# Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture

Tanmay Gupta[1] Amita Kamath[1] Aniruddha Kembhavi[1] Derek Hoiem[2]
[1]PRIOR @ Allen Institute for AI [2]University of Illinois at Urbana-Champaign

CVPR 2022 Oral
Presented by Yujin Lee

2022.11.07

# N-purpose systems

- Most of the computer vision architectures
- limited to N *predefined* set of task(s) and challenging to adapt to new tasks.
  - Modification on architecture or learning process required.
  - Lack of Generality even though N is larger than 1.

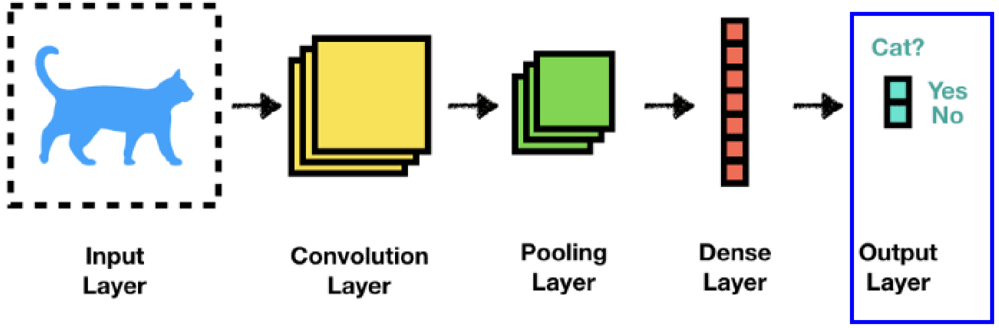
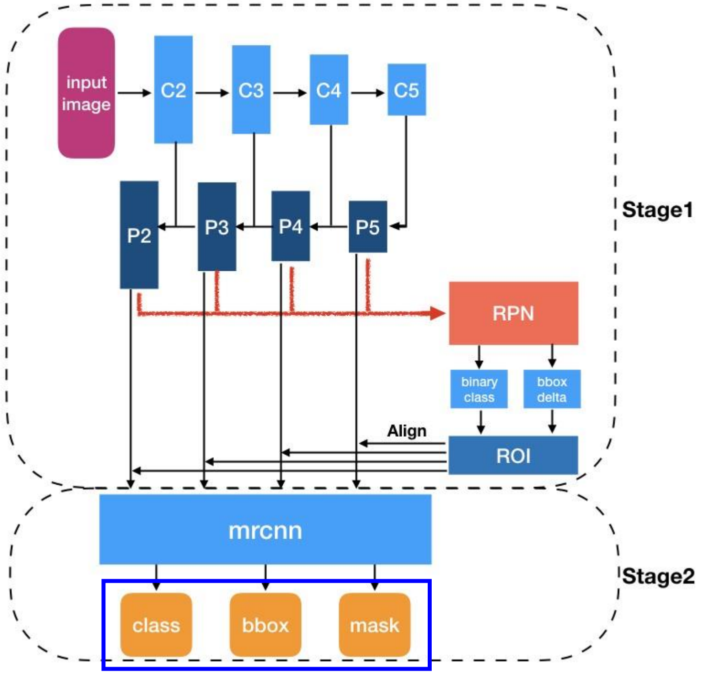


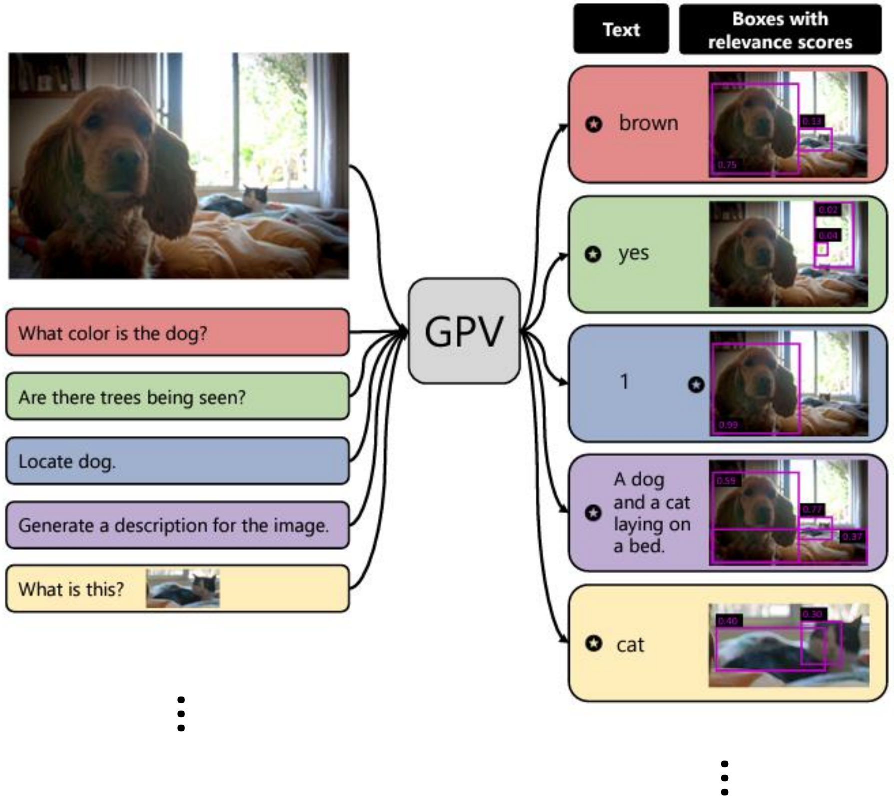

Image Classification
(Single-purpose, N = 1)




