Poster Session: TUE-AM-256

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks

Hao Li*, Jinguo Zhu*, Xiaohu Jiang*, Xizhou Zhu+,

Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, Jifeng Dai

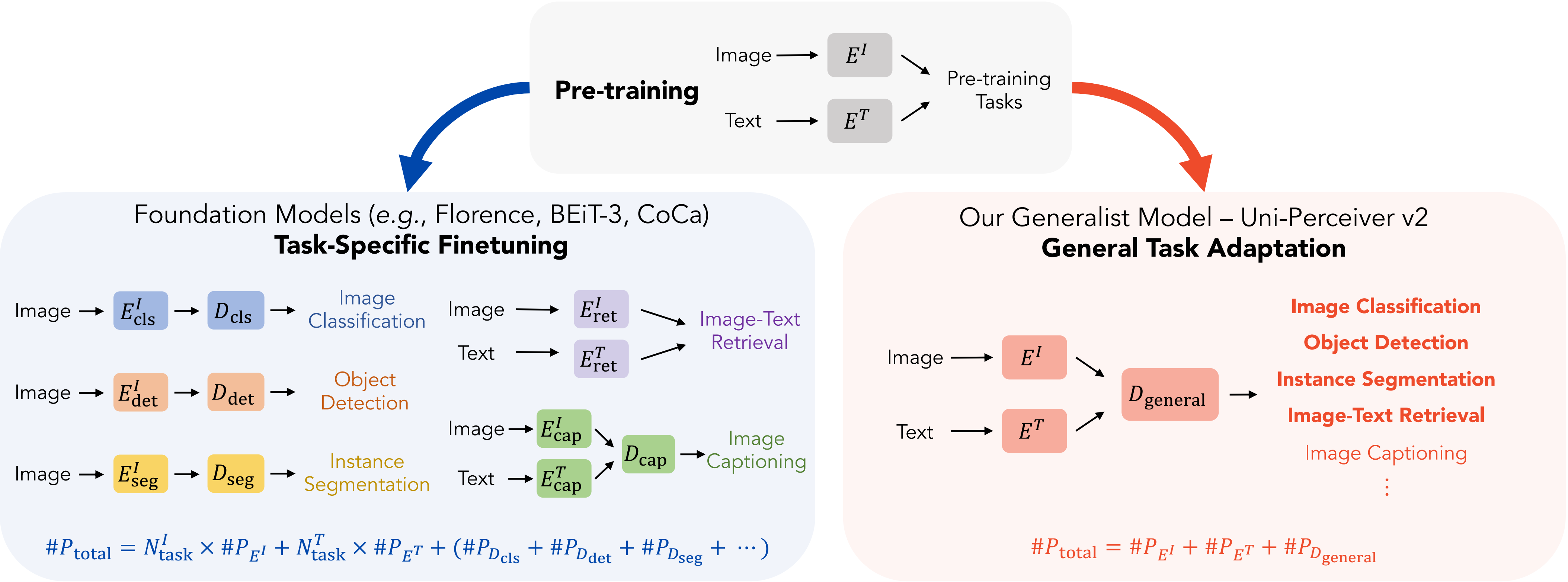
* Co-first Authors + Corresponding Author

