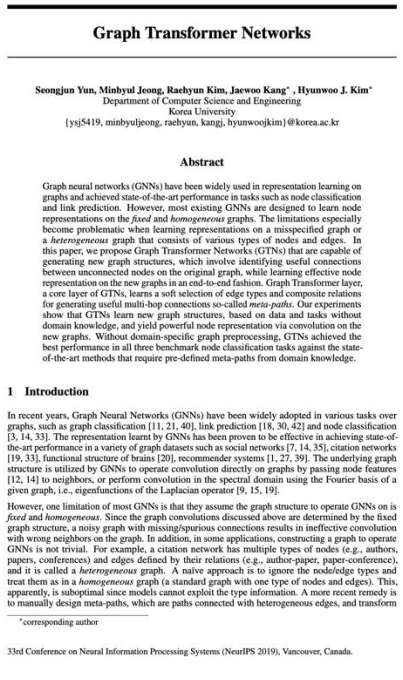


Evaluation:report (tentative)

- Evaluation (2024) – KOR or ENG
  - basic points (overall) (2): length (0.5), format (0.5), clarity of writing (1)
  - Introduction (5): motivation (2), problem definition (2), concise description of contribution (1)
  - Methods (5): significance/novelty (2), figure (1), reproducibility (2)-algorithm
  - Experiments (7): dataset (1), computer resource (CPU,GPU, OS, pytorch etc.) & experimental design (1), quantitative results (1), qualitative results (1), Figures (plots)/Tables and their analysis (2), discussion why the proposed method is successful or unsuccessful (1)
  - Future direction (1).
  - Github history (2)
  - Overleaf history (2)
  - (Bonus+1) pre-trained foundation models beyond ImageNet-pretrained CNNs (distillation, adaption, pseudo-labeling, baseline etc.), CLIP, BERT, RoBERTa

Good Examples for your final report

- Please check papers from top-tier AI conferences



Graph Transformer Networks

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang*, Hyunwoo J. Kim*
Department of Computer Science and Engineering
Korea University
{ysj5419, minbyuljeong, raehyun, kangj, hyunwoojkim}@korea.ac.kr

Abstract

Graph neural networks (GNNs) have been widely used in representation learning on graphs and achieved state-of-the-art performance in tasks such as node classification and link prediction. However, most existing GNNs are designed to learn node representations on the *fixed* and *homogeneous* graphs. The limitations especially become problematic when learning representations on a misspecified graph or a *heterogeneous* graph that consists of various types of nodes and edges. In this paper, we propose Graph Transformer Networks (GTNs) that are capable of generating new graph structures, which involve identifying useful connections between unconnected nodes on the original graph, while learning effective node representation on the new graphs in an end-to-end fashion. Graph Transformer layer, a core layer of GTNs, learns a soft selection of edge types and composite relations for generating useful multi-hop connections so-called *meta-paths*. Our experiments show that GTNs learn new graph structures, based on data and tasks without domain knowledge, and yield powerful node representation via convolution on the new graphs. Without domain-specific graph preprocessing, GTNs achieved the best performance in all three benchmark node classification tasks against the state-of-the-art methods that require pre-defined meta-paths from domain knowledge.

1 Introduction

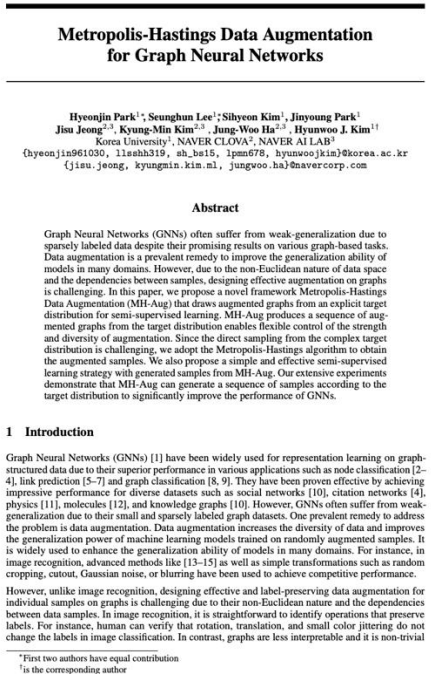
In recent years, Graph Neural Networks (GNNs) have been widely adopted in various tasks over graphs, such as graph classification [11, 21, 40], link prediction [18, 30, 42] and node classification [3, 14, 33]. The representation learnt by GNNs has been proven to be effective in achieving state-of-the-art performance in a variety of graph datasets such as social networks [7, 14, 35], citation networks [19, 33], functional structure of brains [20], recommender systems [1, 27, 39]. The underlying graph structure is utilized by GNNs to operate convolution directly on graphs by passing node features [12, 14] to neighbors, or perform convolution in the spectral domain using the Fourier basis of a given graph, i.e., eigenfunctions of the Laplacian operator [9, 15, 19].

However, one limitation of most GNNs is that they assume the graph structure to operate GNNs on is *fixed* and *homogeneous*. Since the graph convolutions discussed above are determined by the fixed graph structure, a noisy graph with missing/spurious connections results in ineffective convolution with wrong neighbors on the graph. In addition, in some applications, constructing a graph to operate GNNs is not trivial. For example, a citation network has multiple types of nodes (e.g., authors, papers, conferences) and edges defined by their relations (e.g., author-paper, paper-conference), and it is called a *heterogeneous* graph. A naive approach is to ignore the node/edge types and treat them as in a *homogeneous* graph (a standard graph with one type of nodes and edges). This, apparently, is suboptimal since models cannot exploit the type information. A more recent remedy is to manually design meta-paths, which are paths connected with heterogeneous edges, and transform

*corresponding author

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

NeurIPS 19



Metropolis-Hastings Data Augmentation for Graph Neural Networks

Hyeonjin Park[1]*, Seunghun Lee[1]*, Sihyeon Kim[1], Jinyoung Park[1]
Jisu Jeong[2,3], Kyung-Min Kim[2,3], Jung-Woo Ha[2,3], Hyunwoo J. Kim[1]†
Korea University[1], NAVER CLOVA[2], NAVER AI LAB[3]
{hyeonjin961030, llsshh319, sh_bs15, lpmn678, hyunwoojkim}@korea.ac.kr
{jisu.jeong, kyungmin.kim.ml, jungwoo.ha}@navercorp.com

Abstract

Graph Neural Networks (GNNs) often suffer from weak-generalization due to sparsely labeled data despite their promising results on various graph-based tasks. Data augmentation is a prevalent remedy to improve the generalization ability of models in many domains. However, due to the non-Euclidean nature of data space and the dependencies between samples, designing effective augmentation on graphs is challenging. In this paper, we propose a novel framework Metropolis-Hastings Data Augmentation (MH-Aug) that draws augmented graphs from an explicit target distribution for semi-supervised learning. MH-Aug produces a sequence of augmented graphs from the target distribution enables flexible control of the strength and diversity of augmentation. Since the direct sampling from the complex target distribution is challenging, we adopt the Metropolis-Hastings algorithm to obtain the augmented samples. We also propose a simple and effective semi-supervised learning strategy with generated samples from MH-Aug. Our extensive experiments demonstrate that MH-Aug can generate a sequence of samples according to the target distribution to significantly improve the performance of GNNs.

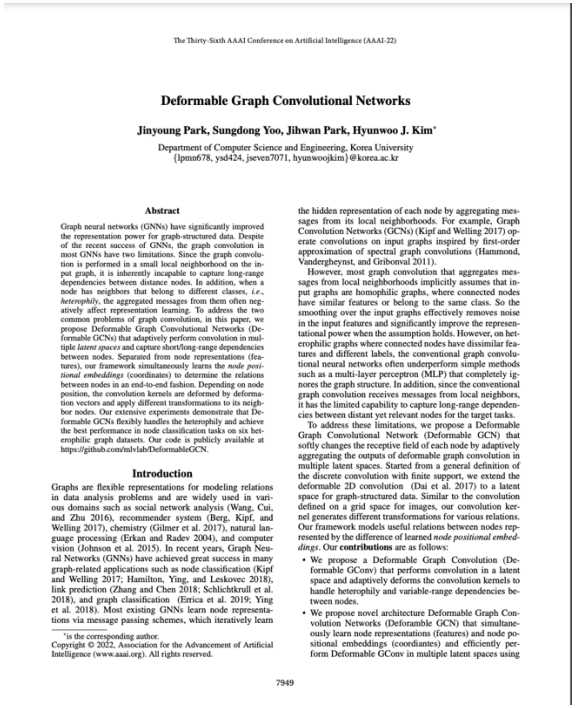
1 Introduction

Graph Neural Networks (GNNs) [1] have been widely used for representation learning on graph-structured data due to their superior performance in various applications such as node classification [2–4], link prediction [5–7] and graph classification [8, 9]. They have been proven effective by achieving impressive performance for diverse datasets such as social networks [10], citation networks [4], physics [11], molecules [12], and knowledge graphs [10]. However, GNNs often suffer from weak-generalization due to their small and sparsely labeled graph datasets. One prevalent remedy to address the problem is data augmentation. Data augmentation increases the diversity of data and improves the generalization power of machine learning models trained on randomly augmented samples. It is widely used to enhance the generalization ability of models in many domains. For instance, in image recognition, advanced methods like [13–15] as well as simple transformations such as random cropping, cutout, Gaussian noise, or blurring have been used to achieve competitive performance.

However, unlike image recognition, designing effective and label-preserving data augmentation for individual samples on graphs is challenging due to their non-Euclidean nature and the dependencies between data samples. In image recognition, it is straightforward to identify operations that preserve labels. For instance, human can verify that rotation, translation, and small color jittering do not change the labels in image classification. In contrast, graphs are less interpretable and it is non-trivial

*First two authors have equal contribution
†is the corresponding author

NeurIPS 21



The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

Deformable Graph Convolutional Networks

Jinyoung Park, Sungdong Yoo, Jihwan Park, Hyunwoo J. Kim*
Department of Computer Science and Engineering, Korea University
{lpmn678, ysd424, jseven7071, hyunwoojkim}@korea.ac.kr

Abstract

Graph neural networks (GNNs) have significantly improved the representation power for graph-structured data. Despite of the recent success of GNNs, the graph convolution in most GNNs have two limitations. Since the graph convolution is performed in a small local neighborhood on the input graph, it is inherently incapable to capture long-range dependencies between distance nodes. In addition, when a node has neighbors that belong to different classes, *i.e., heterophily*, the aggregated messages from them often negatively affect representation learning. To address the two common problems of graph convolution, in this paper, we propose Deformable Graph Convolutional Networks (Deformable GCNs) that adaptively perform convolution in multiple *latent spaces* and capture short/long-range dependencies between nodes. Separated from node representations (features), our framework simultaneously learns the *node positional embeddings* (coordinates) to determine the relations between nodes in an end-to-end fashion. Depending on node position, the convolution kernels are deformed by deformation vectors and apply different transformations to its neighbor nodes. Our extensive experiments demonstrate that Deformable GCNs flexibly handles the heterophily and achieve the best performance in node classification tasks on six heterophilic graph datasets. Our code is publicly available at https://github.com/mlvlab/DeformableGCN.

Introduction

Graphs are flexible representations for modeling relations in data analysis problems and are widely used in various domains such as social network analysis (Wang, Cui, and Zhu 2016), recommender system (Berg, Kipf, and Welling 2017), chemistry (Gilmer et al. 2017), natural language processing (Erkan and Radev 2004), and computer vision (Johnson et al. 2015). In recent years, Graph Neural Networks (GNNs) have achieved great success in many graph-related applications such as node classification (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2018), link prediction (Zhang and Chen 2018; Schlichtkrull et al. 2018), and graph classification (Errica et al. 2019; Ying et al. 2018). Most existing GNNs learn node representations via message passing schemes, which iteratively learn the hidden representation of each node by aggregating messages from its local neighborhoods. For example, Graph Convolution Networks (GCNs) (Kipf and Welling 2017) operate convolutions on input graphs inspired by first-order approximation of spectral graph convolutions (Hammond, Vandergheynst, and Gribonval 2011).