Mask R-CNN
:Detection, InstSeg
(multi-purpose, N = 2)

https://towardsdatascience.com/convolutional-neural-network-a-step-by-step-guide-a8b4c88d6943
He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

# General purpose systems



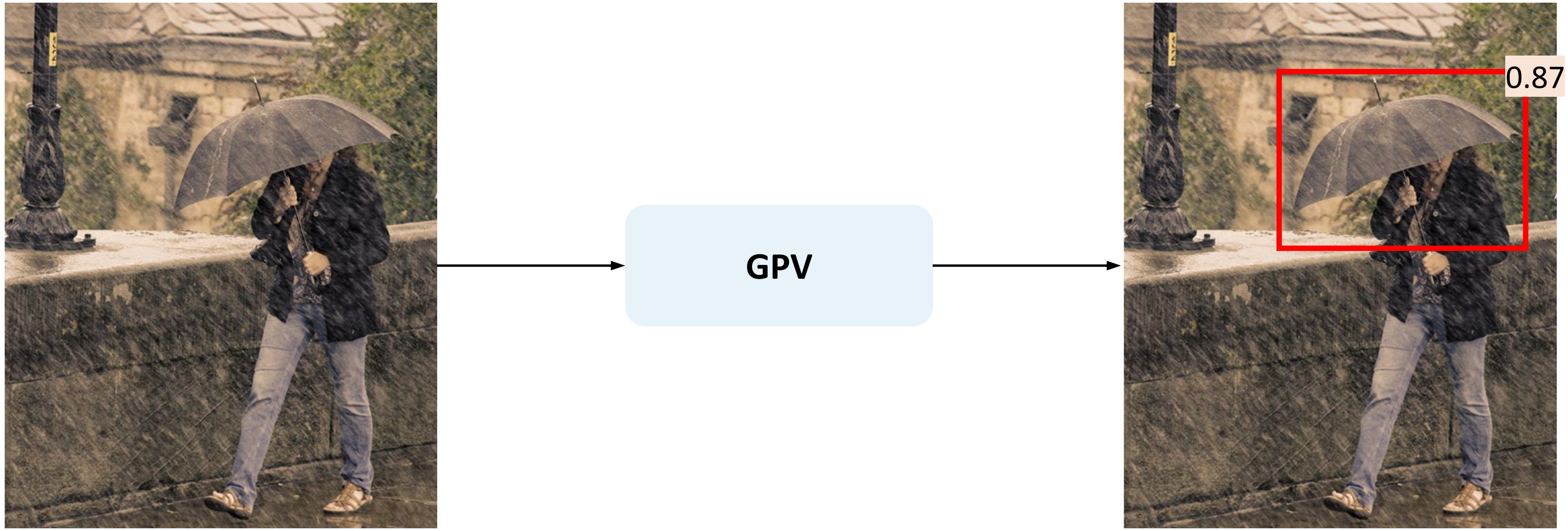

- Designed to carry out many vision tasks.
  → *<u>not limited to predefined tasks</u> at the time of design.*

- Constrained only by its input modalities, memory/instructions, and output modalities.
  → *Highly Flexible*

Gupta, Tanmay, et al. "Towards general purpose vision systems." *arXiv preprint arXiv:2104.00743* (2021).

# GPV-1: Towards General Purpose Vision Systems

- An end-to-end trainable task-agnostic vision-language architecture.
- Task Query: task given in a natural language.
- Each query drawing out a different response using <u>output heads that are shared across tasks.</u>
- **Generality of Architecture, Concepts across Skills, and Learning.**




Locate an umbrella.
(Localization)

1

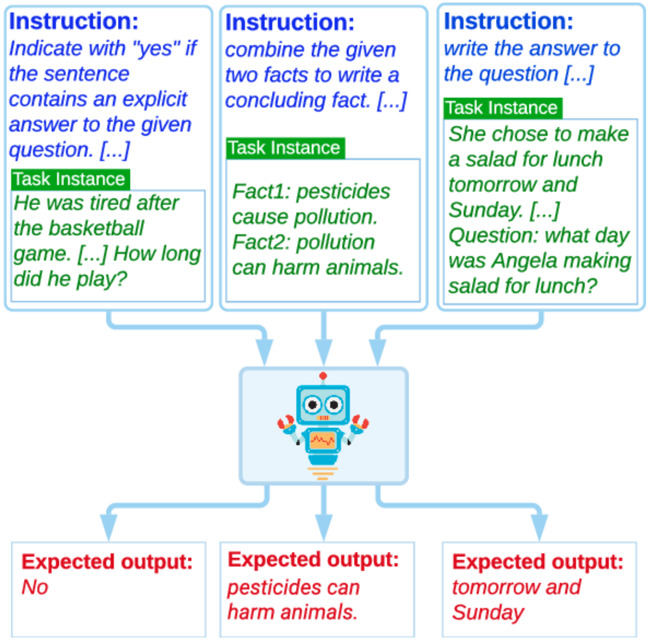
https://cocodataset.org/#explore

# Generality of Architecture

- Learn any task within a broad domain without change to network architecture
- Leveraging **Encoder-Decoder** Architecture
  - Applicable to a wide range of tasks
- Learning from **Task-Description**
  - Task Description → Sequence of text tokens (eventually fed into a text encoder)
  - Enables GPV-1 to be **task-agnostic**