CVPR 2023 Highlight Paper

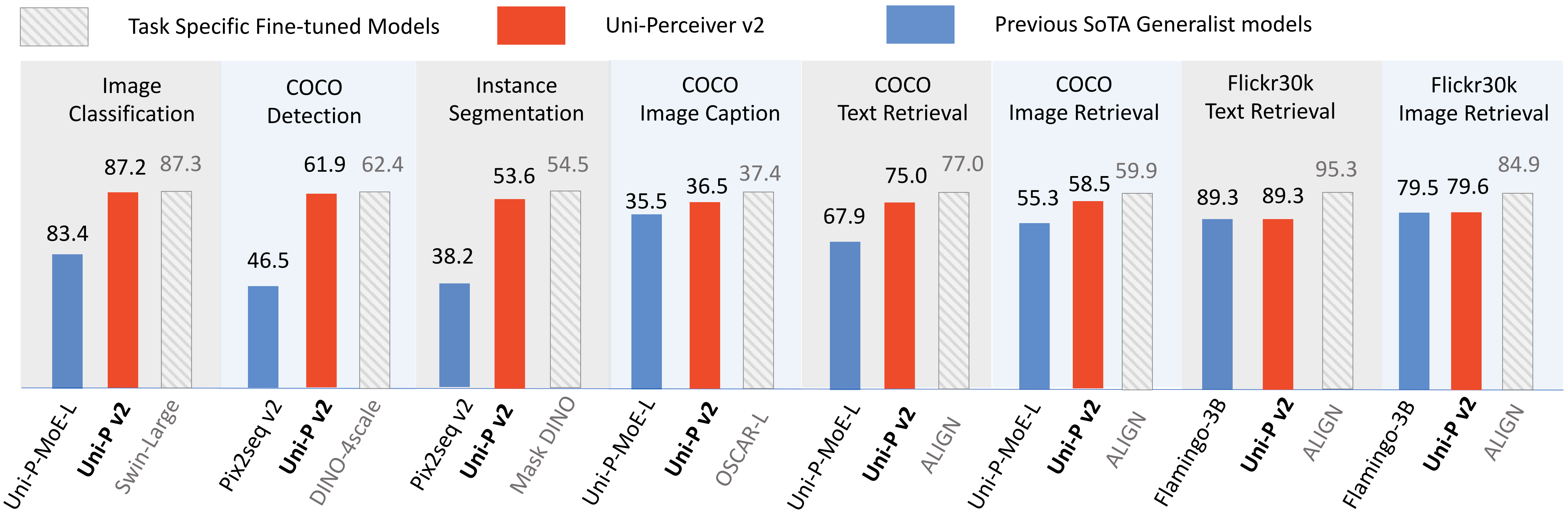
商汤 sensetime

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

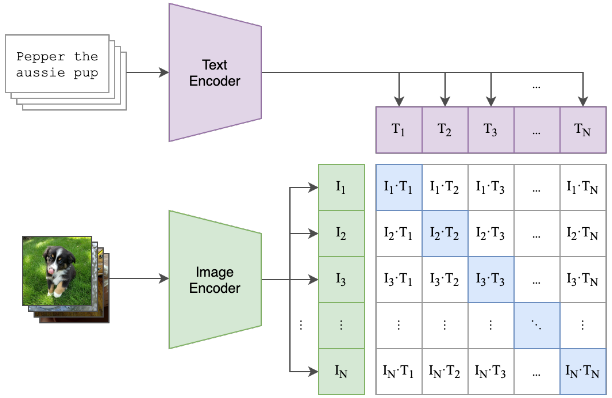
- **Uni-Perceiver v2**： A generalist model for large-scale vision and vision-language tasks
  - Handles a broad range of vision / vision-language tasks **without finetuning**
  - **Outperforms all existing generalist models** in both versatility and performance
  - Achieves competitive performance compared with **commonly-recognized task-specific strong baselines**



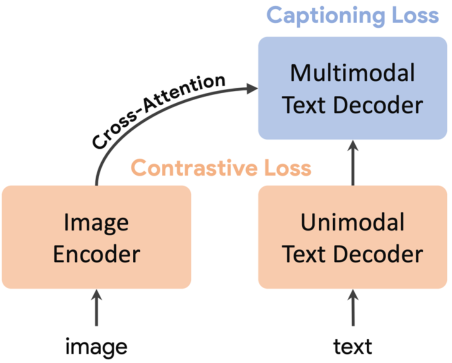
- **Uni-Perceiver v2**： A generalist model for large-scale vision and vision-language tasks
  - Handles a broad range of vision / vision-language tasks **without finetuning**
  - **Outperforms all existing generalist models** in both versatility and performance
  - Achieves competitive performance compared with **commonly-recognized task-specific strong baselines**



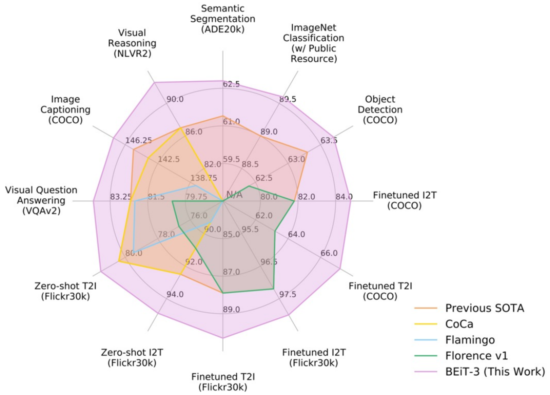
- **Foundation models** pretrained on large-scale image-text pairs show strong performance on a series of downstream tasks



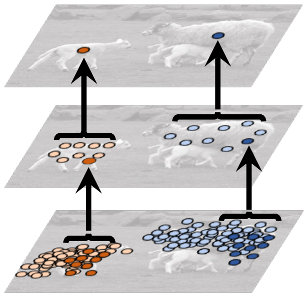
CLIP



CoCa

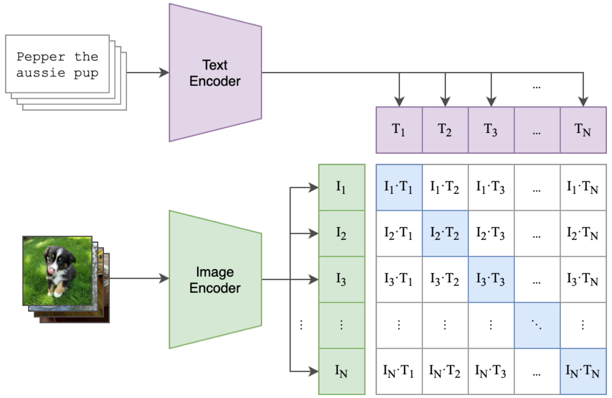


BEiT-3

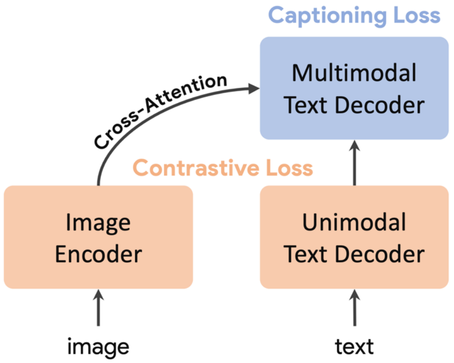


InternImage

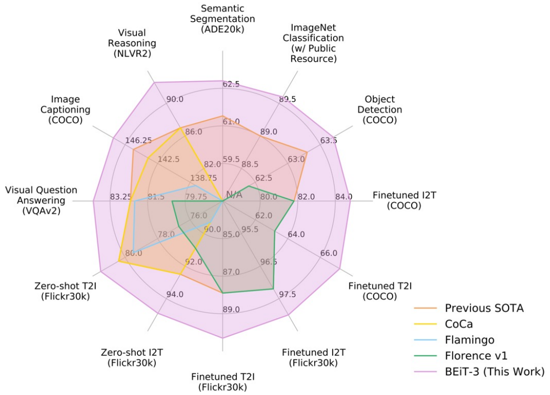
- **Foundation models** pretrained on large-scale image-text pairs show strong performance on a series of downstream tasks
- **Foundation models are not general enough** – they need finetuning