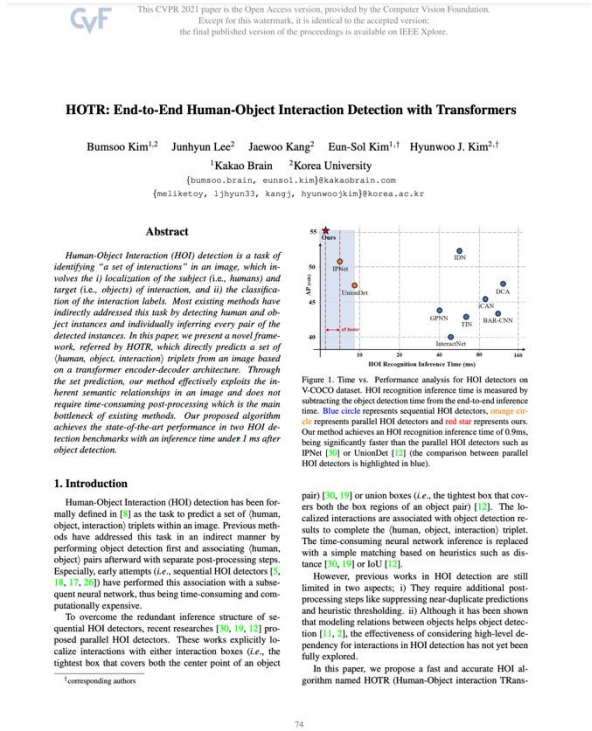
However, most graph convolution that aggregates messages from local neighborhoods implicitly assumes that input graphs are homophilic graphs, where connected nodes have similar features or belong to the same class. So the smoothing over the input graphs effectively removes noise in the input features and significantly improve the representational power when the assumption holds. However, on heterophilic graphs where connected nodes have dissimilar features and different labels, the conventional graph convolutional neural networks often underperform simple methods such as a multi-layer perceptron (MLP) that completely ignores the graph structure. In addition, since the conventional graph convolution receives messages from local neighbors, it has the limited capability to capture long-range dependencies between distant yet relevant nodes for the target tasks.

To address these limitations, we propose a Deformable Graph Convolutional Network (Deformable GCN) that softly changes the receptive field of each node by adaptively aggregating the outputs of deformable graph convolution in multiple latent spaces. Started from a general definition of the discrete convolution with finite support, we extend the deformable 2D convolution (Dai et al. 2017) to a latent space for graph-structured data. Similar to the convolution defined on a grid space for images, our convolution kernel generates different transformations for various relations. Our framework models useful relations between nodes represented by the difference of learned *node positional embeddings*. Our **contributions** are as follows:

- We propose a Deformable Graph Convolution (Deformable GConv) that performs convolution in a latent space and adaptively deforms the convolution kernels to handle heterophily and variable-range dependencies between nodes.
- We propose novel architecture Deformable Graph Convolution Networks (Deformable GCN) that simultaneously learn node representations (features) and node positional embeddings (coordiantes) and efficiently perform Deformable GConv in multiple latent spaces using

*is the corresponding author.

AAAI 22



HOTR: End-to-End Human-Object Interaction Detection with Transformers

Bumsoo Kim[1,2] Junhyun Lee[2] Jaewoo Kang[2] Eun-Sol Kim[1,†] Hyunwoo J. Kim[2,†]
[1]Kakao Brain [2]Korea University
{bumsoo.brain, eunsol.kim}@kakaobrain.com
{meliketoy, ljhyun33, kangj, hyunwoojkim}@korea.ac.kr

Abstract

Human-Object Interaction (HOI) detection is a task of identifying "a set of interactions" in an image, which involves the i) localization of the subject (i.e., humans) and target (i.e., objects) of interaction, and ii) the classification of the interaction labels. Most existing methods have indirectly addressed this task by detecting human and object instances and individually inferring every pair of the detected instances. In this paper, we present a novel framework, referred by HOTR, which directly predicts a set of ⟨human, object, interaction⟩ triplets from an image based on a transformer encoder-decoder architecture. Through the set prediction, our method effectively exploits the inherent semantic relationships in an image and does not require time-consuming post-processing which is the main bottleneck of existing methods. Our proposed algorithm achieves the state-of-the-art performance in two HOI detection benchmarks with an inference time under 1 ms after object detection.

Figure 1. Time vs. Performance analysis for HOI detectors on V-COCO dataset. HOI recognition inference time is measured by subtracting the object detection time from the end-to-end inference time. Blue circle represents sequential HOI detectors, orange circle represents parallel HOI detectors and red star represents ours. Our method achieves an HOI recognition inference time of 0.9ms, being significantly faster than the parallel HOI detectors such as IPNet [30] or UnionDet [12] (the comparison between parallel HOI detectors is highlighted in blue).

1. Introduction

Human-Object Interaction (HOI) detection has been formally defined in [8] as the task to predict a set of ⟨human, object, interaction⟩ triplets within an image. Previous methods have addressed this task in an indirect manner by performing object detection first and associating ⟨human, object⟩ pairs afterward with separate post-processing steps. Especially, early attempts (*i.e.*, sequential HOI detectors [5, 18, 17, 26]) have performed this association with a subsequent neural network, thus being time-consuming and computationally expensive.

To overcome the redundant inference structure of sequential HOI detectors, recent researches [30, 19, 12] proposed parallel HOI detectors. These works explicitly localize interactions with either interaction boxes (*i.e.*, the tightest box that covers both the center point of an object pair) [30, 19] or union boxes (*i.e.*, the tightest box that covers both the box regions of an object pair) [12]. The localized interactions are associated with object detection results to complete the ⟨human, object, interaction⟩ triplet. The time-consuming neural network inference is replaced with a simple matching based on heuristics such as distance [30, 19] or IoU [12].

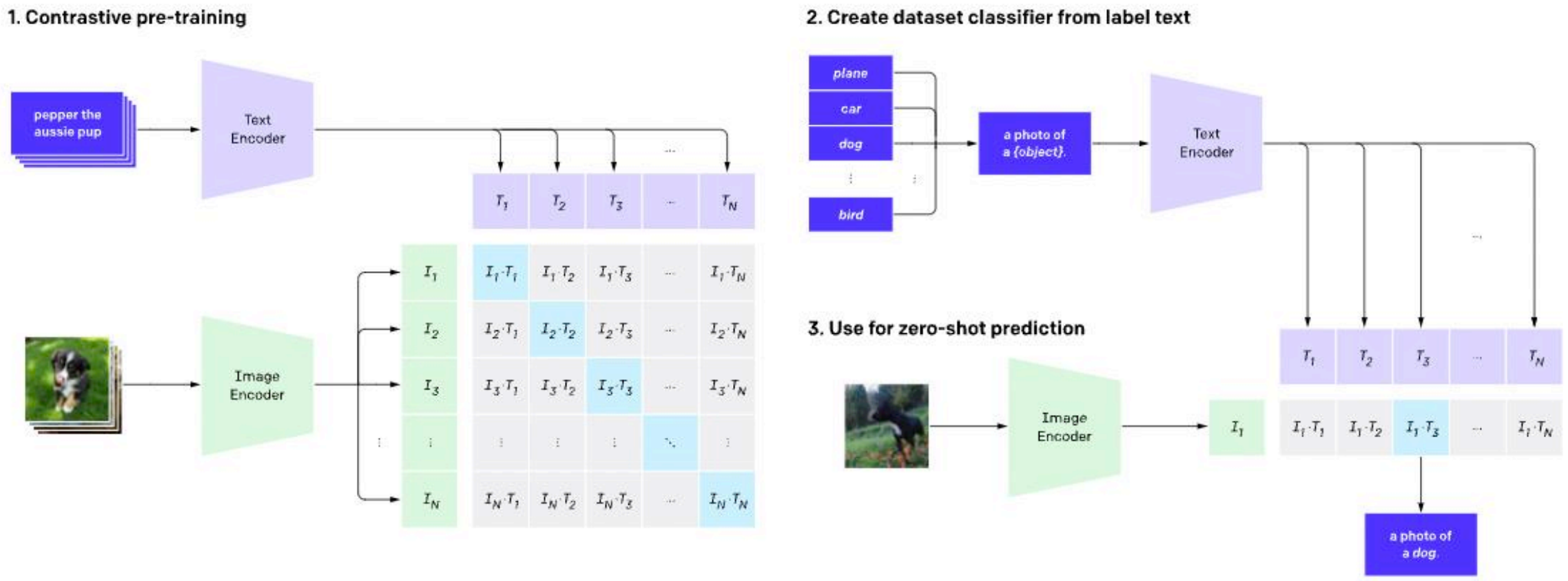
However, previous works in HOI detection are still limited in two aspects; i) They require additional post-processing steps like suppressing near-duplicate predictions and heuristic thresholding. ii) Although it has been shown that modeling relations between objects helps object detection [11, 2], the effectiveness of considering high-level dependency for interactions in HOI detection has not yet been fully explored.

In this paper, we propose a fast and accurate HOI algorithm named HOTR (Human-Object interaction TRans-

†corresponding authors

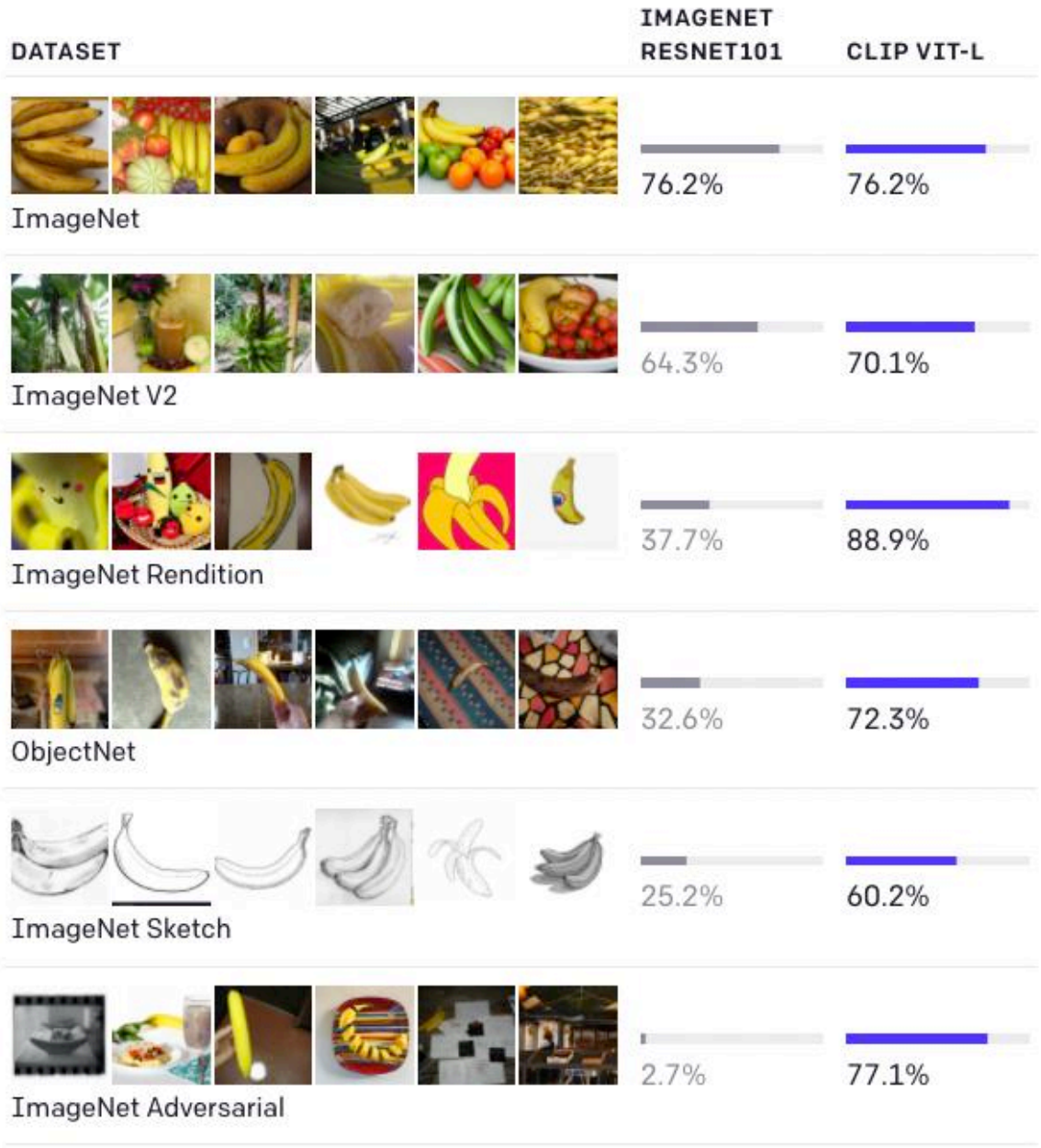
CVPR 21 (Oral)

Foundation Models
CLIP: Connecting Text and Images



CLIP: Connecting Text and Images

- CLIP which efficiently learns visual concepts from natural language supervision.
- CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the “zero-shot” capabilities of GPT-2 and GPT-3.



CALTECH-101

kangaroo (99.8%) Ranked 1 out of 102

✓ a photo of a **kangaroo**.

× a photo of a **gerenuk**.

× a photo of a **emu**.

× a photo of a **wild cat**.