T5

Templated Task Description

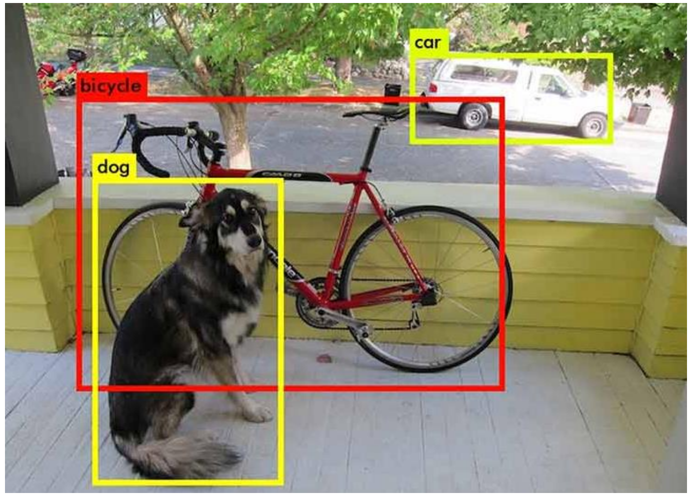



Natural Language Task Description

https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html
Mishra, Swaroop, et al. "Natural instructions: Benchmarking generalization to new tasks from natural language instructions." (2021).

# Terms

- **Concepts**
  - Nouns
  - e.g. car, person, dog, …
- **Skills**
  - Operations that we wish to perform on the given inputs
  - e.g. classification, object detection, image captioning, …
- **Tasks**
  - Predefined combinations of a set of skills performed on a set of concepts
  - e.g. COCO Object Detection task involves the skill of object detection across 80 concepts




✓ Concepts: dog, bicycle, car
✓ Skills: object detection
✓ Tasks: object detection on dog, bicycle, car

https://machinethink.net/blog/object-detection-with-yolo/

# Generality of Concepts Across Skills

- Ability to perform tasks in skill-concepts combinations not seen during training
- If well generalized, should perform well on unseen tasks