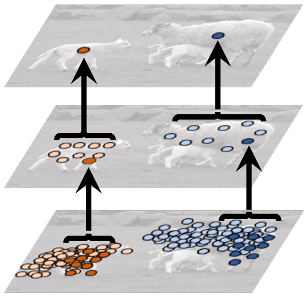
CLIP



CoCa

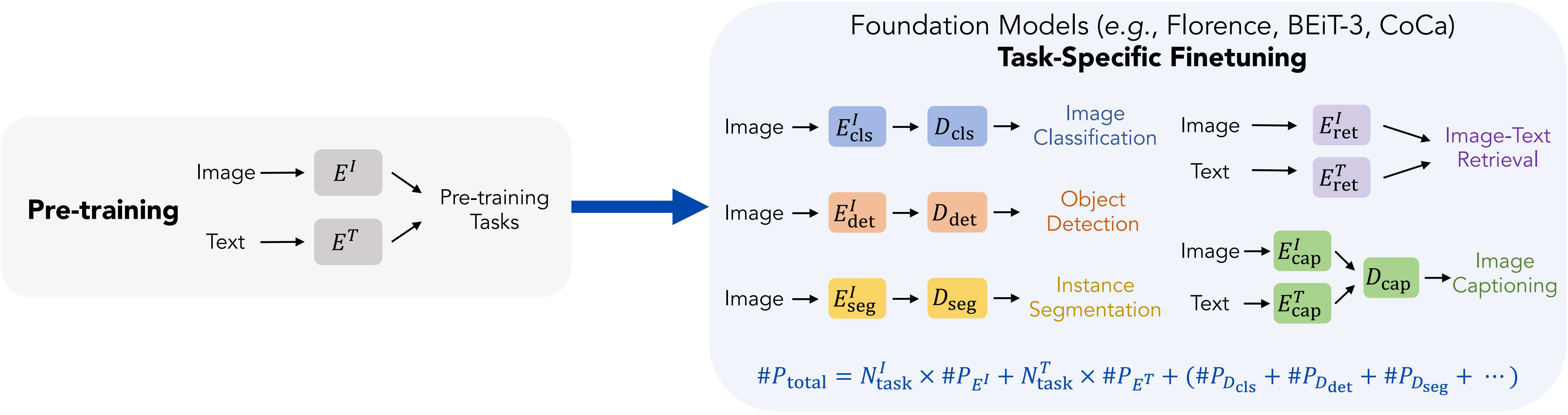


BEiT-3



InternImage

- **Foundation models** pretrained on large-scale image-text pairs show strong performance on a series of downstream tasks
- **Foundation models are not general enough** – they need finetuning
  - Enough data needs to be collected and labeled for training on each downstream task
  - Task modules (*e.g.*, detection heads) need to be designed and trained
  - Thousands of models for thousands of tasks / real-world scenarios



- **Foundation models** pretrained on large-scale image-text pairs show strong performance on a series of downstream tasks
- **Foundation models are not general enough** – they need finetuning
  - Enough data needs to be collected and labeled for training on each downstream task
  - Task modules (*e.g.*, detection heads) need to be designed and trained
  - Thousands of models for thousands of tasks / real-world scenarios
- How to design **a generalist model** capable of handling different tasks **without finetuning**?

- How to design **a generalist model** capable of handling different tasks **without finetuning**?
- **Difficulties:**
  - Different tasks have **different representations and output forms**
  - Different tasks may **conflict with each other** with shared parameters
  - Multi-task joint training requires **trade-off between tasks, which is tricky**