× a photo of a **scorpion**.

IMAGENET-R (RENDITION)

Siberian Husky (76.0%) Ranked 1 out of 200

✓ a photo of a **siberian husky**.

× a photo of a **german shepherd dog**.

× a photo of a **collie**.

× a photo of a **border collie**.

× a photo of a **rottweiler**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23

✓ a photo of a **airplane**.

× a photo of a **bird**.

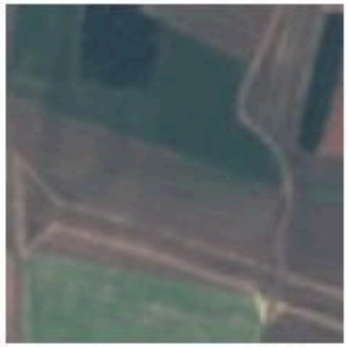
× a photo of a **bear**.

× a photo of a **giraffe**.

× a photo of a **car**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10



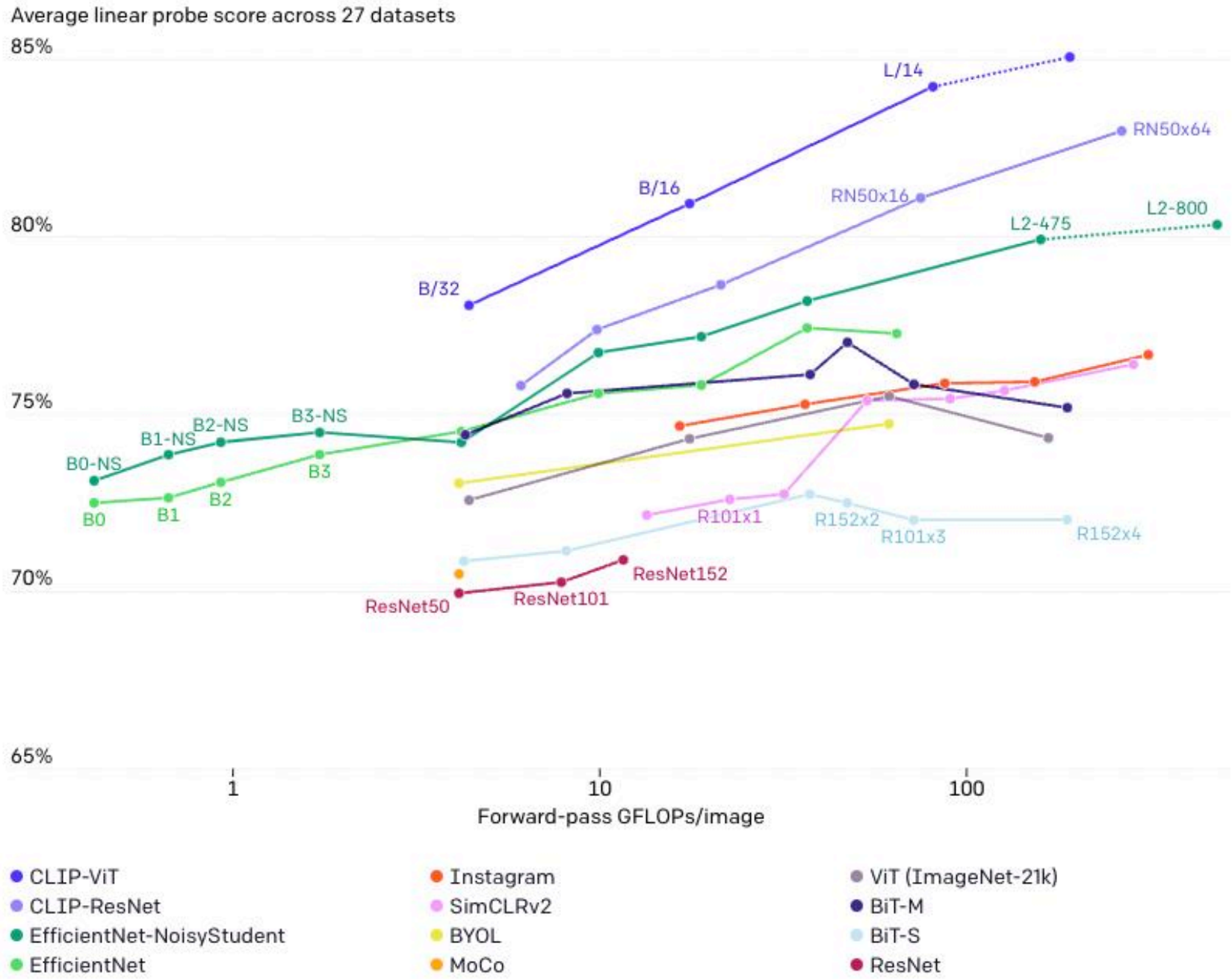
× a centered satellite photo of **permanent crop land**.

× a centered satellite photo of **pasture land**.

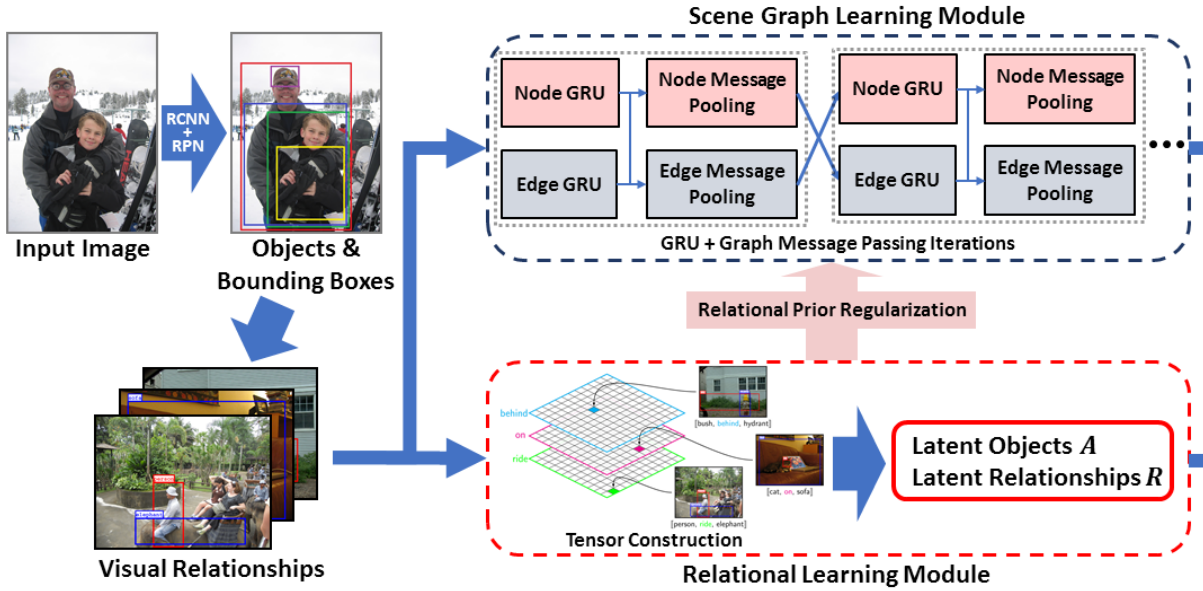
× a centered satellite photo of **highway or road**.

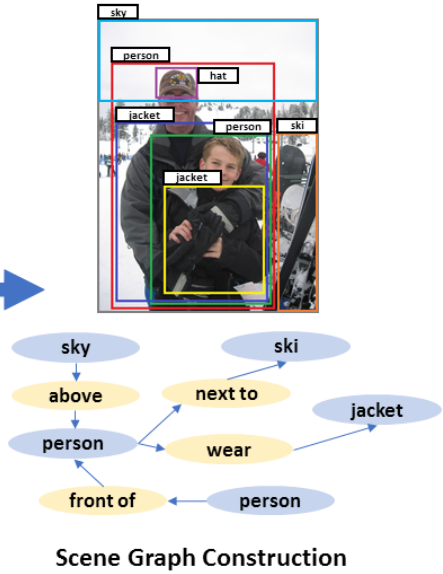
✓ a centered satellite photo of **annual crop land**.

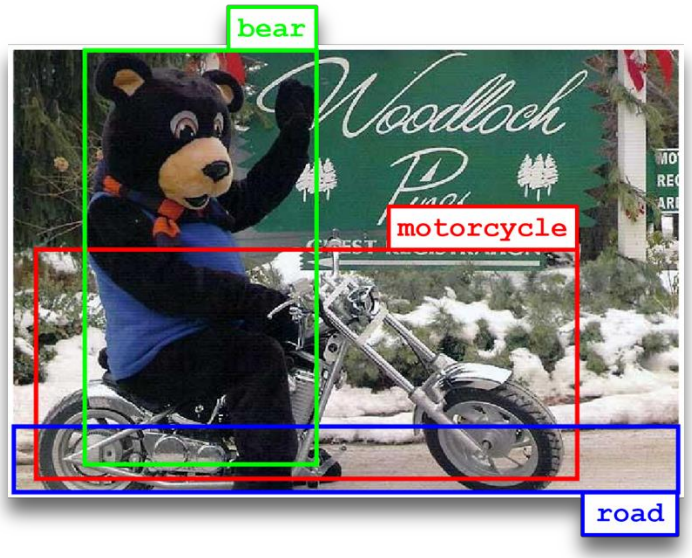
× a centered satellite photo of **brushland or shrubland**.

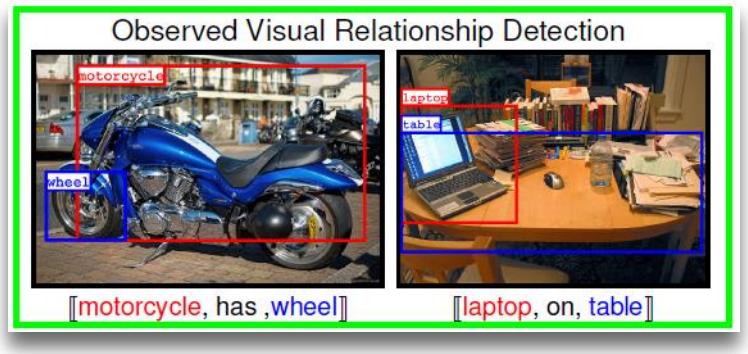

Average linear probe score across 27 datasets
85%
80%
75%
70%
65%
L/14
RN50x64
B/16
RN50x16
L2-800
L2-475
B/32
B3-NS
B2-NS
B1-NS
B0-NS
B3
B2
B1
B0
R101x1
R152x2
R101x3
R152x4
ResNet152
ResNet50
ResNet101
1
10
100
Forward-pass GFLOPs/image
CLIP-ViT
CLIP-ResNet
EfficientNet-NoisyStudent
EfficientNet
Instagram
SimCLRv2
BYOL
MoCo
ViT (ImageNet-21k)
BiT-M
BiT-S
ResNet

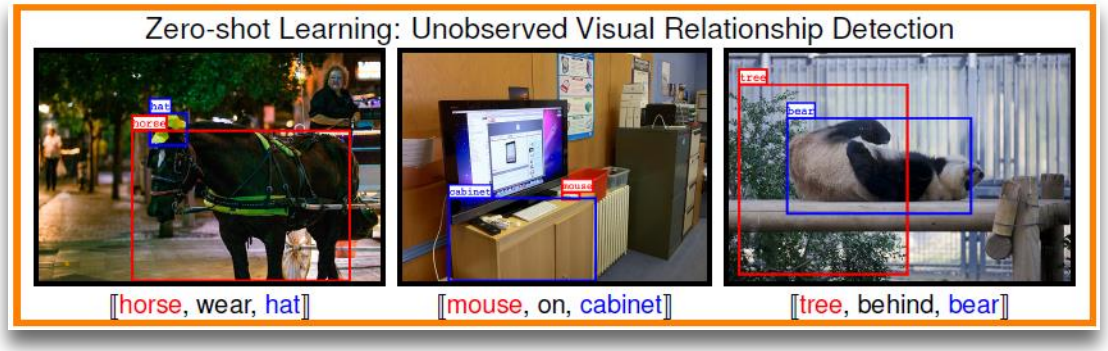
Scene understanding
Scene graph generation