Q. “What is this object?”

A. “Dog” ✓

- ➢ Skill: Classification
- ➢ Concepts: Dog

Q. “Is cat sleeping?”

A. “No” ✓

- ➢ Skill: VQA
- ➢ Concepts: Cat

Q. “Is the dog white”

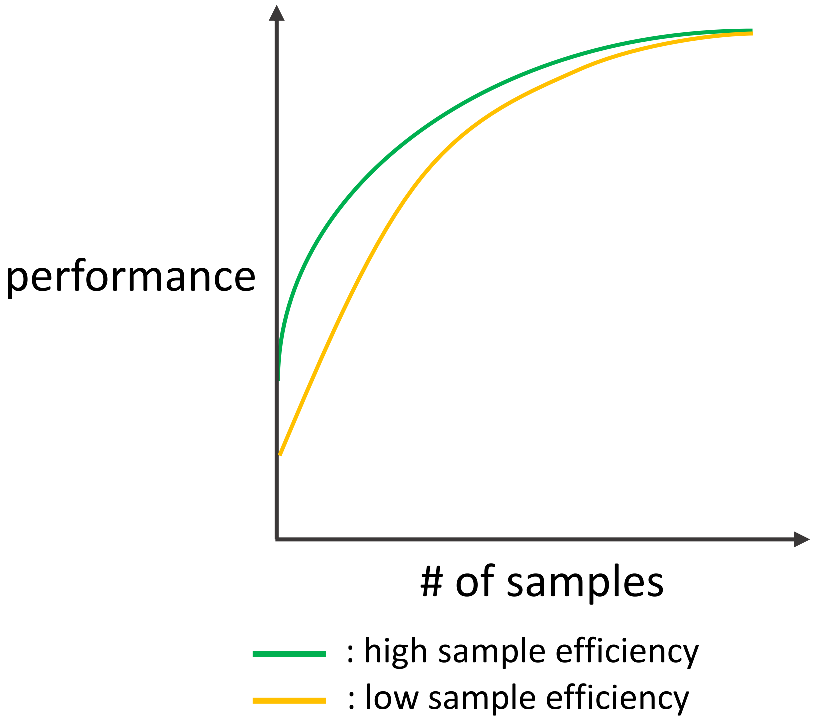
**A. ?**

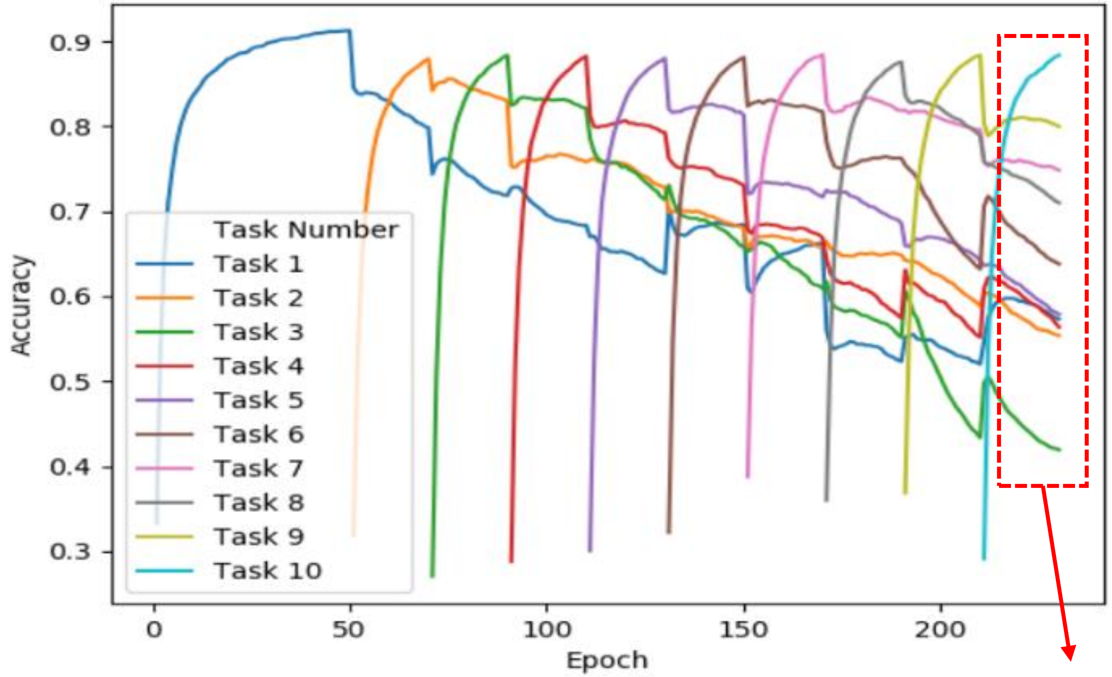
- ➢ Skill: VQA
- ➢ Concepts: Dog

| *Performance* | *Dog* | *Cat* |
|---|---|---|
| *VQA* | **?** | ✓ |
| *Classification* | ✓ | **?** |

# Generality of Learning

- General purpose architecture should be able to ***learn new tasks…***
  - **with sample-efficiency:** to learn with a smaller number of samples
  - **without catastrophic forgetting:** not to forget previous tasks