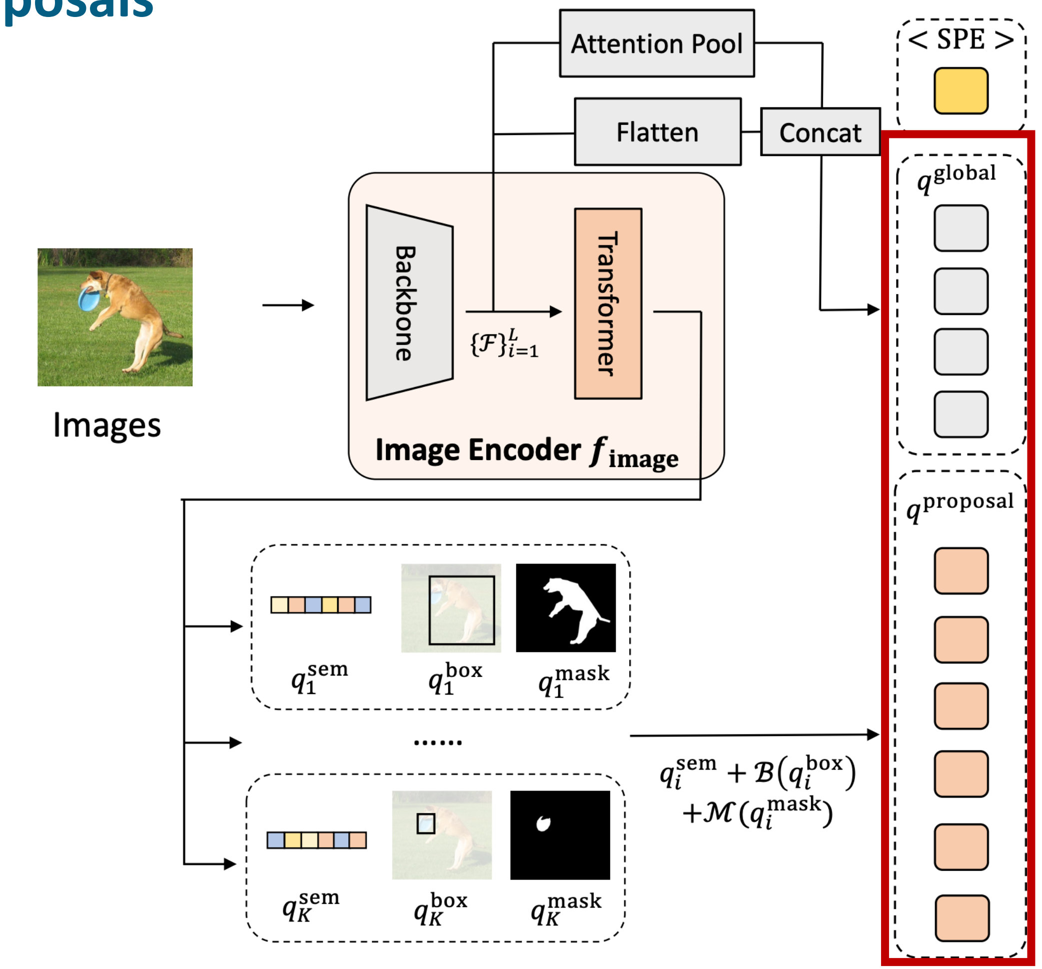
- **Difficulty #1:** Different tasks have **different representations and output forms**
- Representation: Encoding images as **general region proposals**

$$f_{\text{image}}(x) = \text{Concat}\left(\{q_i^{\text{global}}\}_{i=1}^{M}\,,\;\{q_j^{\text{proposal}}\}_{j=1}^{N}\right)$$

where

$$q_j^{\text{proposal}} = q_j^{\text{sem}} + \mathcal{B}(q_j^{\text{box}}) + \mathcal{M}(q_j^{\text{mask}})$$

$$q^{\text{global}} = \text{Concat}\Big(\{\,\text{AttnPool}_i(\mathcal{F}_L)\}_{i=1}^{M'}\,,\;\text{Flatten}(\mathcal{F}_L)\Big)$$



- **Difficulty #1:** Different tasks have **different representations and output forms**
- Representation: Encoding images as **general region proposals**
- Output: Employing the **unified task formulation** of Uni-Perceiver

In Uni-Perceiver, different tasks are identified as **different input set X and candidate output set Y**. Given $x \in X$, the task is defined as **finding $y \in Y$ with the maximum likelihood x**.

The likelihood between input x and target y

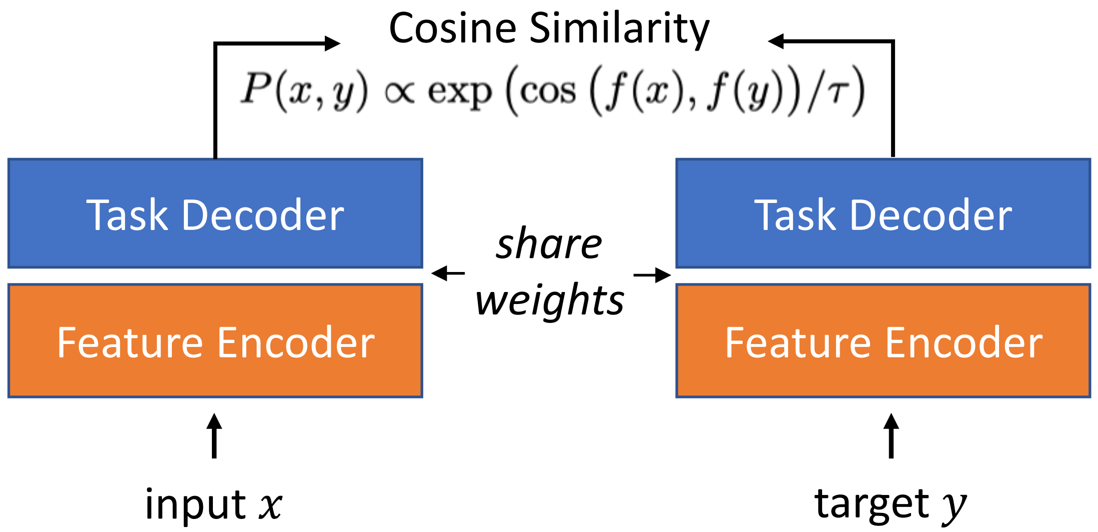
$$P(x, y) \propto \exp\left(\cos\left(f(x), f(y)\right)/\tau\right)$$