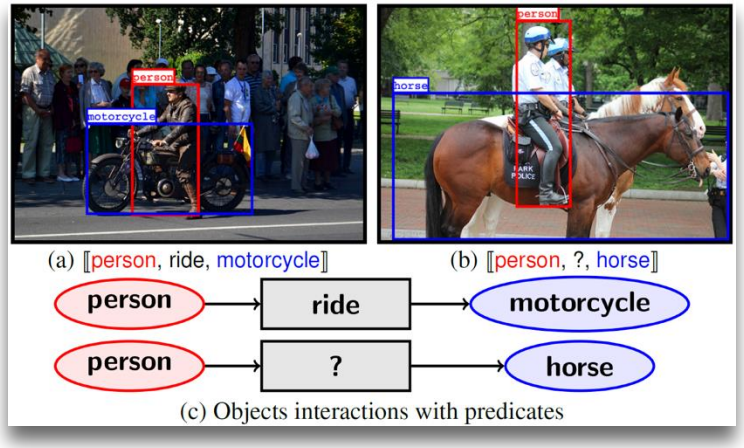




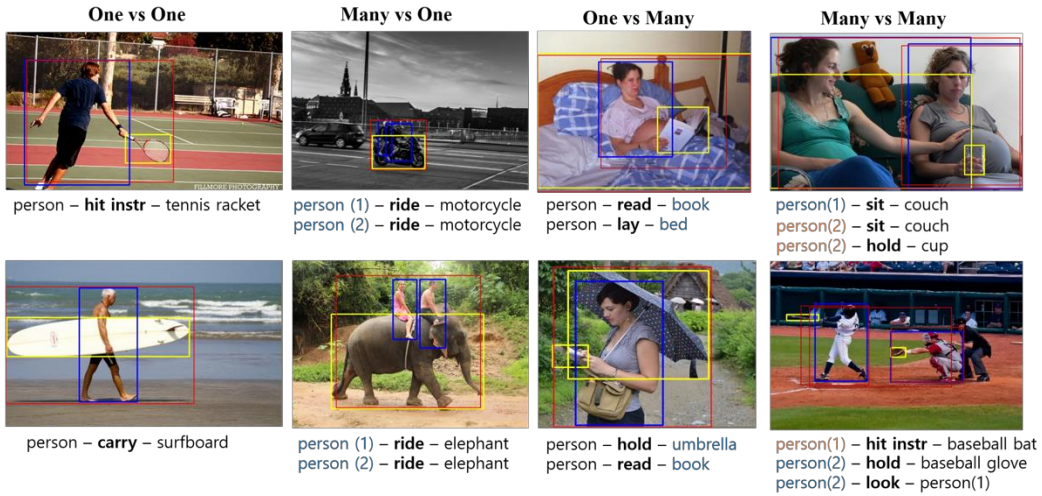




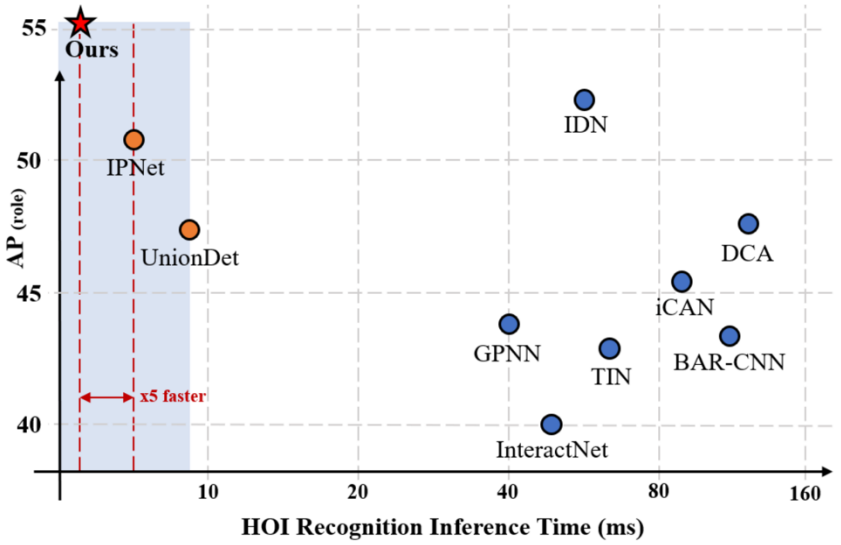
Scene understanding and object detection

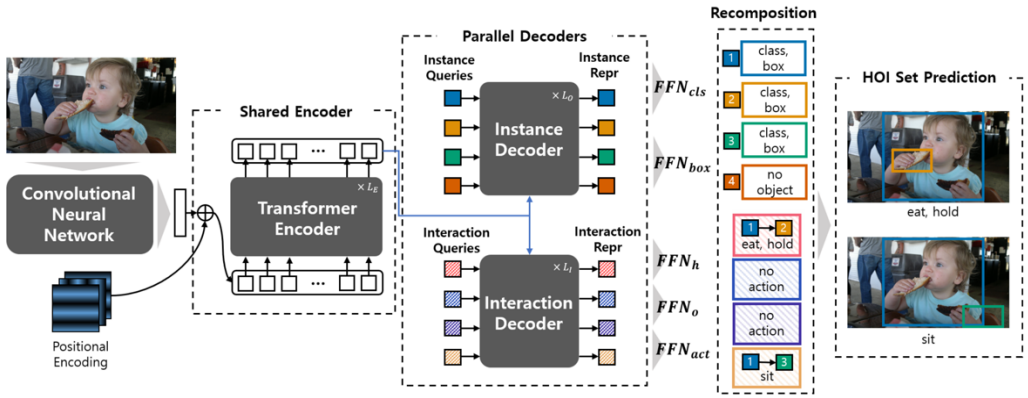


[CVPR18] Tensorize

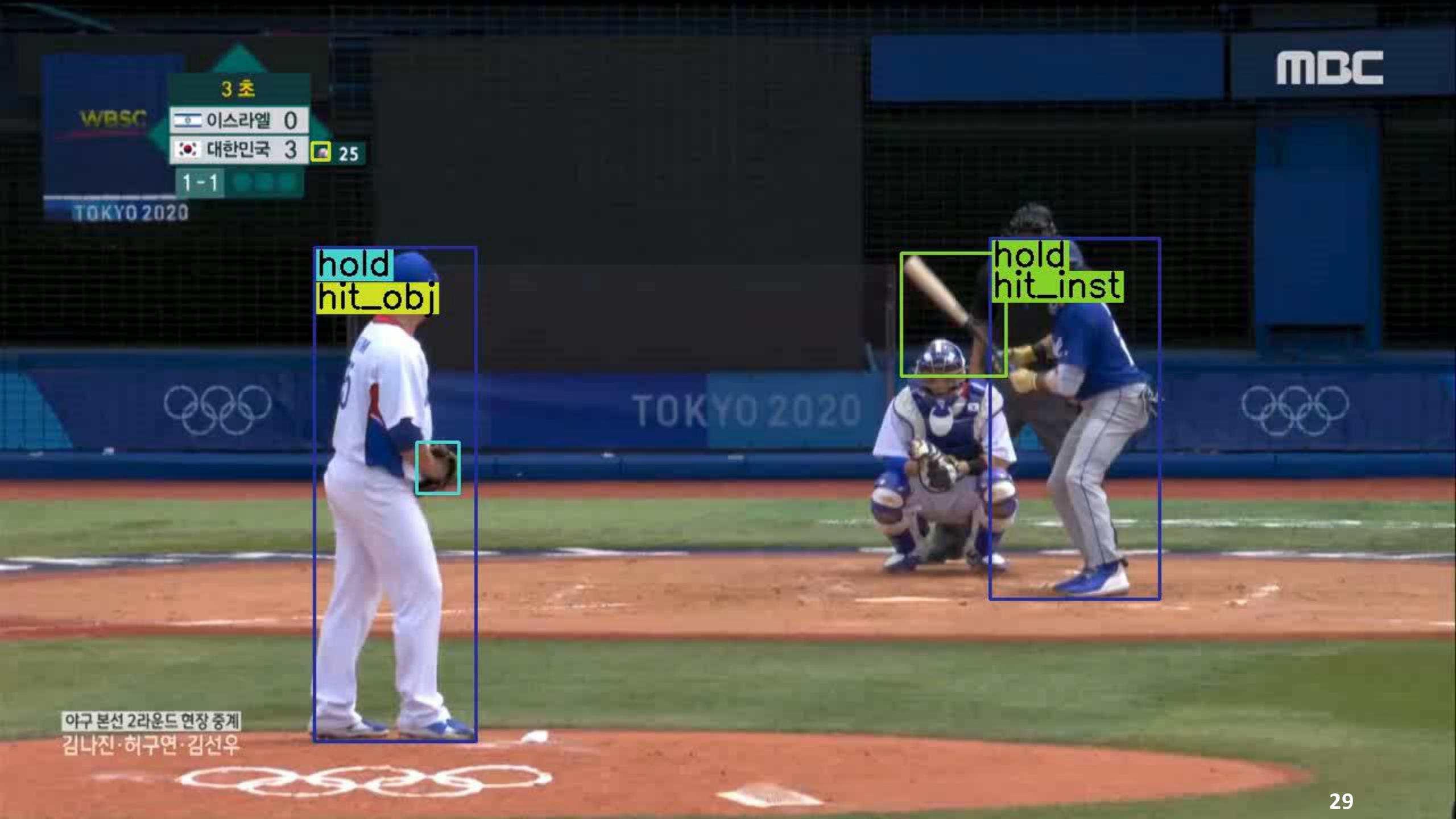


[ECCV20] UnionDet





[CVPR ’21 (ORAL)] HOTR


3 초
이스라엘 0
대한민국 3
25
1-1
TOKYO 2020
WBSC
MBC
hold
hit_obj
hold
hit_inst
TOKYO 2020
야구 본선 2라운드 현장 중계
김나진·허구연·김선우

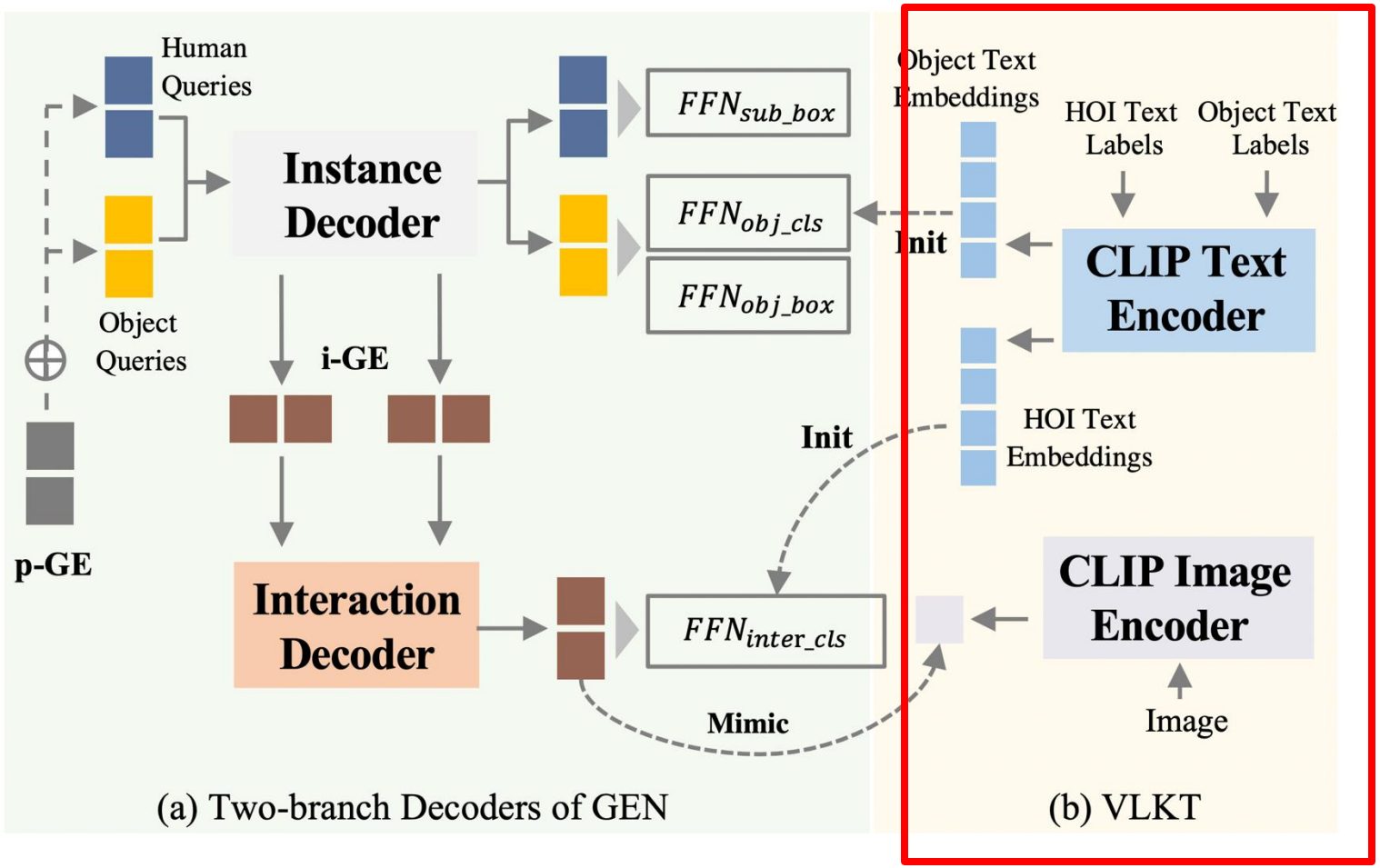
물체 탐지 강의 (데이터 과학원)