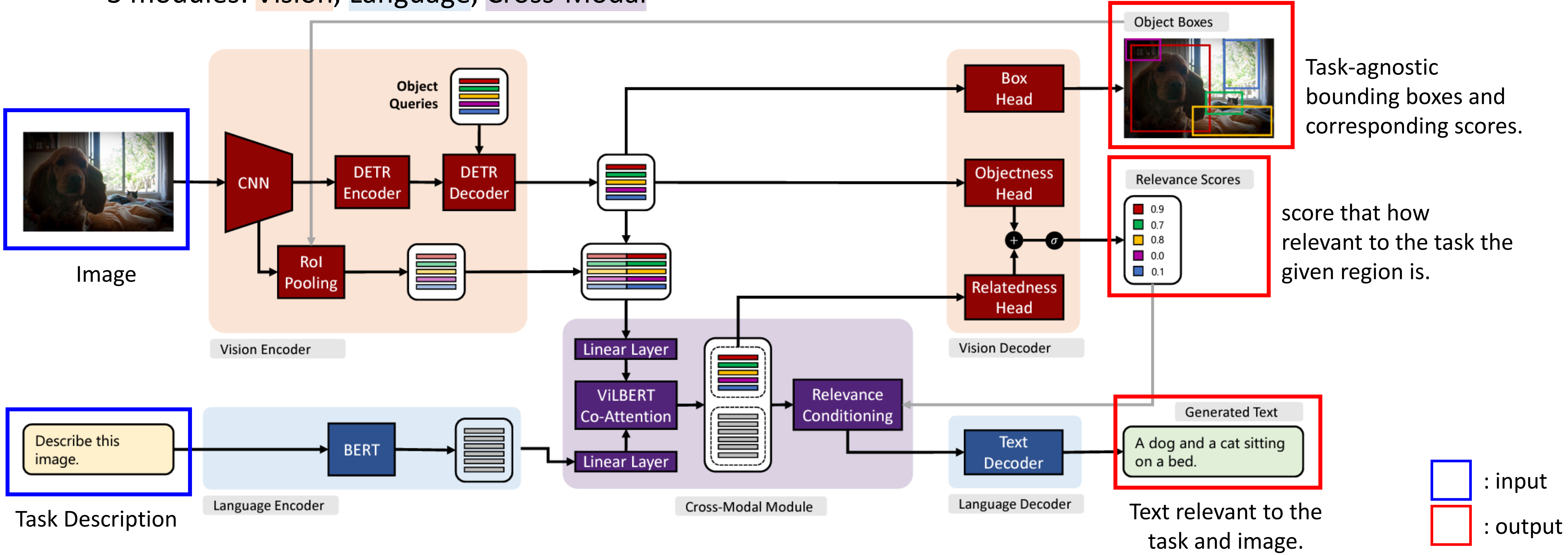



Forgot previous tasks
when training task 10

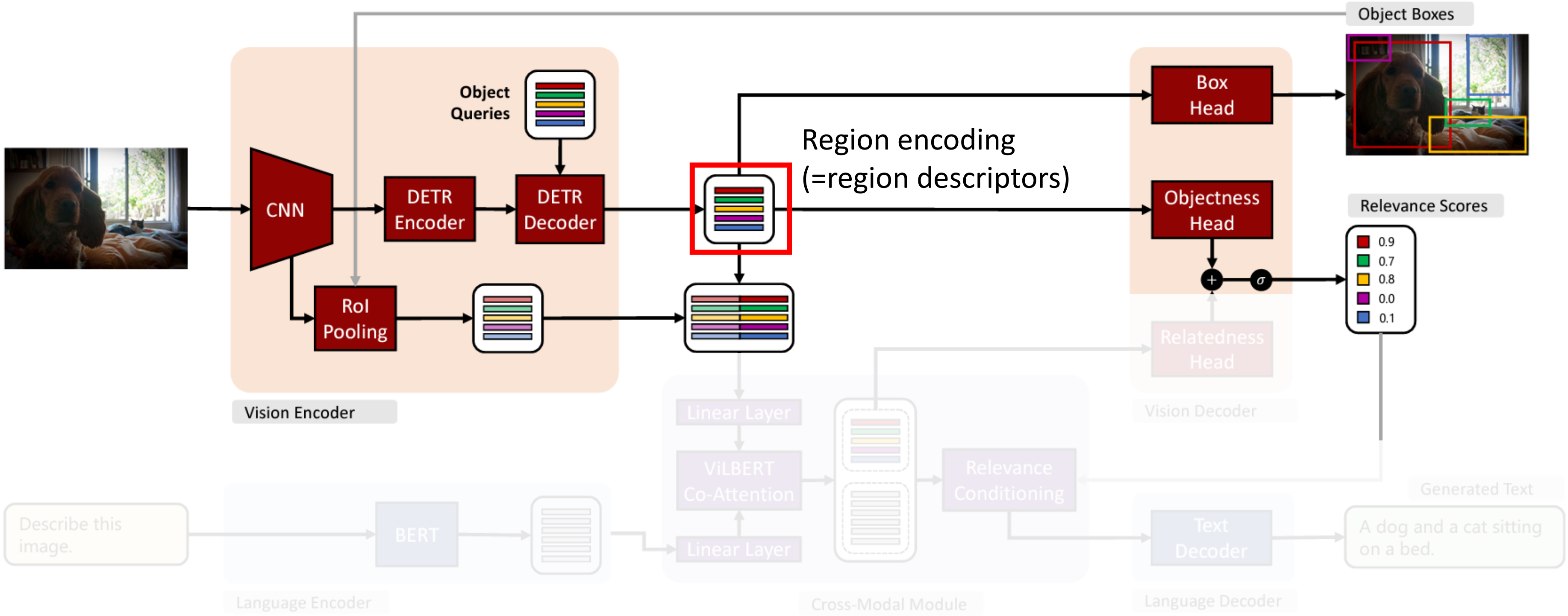
# GPV-1: Architecture

- **Input**: Image, **Text (indicates which task to be performed)**
- **Output**: Image, Text, Object Boxes, Relevance Scores
  - # of output heads corresponds to the # of output modalities (<< # of classes)
- 3 modules: Vision, Language, Cross-Modal



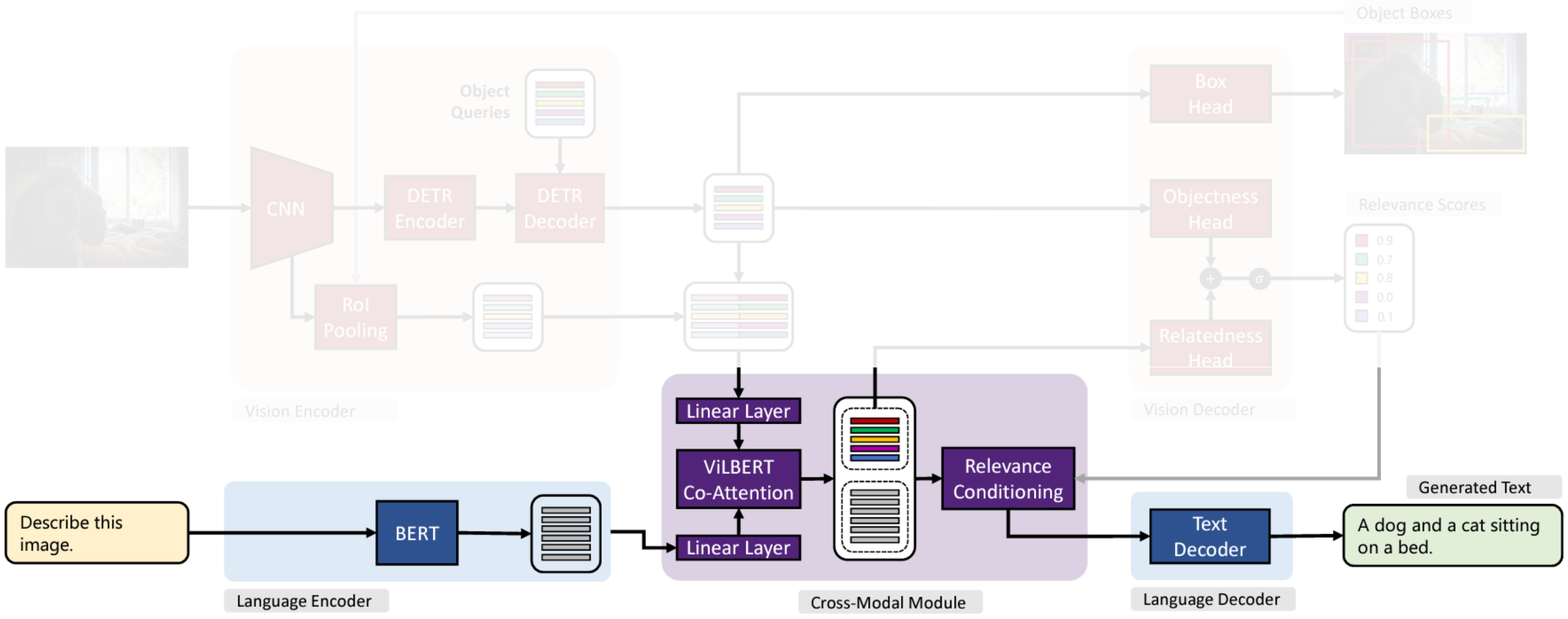
# Vision Modules

- Encoder: CNN Backbone + DETR + RoI pooling on features from CNN
- Decoder
  - Box Head: predict R(=100) bounding boxes from region descriptors
  - Objectness Head: binary objectness classification layer (*objectness: whether it has an object or not)
- Vision Encoder and Decoder initialized with pre-trained DETR and finetuned