Given x, the target $\hat{y}$ with the maximum likelihood

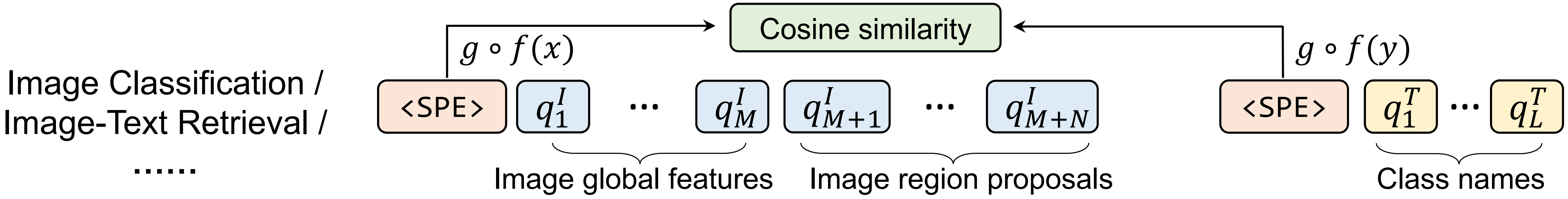
$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(x, y)$$

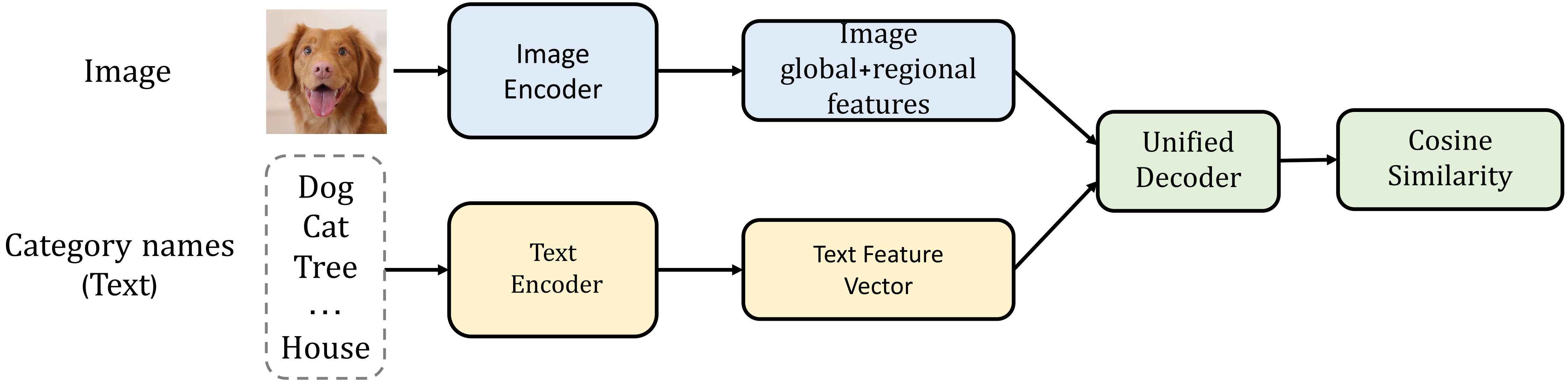
Loss function for multi-task joint training

$$L = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\{x,y\} \in \{\mathcal{X}_i, \mathcal{Y}_i\}} \left[-\log \frac{P(x, y)}{\sum_{z \in \mathcal{Y}_i} P(x, z)} \right]$$

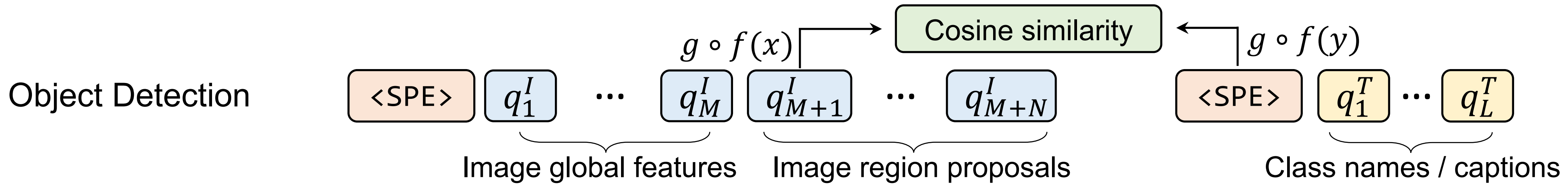


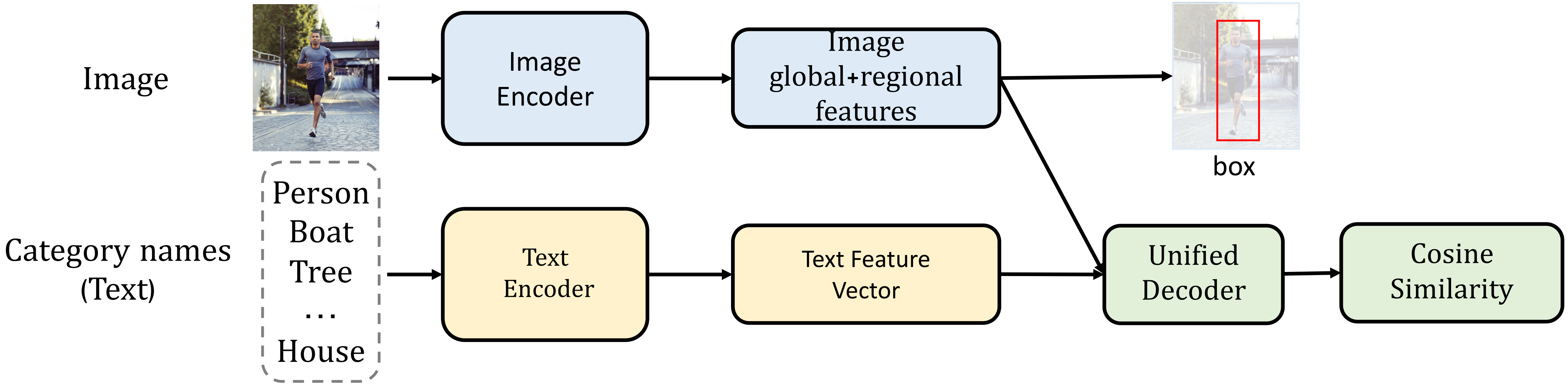
- **Unified Task Formulation of Uni-Perceiver**
  - **Image Classification**