1. 물체 탐지 원리
[LInk]

2. 물체 탐지 기법
[Link]

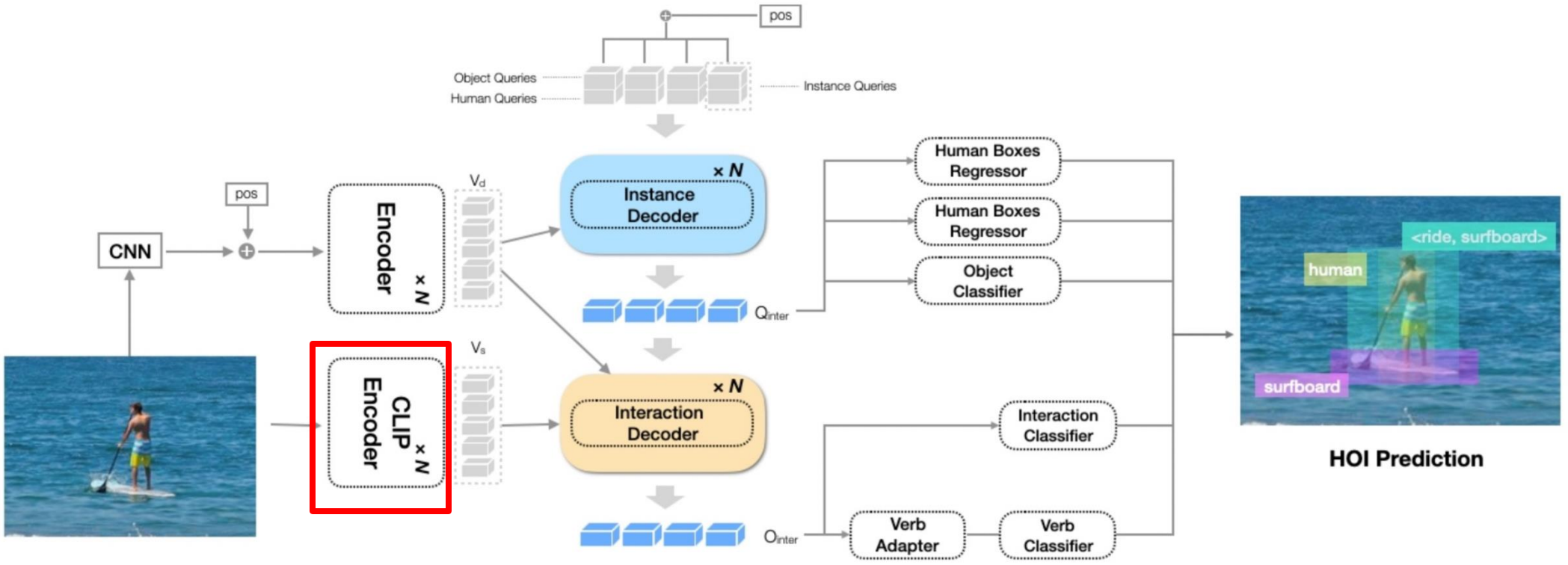
3. 물체간 관계 추론
[Link]

초거대 AI + 장면이해? Latent space alignment



[CVPR22] GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection

초거대 AI + 장면이해?



HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models

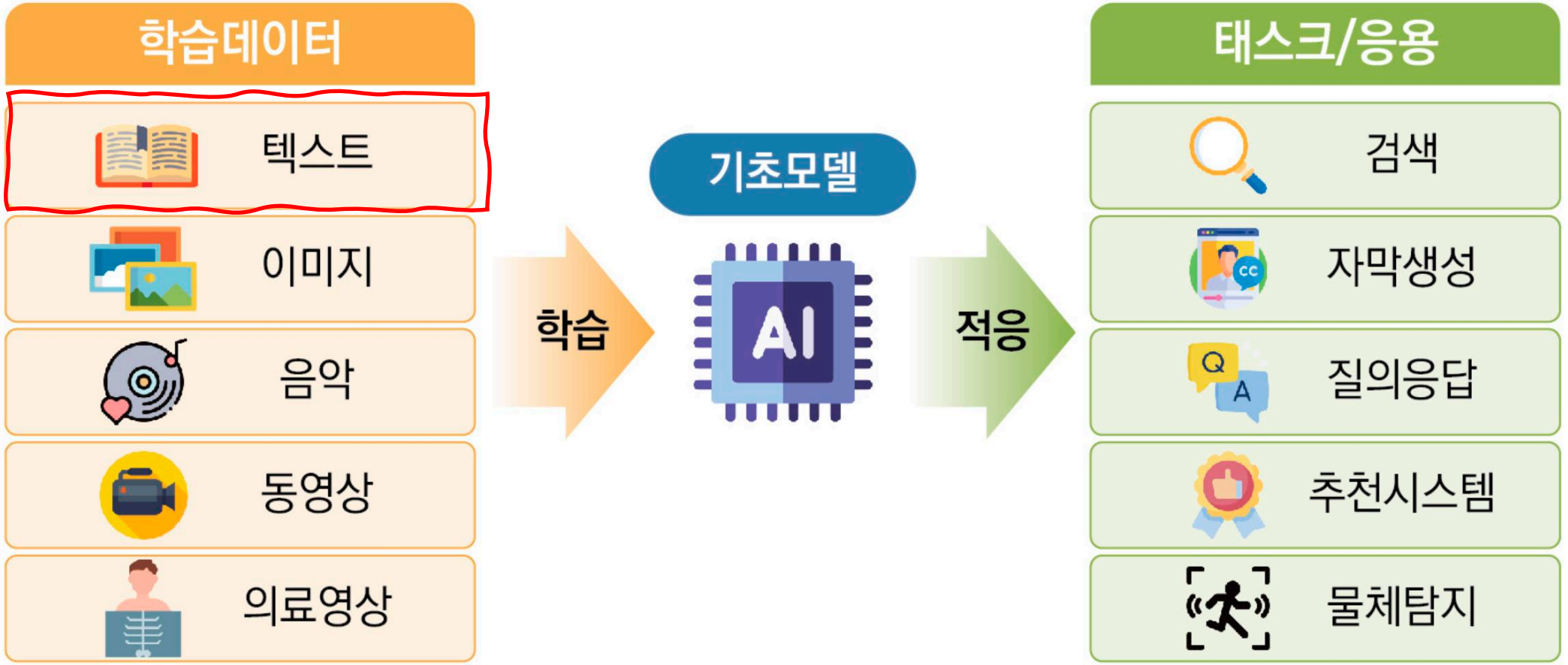
Examples: MLV top-tier conference papers (2023)

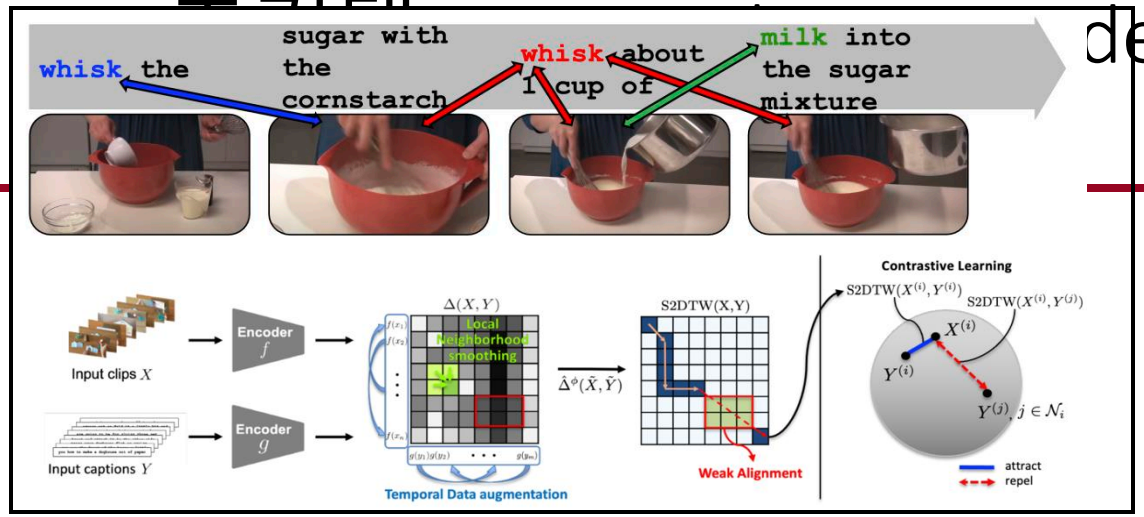
1. [EMNLP23] Large Language Models are Temporal and Causal Reasoners for Video Question Answering
2. [NeurIPS23] NuTrea: Neural Tree Search for Context-guided Multi-hop KGQA
3. [NeurIPS23] Advancing Bayesian Optimization via Learning Smooth Latent Spaces
4. [NeurIPS23] Unconstrained Pose Prior-Free Neural Radiance Field
5. [ICCV23] Open-Vocabulary Video Question Answering: A New Benchmark for Evaluating the Generalizability of Video Question Answering Models
6. [ICCV23] Distribution-Aware Prompt Tuning for Vision-Language Models
7. [ICCV23] Semantic-Aware Template Learning via Part Deformation Consistency
8. [ICCV23] Read-only Prompt Optimization for Vision-Language Few-shot Learning
9. [ICML23] Robust Camera Pose Refinement for Multi-Resolution Hash Encoding
10. [CVPR23] MELTR: Meta Loss Transformer for Learning to Fine-tune Video Foundation Models
11. [CVPR23] Self-positioning Point-based Transformer for Point Cloud Understanding
12. [AAAI23] Relation-aware Language-Graph Transformer for Question Answering

Examples: MLV top-tier conference papers (2024)

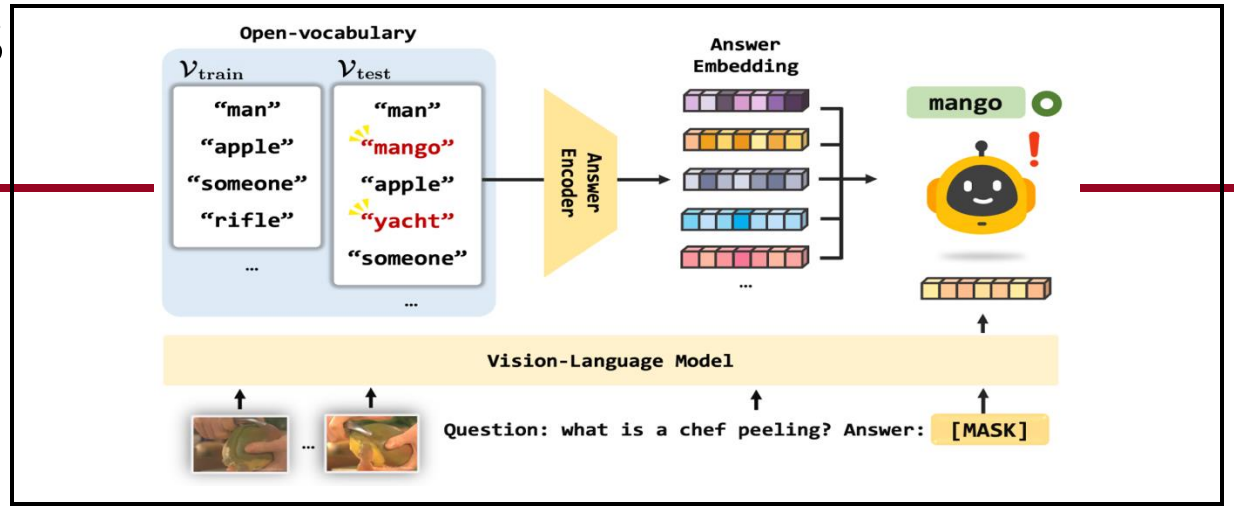
1. [NeurIPS] LLaMo: Large Language Model-based Molecular Graph Assistant
2. [NeurIPS] Alleviating Attention Bias for Visual-Informed Text Generation
3. [NeurIPS] Constant Acceleration Flow
4. [NeurIPS] Inversion-based Latent Bayesian Optimization
5. [EMNLP] Generative Subgraph Retrieval for Knowledge Graph–Grounded Dialog Generation
6. [ECCV] Diffusion Prior-Based Amortized Variational Inference for Noisy Inverse Problems
7. [ECCV] Understanding Multi-compositional learning in Vision and Language models via Category Theory
8. [ICML] Stochastic Conditional Diffusion Models for Robust Semantic Image Synthesis
9. [CVPR] vid-TLDR: Training Free Token merging for Light-weight Video Transformer
10. [CVPR] Multi-criteria Token Fusion with One-step-ahead Attention for Efficient Vision
11. [CVPR] Groupwise Query Specialization and Quality-Aware Multi-Assignment for Transformer-based Visual Relationship Detection
12. [CVPR] Prompt Learning via Meta-Regularization
13. [CVPR] Retrieval-Augmented Open-Vocabulary Object Detection
14. [ICLR] Domain-agnostic Latent Diffusion Models for Synthesizing High-Quality Implicit Neural Representations