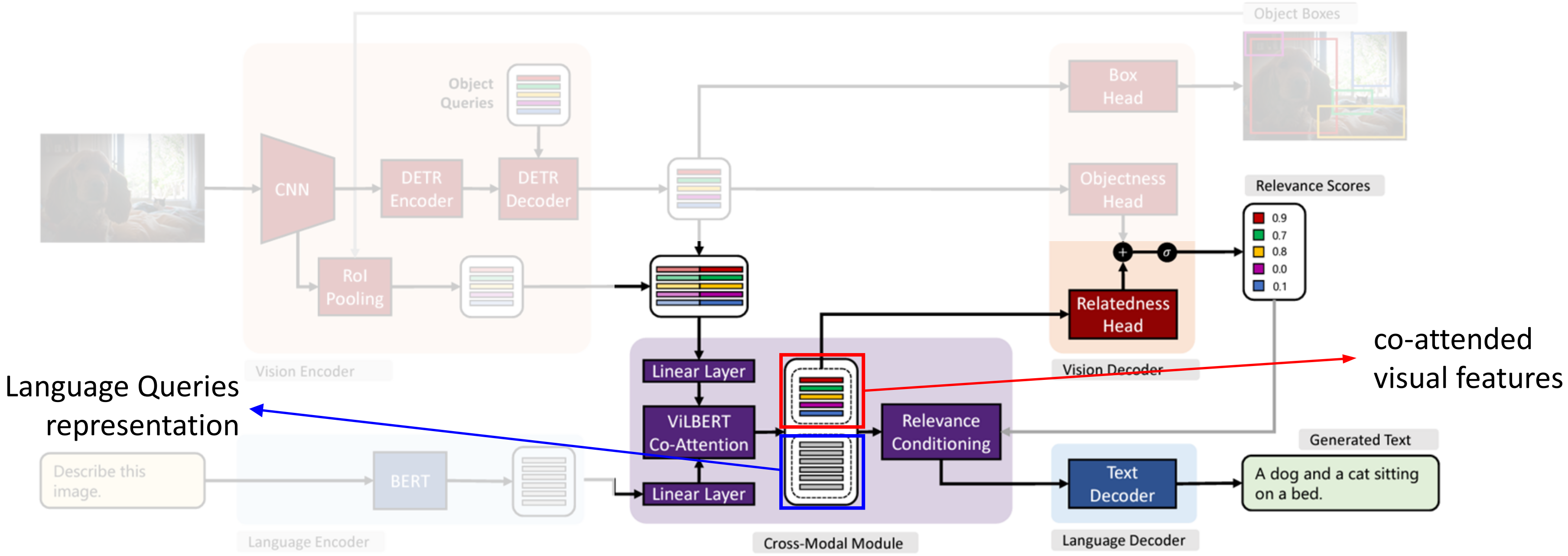
# Language Modules

- Encoder: encode the given task description.
  - Sub-word tokenization : robust to out-of-vocabulary words
  - Pre-trained BERT: handling paraphrases and zero-shot generalization to novel task descriptions.
- Decoder: outputs <u>words to classify, describe, or answer the input</u>.



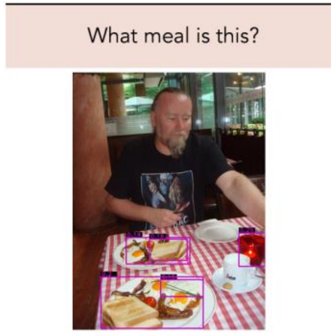
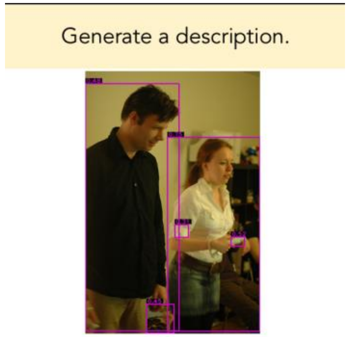
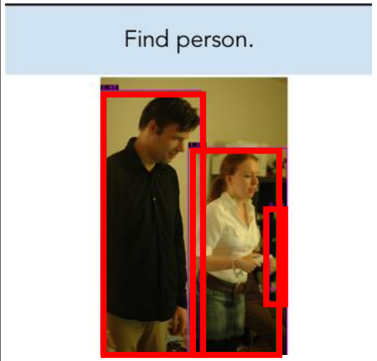
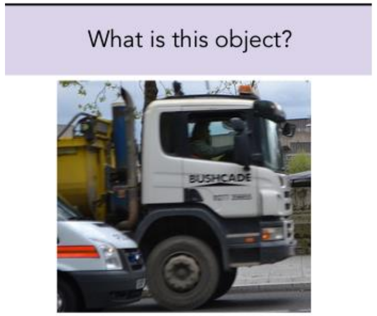
# Cross-Modal modules

- **Co-attention** from ViLBERT
  - cross-contextualize representations from the visual and language encoders
- **Relatedness head**
  - learns to indicate relevance of regions to the task description..
- **Relevance Conditioning**
  - modulates the co-attended visual features with relevance scores.
  - enables supervision from the text decoder to affect the relatedness and objectness heads.