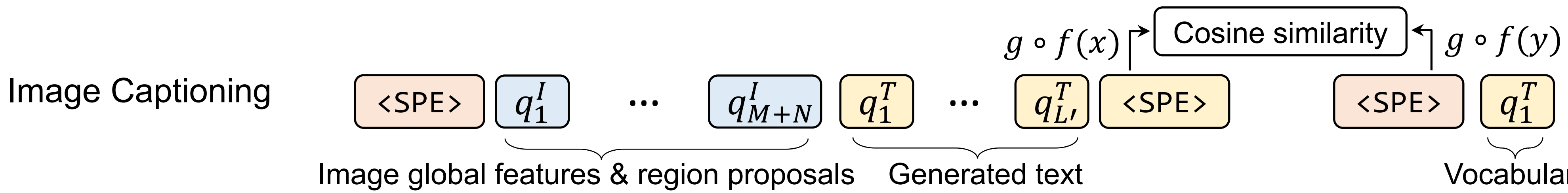


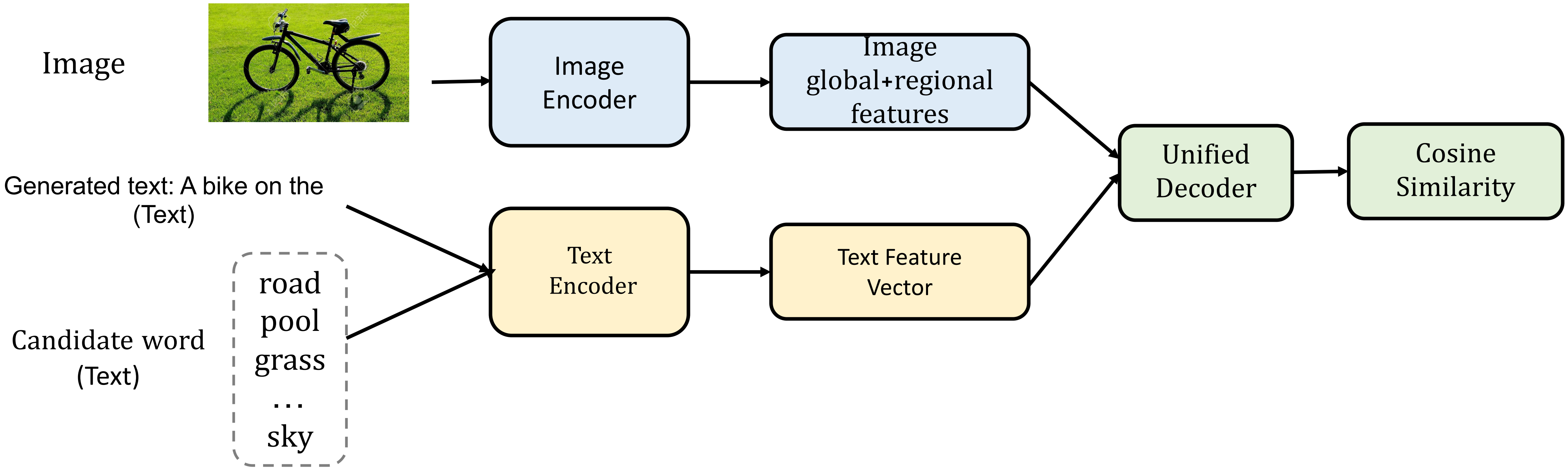
- **Unified Task Formulation of Uni-Perceiver**
  - **Object Detection**