Foundation Models (기초모델)

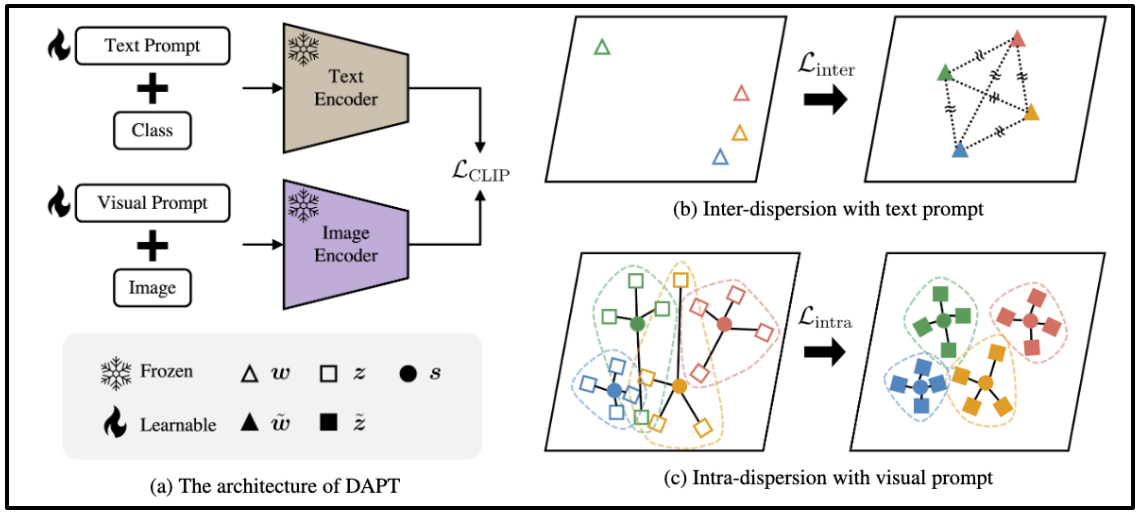




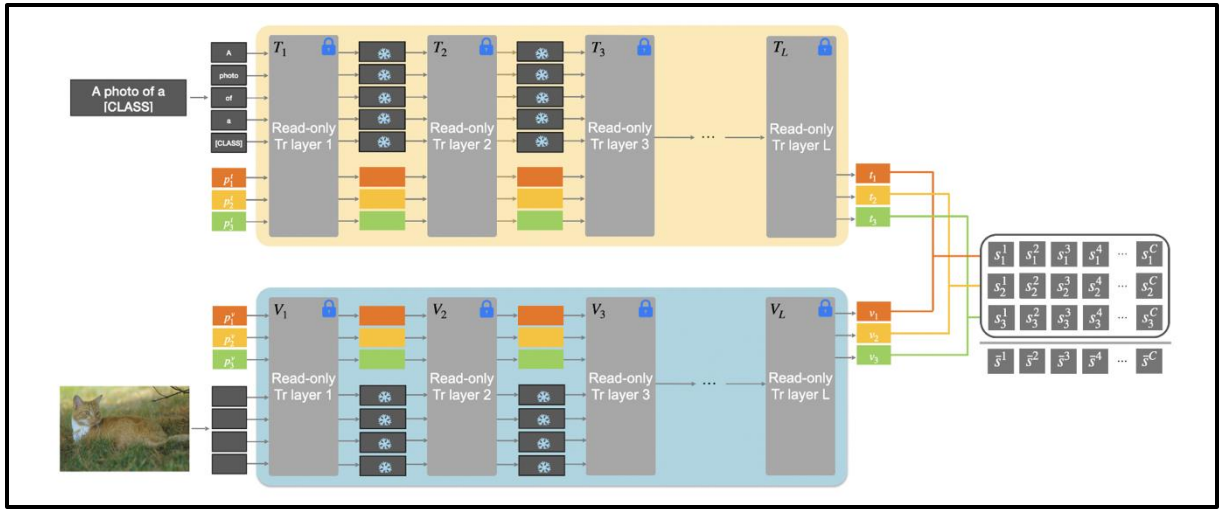
[CVPR22] VT-TWIN. (1억개 유투브 클립학습. 비디오 기초모델)



[ICCV23] 거대 동영상 기초모델 기반 오픈-사전 비디오 QA 시스템



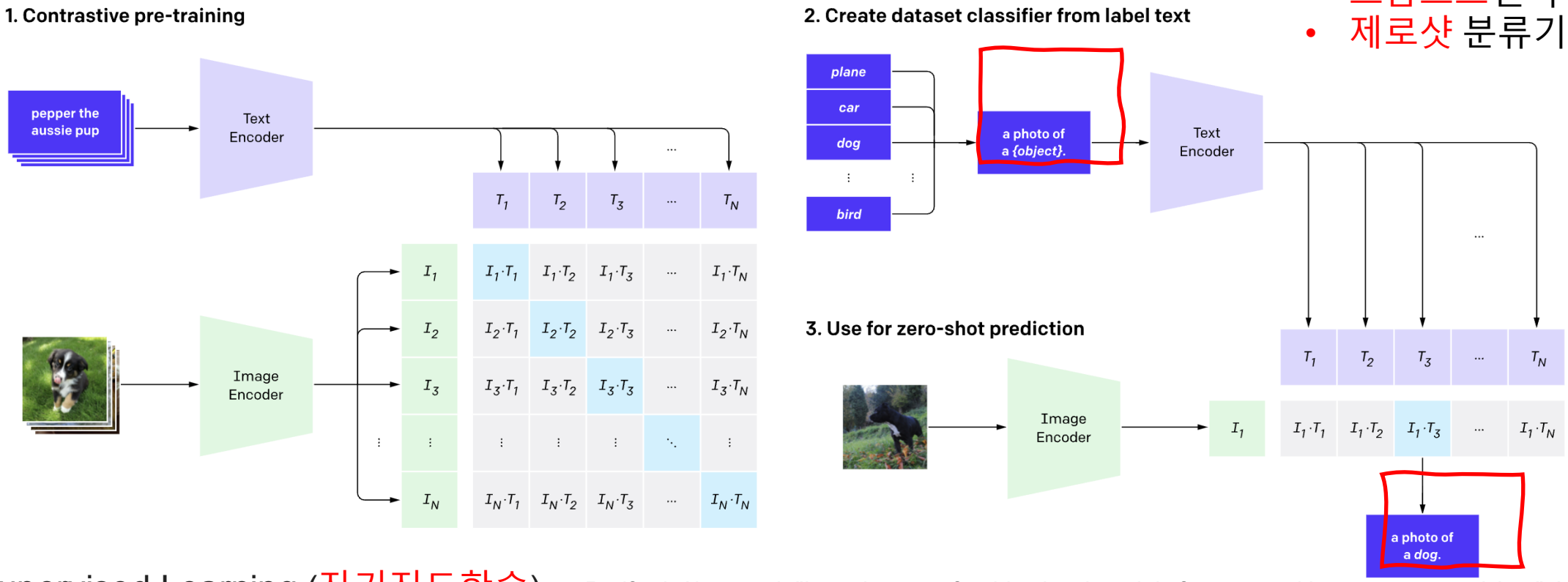
[ICCV23] DAPT. CLIP 기반 few-shot learning



[ICCV23] RPT. 기초모델 위한 읽기 중심 프롬프트 튜닝

Contrastive Learning (CLIP)

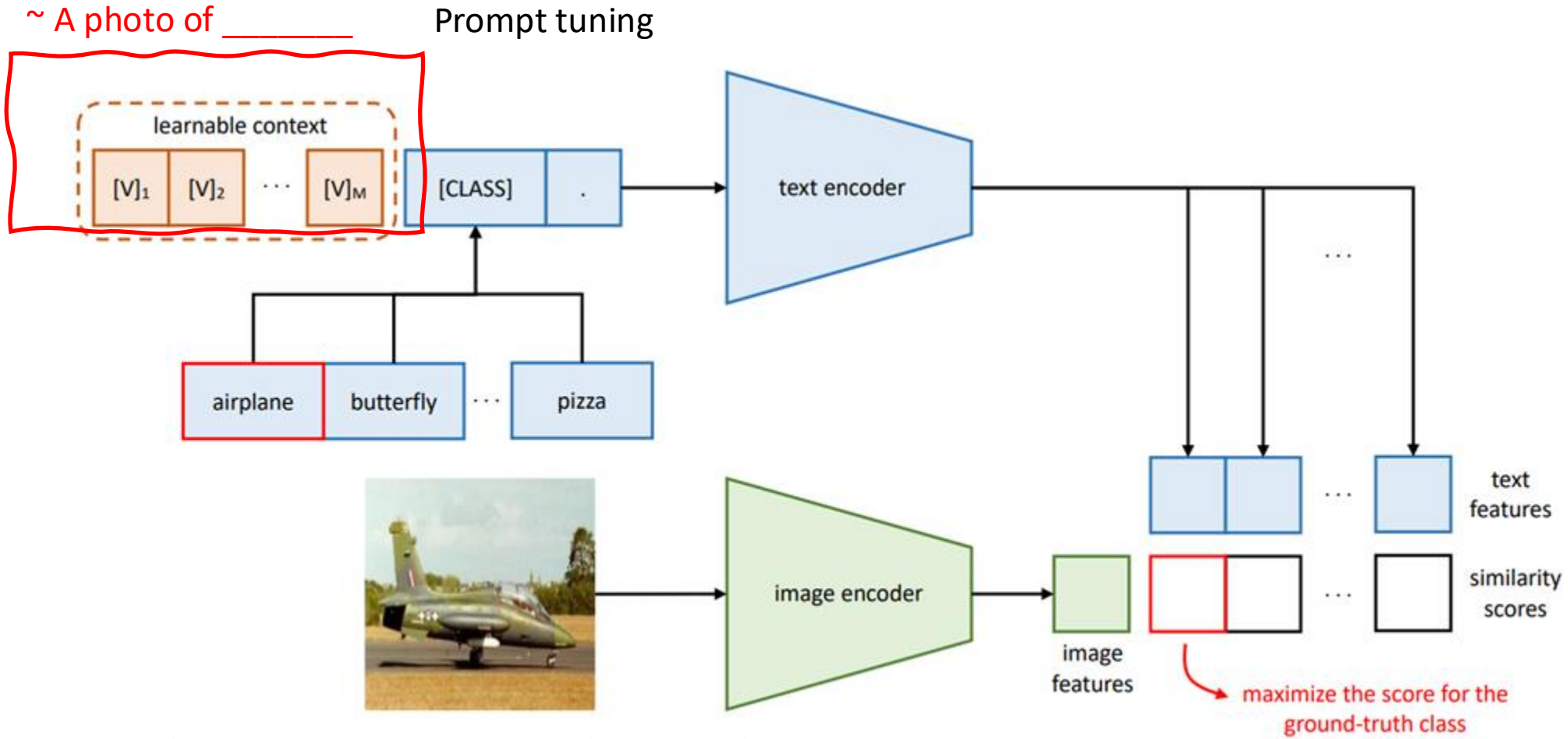
A photo of _____
- 프롬프트엔지니어링
- 제로샷 분류기



Self-Supervised Learning (자기지도학습)