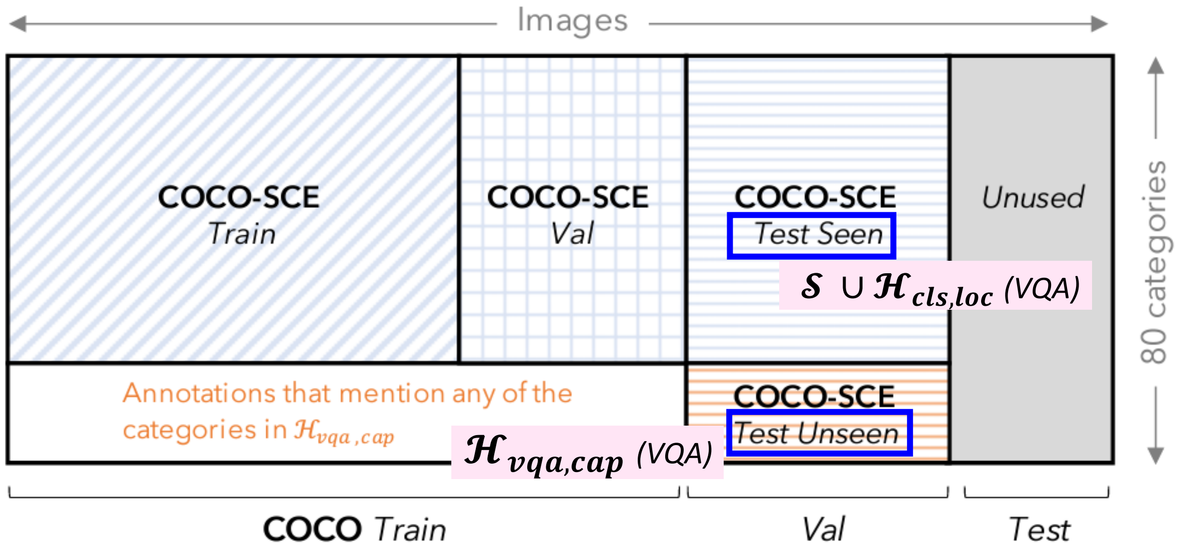
# Tasks in GPV-1

- Jointly Trained on 4 tasks (VQA, Captioning, Localization, and Classification).
  - Each mini-batch consists of a mix of samples from all 4 tasks
- Referring Expressions (a.k.a. RefExp) to test the learning ability of GPV-1

| Skills | VQA (text) | Captioning (text) | Localization (boxes) | Classification (text) | RefExp (box) |
|---|---|---|---|---|---|
| |  *Breakfast* |  *A man and a woman playing a game with remote controllers.* |  |  *Truck* *image patch given |  |
| **Loss** | NLL of the ground truth answer text | NLL of the annotated caption | DETR's Hungarian Loss | NLL of text output | DETR's Hungarian Loss |
| **Evaluation Metrics** | Annotator-agreement weighted answer accuracy | CIDEr-D | mAP | Accuracy | mAP |

# COCO-SCE

- Splitting 80 classes of COCO dataset **to test unseen combinations of concepts - skills**
  - **3 disjoint sets** $\mathcal{H}_{vqa,cap}, \mathcal{H}_{cls,loc}, \mathcal{S}$ specifying which tasks can use them for training and validation
  - $\boldsymbol{\mathcal{H}}_{\boldsymbol{vqa,cap}}$: 10 classes held-out from the VQA and captioning tasks in the train/val sets
  - $\boldsymbol{\mathcal{H}}_{\boldsymbol{cls,loc}}$: 10 different classes held-out from the classification and localization tasks in the train/val sets
  - $\boldsymbol{\mathcal{S}}$: 60 remaining classes not held out from any tasks
- When a category is held out, any <u>annotations</u> containing that word are <u>not used</u> for training or validation.




**Image**

**boat** is held-out for VQA

**Annotation**

"What color is the boat?"→"Orange"

"Is it a sunny day?"→"Yes"

"Locate a boat"

"What is this object?"

https://cocodataset.org/#explore

# Experiments

1. Effectiveness compared to specialized models **(→ Generality of Architecture)**
2. Ability to apply learned skills to unseen concepts for that skill

   **(→ Generality of Concepts across Skills)**
3. Efficiency at learning new skills and retention of previously learned skills

   **(→ Generality of Learning)**
4. Ablations

# Models in Experiments

- **Specialized Models (Baseline)**
  - ViLBERT (VQA), VLP (captioning), Faster-RCNN (localization), Resnet-50 (classification)
- **1-Task GPV-1**
  - trained only on individual task data (no joint training)
- **Multitask GPV-1**
  - joint training on all 4 tasks

# Generality vs. Effectiveness

- Test if the **general-purpose architecture is effective** compared to single specialized models
- In general, Multitask GPV-1 <u>improved performance (at least, comparable)</u> compared to single-task models.