- ## Unified Task Formulation of Uni-Perceiver
  - ### Image Captioning

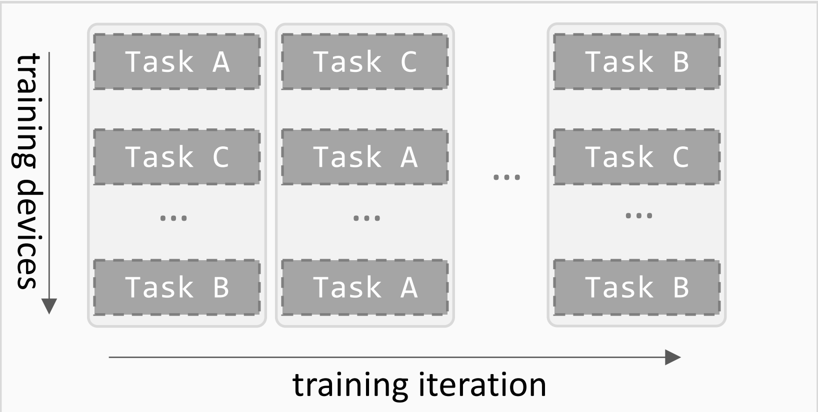




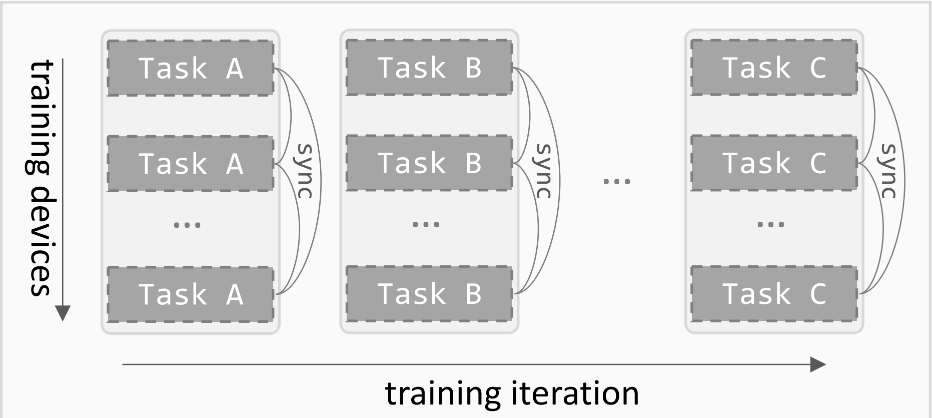
- **Difficulty #2:** Different tasks may **conflict** with shared parameters
- Solution: We employ the **Conditional MoE** proposed in Uni-Perceiver-MoE

Tasks	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
Single Task	50.1	76.1	50.0	37.6	30.2
All Tasks	49.8	76.3	46.0	34.7	28.9
w/o Detection	-	76.6 (+0.3)	47.0 (+1.0)	34.6 (−0.1)	30.4 (+0.5)
w/o Classification	50.1 (+0.3)	-	51.6 (+5.6)	38.6 (+3.9)	25.9 (−3.0)
w/o Retrieval	49.5 (−0.3)	76.3 (+0.0)	-	-	27.4 (−1.5)
w/o Captioning	49.7 (−0.1)	76.3 (+0.0)	51.2 (+5.2)	38.3 (+3.6)	-
All Tasks w/ MoE	49.9 (+0.1)	76.9 (+0.6)	51.3 (+5.3)	38.8 (+4.1)	30.6 (+0.7)

- **Difficulty #3:** Multi-task joint training requires **trade-off between tasks, which is tricky**
- Solution: We propose improved optimization strategy for multi-task training
  - **Unmixed sampling strategy**：All GPUs share the same task in one iteration
    - Increases batch-size, which improves efficiency and performance
    - Reduces the synchronization cost caused by the different iteration time of different tasks
    - **Difficulty:** the gradients differ significantly between iterations, causing training instability



Mixed sampling



Unmixed sampling

- **Difficulty #3:** Multi-task joint training requires **trade-off between tasks, which is tricky**
- Solution: We propose improved optimization strategy for multi-task training
  - **Unmixed sampling strategy**：All GPUs share the same task in one iteration
  - **Task-Balanced Gradient Normalization:** Adaptively normalize the gradients of each task to stabilize the training with unmixed sampling strategy

$$\begin{cases} \mathbf{g}_t \leftarrow \nabla L_{t,k}\left(\theta_{t-1}\right) \\ \mathbf{m}_t = (1-\beta_1)\mathbf{m}_{t-1} + \beta_1 \mathbf{g}_t \\ \mathbf{n}_t = (1-\beta_2)\mathbf{n}_{t-1} + \beta_2 \mathbf{g}_t^2 \\ \theta_t = \theta_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{n}_t} + \varepsilon} \end{cases} \Rightarrow \begin{cases} \mathbf{g}_t \leftarrow \omega_k \frac{\nabla L_{t,k}\left(\theta_{t-1}\right)}{\left\|\nabla L_{t,k}\left(\theta_{t-1}\right)\right\|} \\ \mathbf{m}_t = (1-\beta_1)\mathbf{m}_{t-1} + \frac{\beta_1}{s_k} \mathbf{g}_t \\ \mathbf{n}_t = (1-\beta_2)\mathbf{n}_{t-1} + \frac{\beta_2}{s_k} \mathbf{g}_t^2 \\ \theta_t = \theta_{t-1} - \alpha \frac{\mathbf{m}_t}{\sqrt{\mathbf{n}_t} + \varepsilon} \end{cases}$$