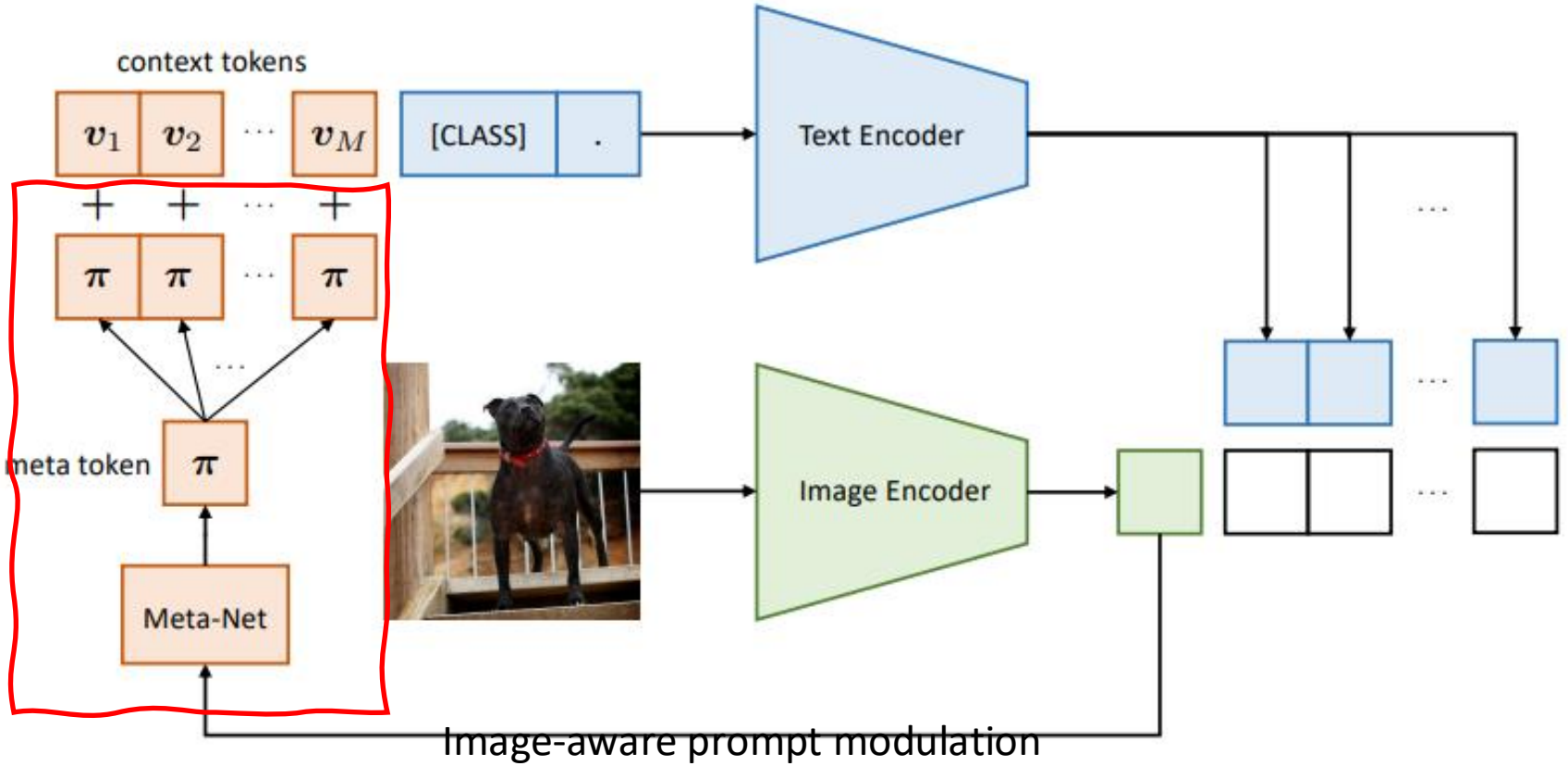
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *ICML* 2021.

Prompt Learning (CoOp)



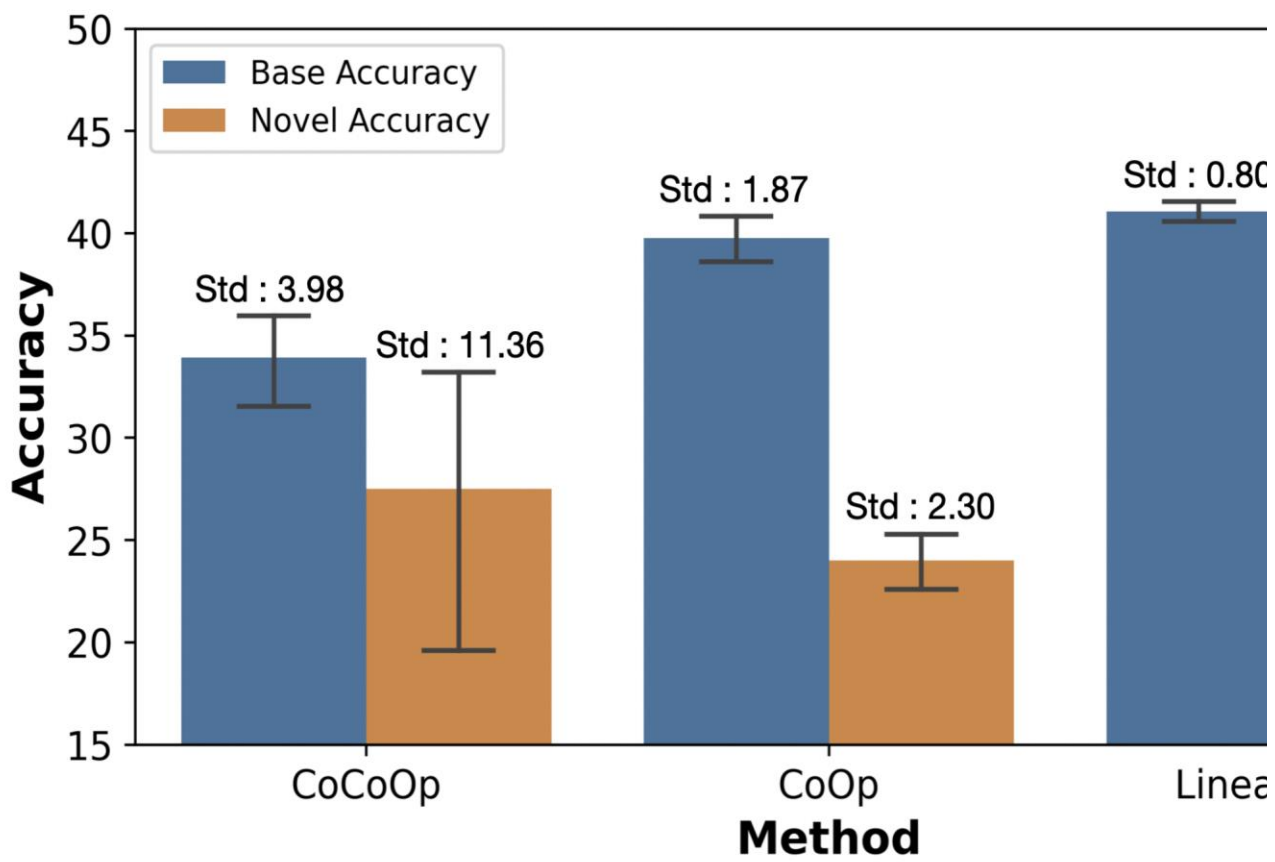
Learning to prompt for vision-language models (IJCV 2022)

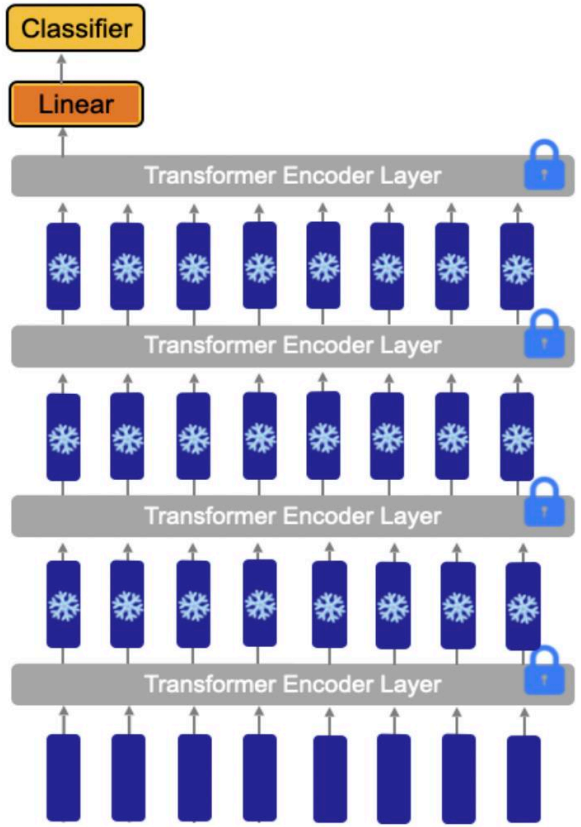
CoCoOp



Conditional prompt learning for vision-language models (CVPR 2022)

Our Observation

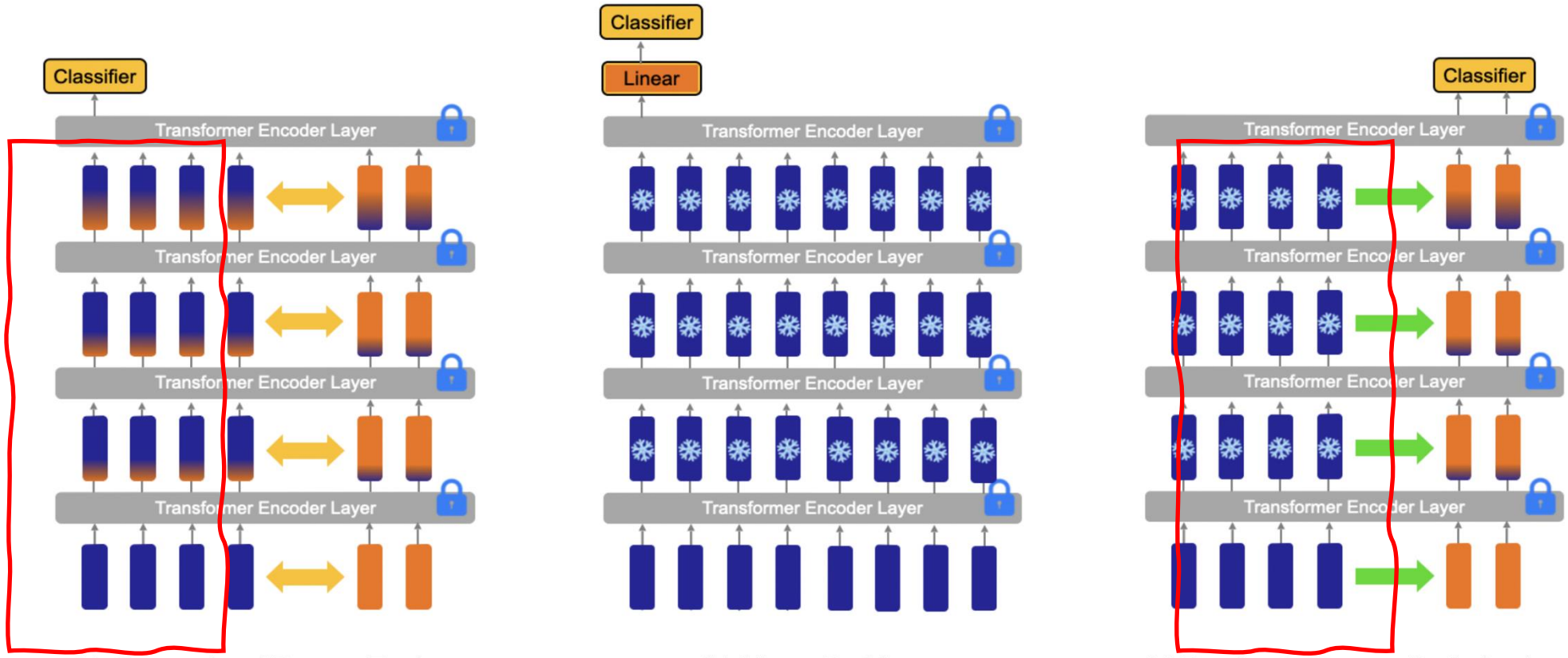




(b) Linear Probing

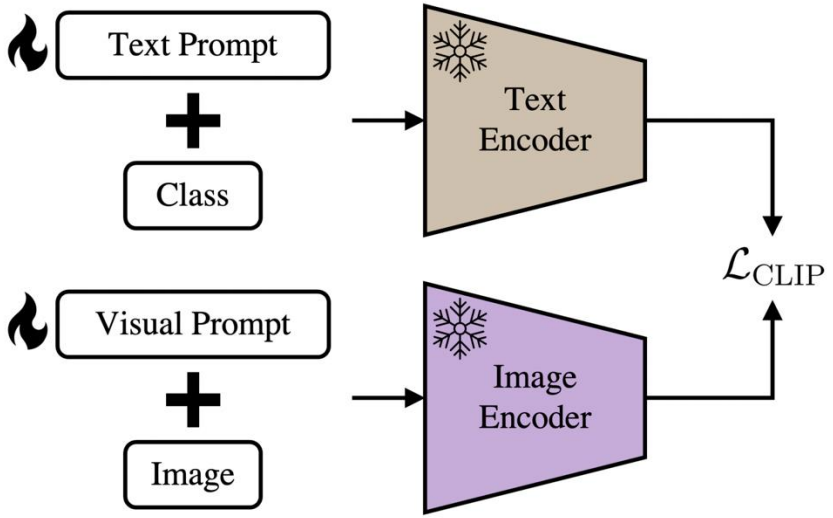
Linear Probe achieved better performance with small VAR (std).