Task-Balanced Gradient Normalization

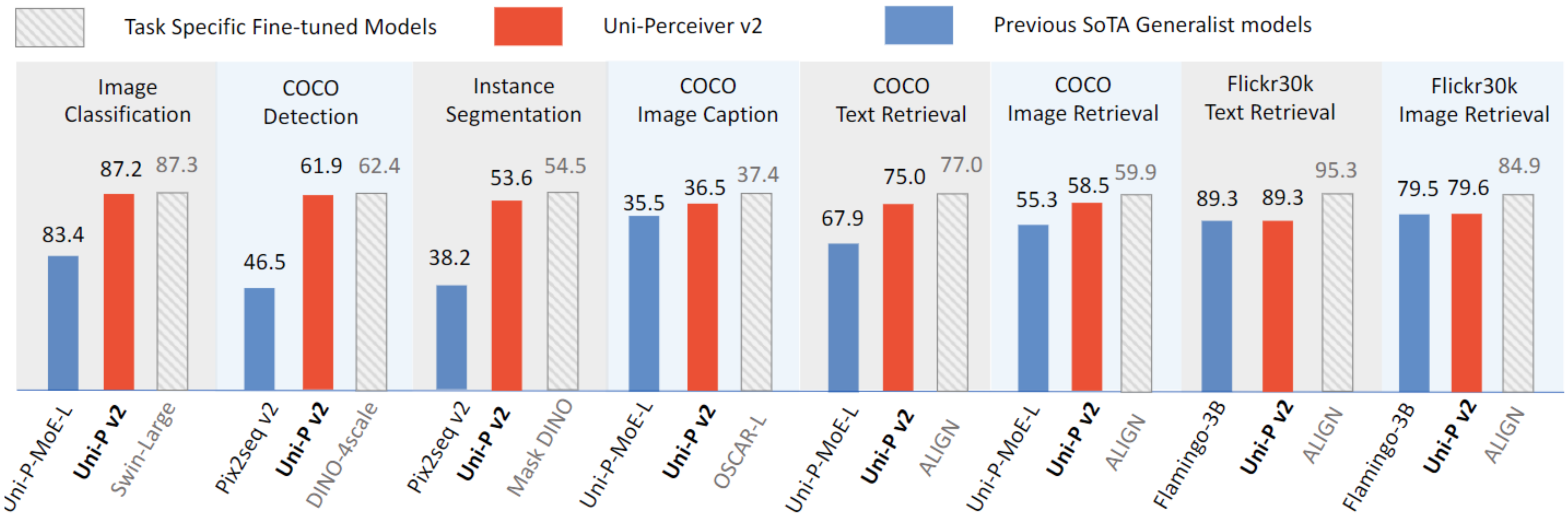
Task Sampling	Gather Feature	TBGN	COCO Detection	ImageNet-1k Classification	COCO Retrieval		COCO Caption
mixed			49.6	76.7	40.1	31.9	27.6
unmixed			49.2	76.6	39.8	30.9	27.5
unmixed	✓		49.3	76.8	50.4	37.3	27.6
unmixed	✓	✓	**49.9**	**76.9**	**51.3**	**38.8**	**30.6**

- **Experiments**

Methods	#params	Image Classification	Object Detection	Instance Segmentation	Image Captioning		Text Retrieval		Image Retrieval	
		ImageNet-1k	COCO	COCO	COCO		COCO	Flickr30k	COCO	Flickr30k
		Acc	mAP	mAP	B@4	CIDEr	R@1	R@1	R@1	R@1
Pix2Seq v2 [5]	132M	-	46.5	38.2	34.9	-	-	-	-	-
UniTab [43]	185M	-	-	-	-	115.8	-	-	-	-
Unified-IO $_{\text{LARGE}}$ [23]	776M	71.8	-	-	-	-	-	-	-	-
Unified-IO $_{\text{XL}}$ [23]	2.9B	79.1	-	-		122.3	-	-	-	-
Flamingo-3B [1]	3.2B	-	-	-	-	-	65.9	89.3	48.0	79.5
Uni-Perceiver $_{\text{BASE}}$ [50]	124M	79.2	-	-	32.0	-	64.9	82.3	50.7	71.1
Uni-Perceiver $_{\text{LARGE}}$ [50]	354M	82.7	-	-	35.3	-	67.8	83.7	54.1	74.2
Uni-Perceiver-MoE $_{\text{BASE}}$ [49]	167M	80.3	-	-	33.2	-	64.6	82.1	51.6	72.4
Uni-Perceiver-MoE $_{\text{LARGE}}$ [49]	505M	83.4	-	-	35.5	-	67.9	83.6	55.3	75.9
Uni-Perceiver-v2 $_{\text{BASE}}$	308M	86.3	58.6	50.6	35.4	116.9	71.8	88.1	55.6	73.8
Uni-Perceiver-v2 $_{\text{LARGE}}$	446M	**87.2** (+3.8)	**61.9** (+15.4)	**53.6** (+15.4)	**36.5** (+1.6)	**122.5** (+0.2)	**75.0** (+7.1)	**89.3** (+0.0)	**58.5** (+3.2)	**79.6** (+0.1)

- Uni-Perceiver v2 **outperforms all existing generalist models**.
- Uni-Perceiver v2 supports core vision tasks (*e.g.*, object detection / instance segmentation) that **existing generalist models do not support**.

- **Experiments**



- Uni-Perceiver v2 achieves competitive performance compared with **commonly-recognized task-specific strong baselines that require fine-tuning**.

- **Uni-Perceiver series**

  - ❖ Uni-Perceiver (CVPR 2022)
    - Proposes the **unified task formulation** and handles a broad range of tasks with **a single model and shared weights**

  - ❖ Uni-Perceiver-MoE (NeurIPS 2022)
    - Proposes conditional MoE that **effectively mitigate the task interference** in multi-task learning

  - ❖ Uni-Perceiver v2 (CVPR 2023)
    - **Outperforms all existing generalist models** in both versatility and performance
    - Achieves competitive performance compared with **commonly-recognized task-specific strong methods**

Code & Models (in progress) : https://github.com/fundamentalvision/Uni-Perceiver