Read-only Prompt Optimization

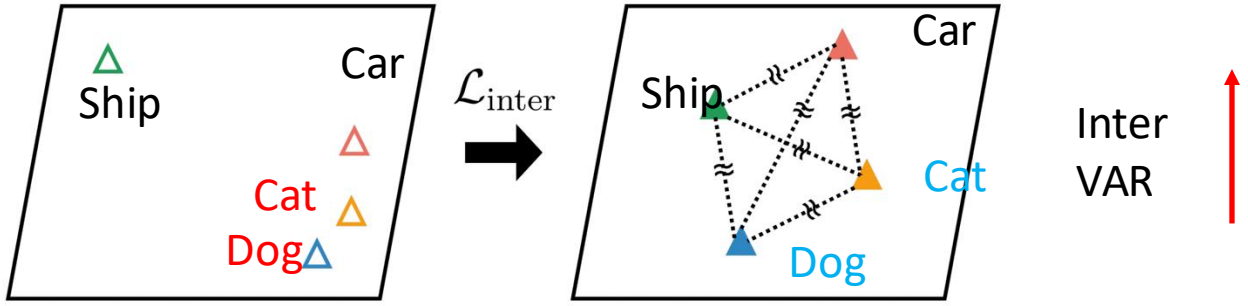


(a) Conventional Prompt Tuning (b) Linear Probing (c) Read-only Prompt Optimization

Distribution-Aware Prompt Tuning

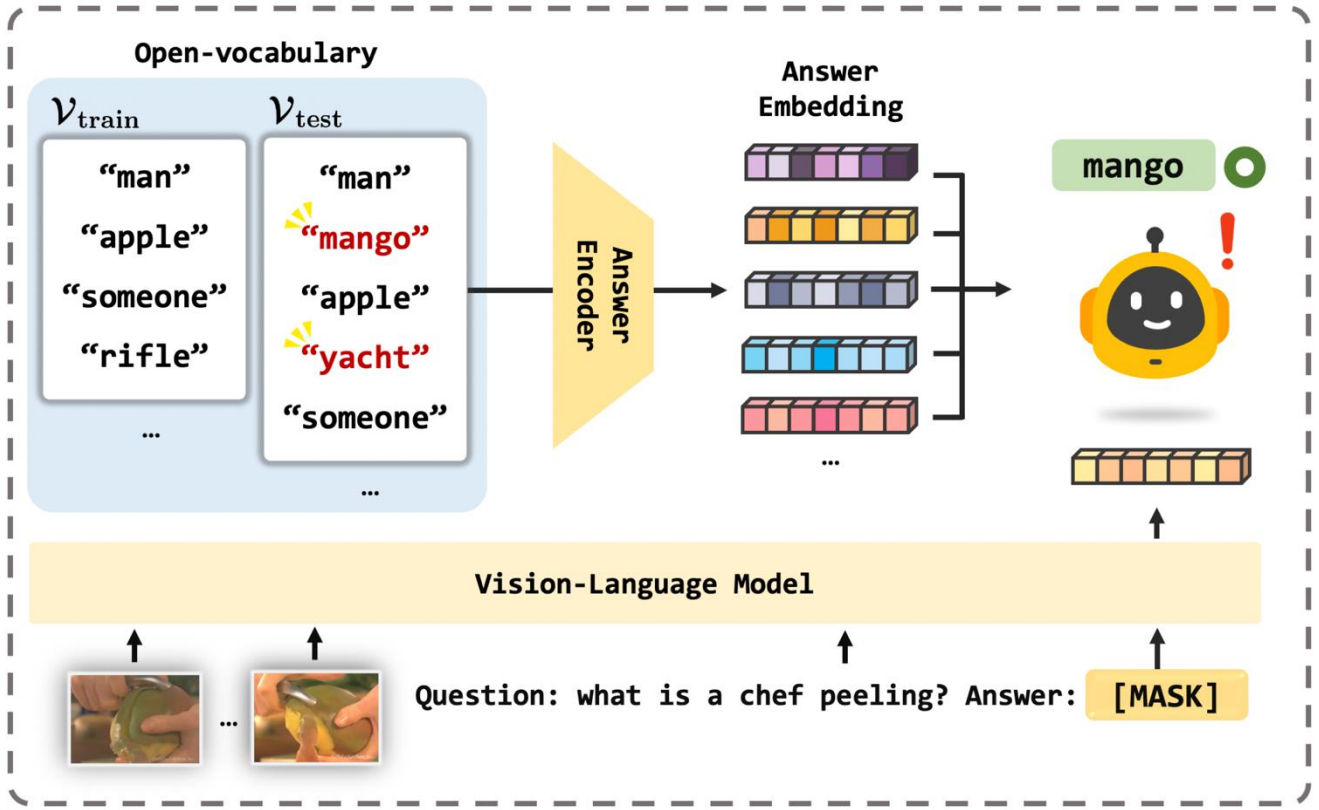


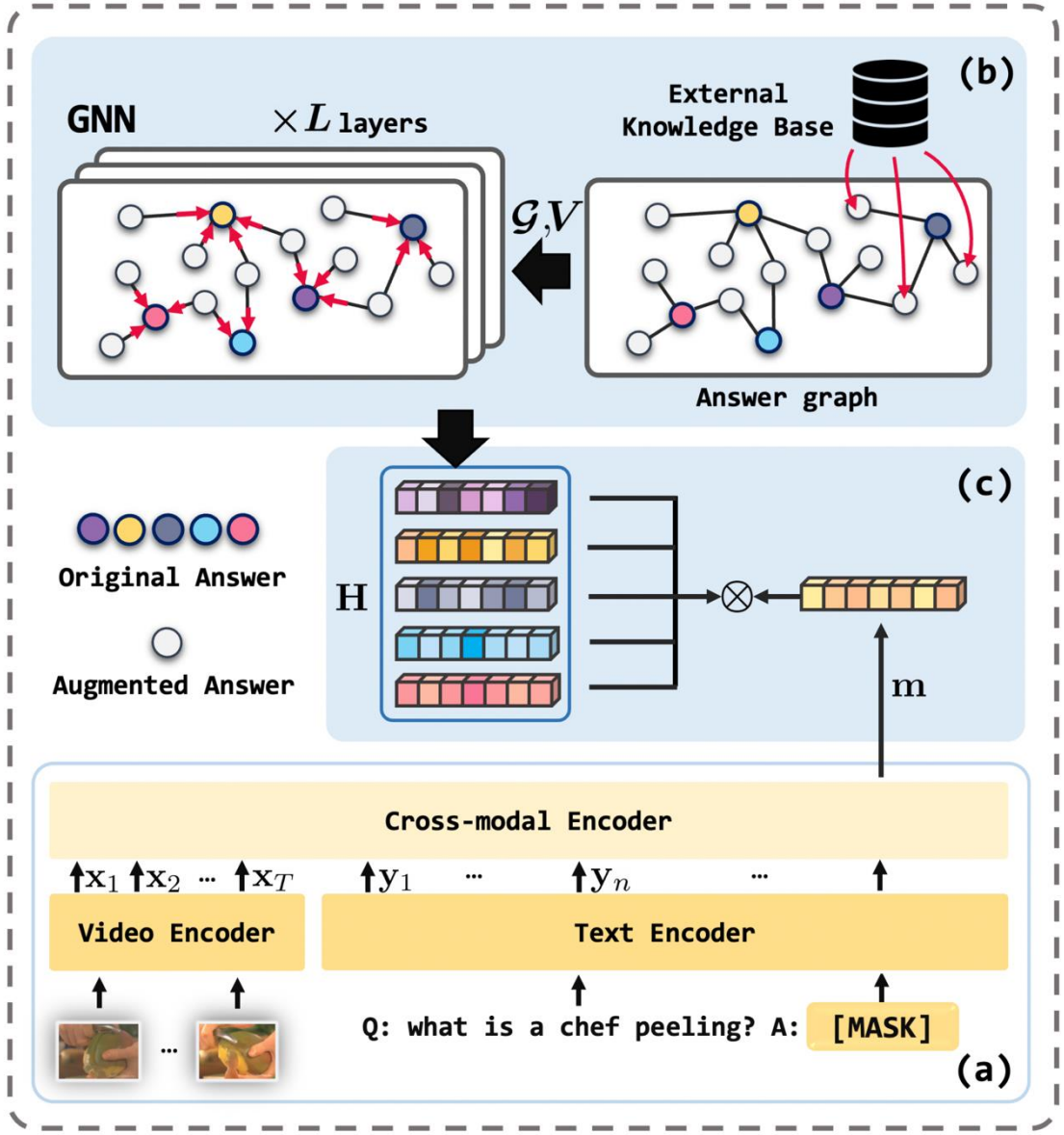
(a) The architecture of DAPT



(b) Inter-dispersion with text prompt

Video Question Answering



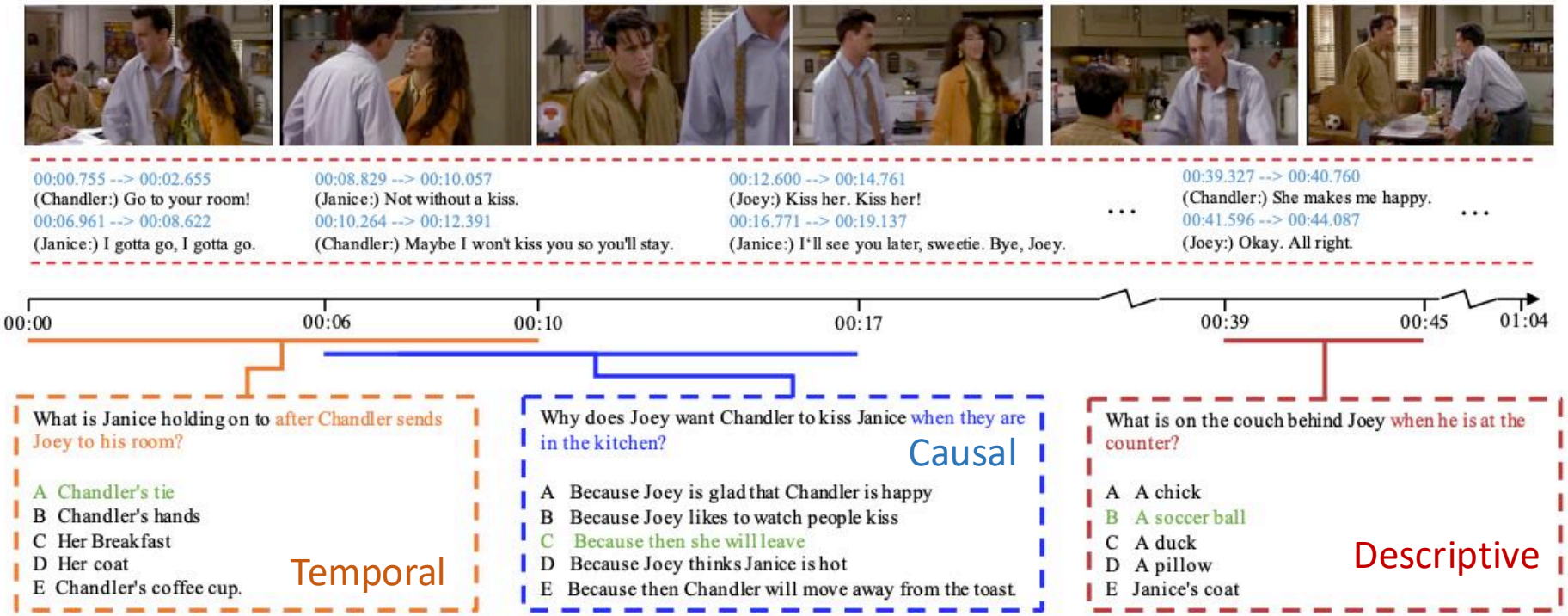


- New Benchmark
- Label smoothing by GNNs
- Vision-Language Model
- Cross-Modal Encoder thanks to Transformers

Video QA

Downstream tasks

- Video question answering



- Dataset:
  - TGIF
  - MSVD
- Evaluation metric:
  - Accuracy

Jie Lei et al. TVQA: Localized, Compositional Video Question Answering. EMNLP, 2018.

Questions?