  ∴ **Generality of GPV-1 is not at the cost of effectiveness.**

| Split | Model | VQA | Cap. | Loc. | Class. |
|---|---|---|---|---|---|
| COCO-SCE | [a] Specialized Model | 56.6 | 0.832 | 62.4 | 75.2 |
| | [b] 1-Task GPV-1 | 55.9 | 0.855 | **64.8** | 75.3 |
| | [c] Multitask GPV-1 | **58.8** | **0.908** | 64.7 | **75.4** |
| COCO | [d] Specialized Model | 60.1 | 0.961 | **75.2** | 83.3 |
| No Held-out | [e] Multitask GPV-1 | **62.5** | **1.023** | 73.0 | **83.6** |

**Table 1**: Comparison to special purpose baselines

# Skill-Concept Generalization

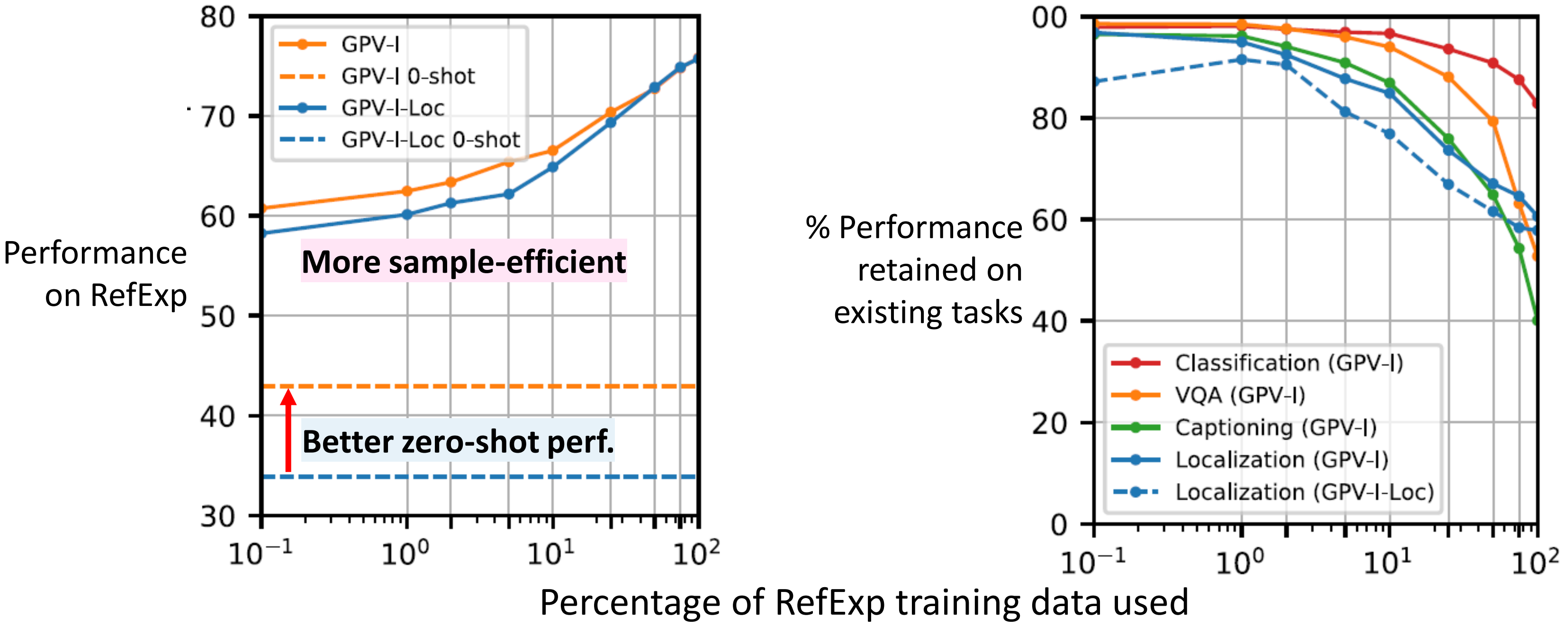
- Handling **unseen skill-concept combinations during training**
- 1-Task GPV-1: no access to held-out concepts
  - A baseline to account for learned priors and dataset biases by the GPV-1 architecture
- Multitask GPV-1 Oracle: trained on the COCO training split
  - Model exposed to held-out concepts for all tasks
  - a loose upper bound for the "unseen" split.
- **General-purpose architecture > Specialized models (especially for "Unseen")**
  - ✓ Multitask GPV-1 is more beneficial to VQA and Captioning compared to Localization and Classification

| Model | VQA | | | Captioning | | | Localization | | | Classification | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* |
| [a] Specialized Model | 56.6 | 57.2 | 45.2 | 0.832 | 0.867 | 0.501 | 62.4 | 68.1 | 7.4 | 75.2 | 83.0 | 0.0 |
| [b] 1-Task GPV-1 | 55.9 | 56.5 | 41.9 | 0.855 | 0.891 | 0.524 | **64.8** | **69.8** | 16.4 | 75.3 | **83.1** | 0.0 |
| [c] Multitask GPV-1 | **58.8** | **59.3** | **47.7** | **0.908** | **0.944** | **0.560** | 64.7 | 68.8 | **25.0** | **75.4** | 82.6 | **5.4** |
| [d] Multitask GPV-1 **Oracle** | 61.4 | 61.3 | 64.0 | 1.018 | 0.997 | 0.939 | 73.0 | 72.7 | 76.0 | 83.6 | 83.4 | 85.7 |

**Table 2**: Skill-Concept Generalization
(Results on COCO-SCE, Test is Full COCO-SCE test split)

# Learning Generalization

- Test if GPV-1 learn **new skills sample-efficiently without forgetting previous-learned skills.**
  - New task: Referring Expressions (fine-tuned on RefCOCO+ dataset)
  - GPV-1 (Multi-task) vs. GPV-1-Loc (pre-trained only on the localization task)
- Multitask GPV-1: <u>Better zero-shot performance and better sample-efficiency (left figure)</u>
  - Better starting point with the learning of attributes and additional nouns
- Multitask GPV-1 <u>alleviates forgetting (right figure)</u>



# Ablations

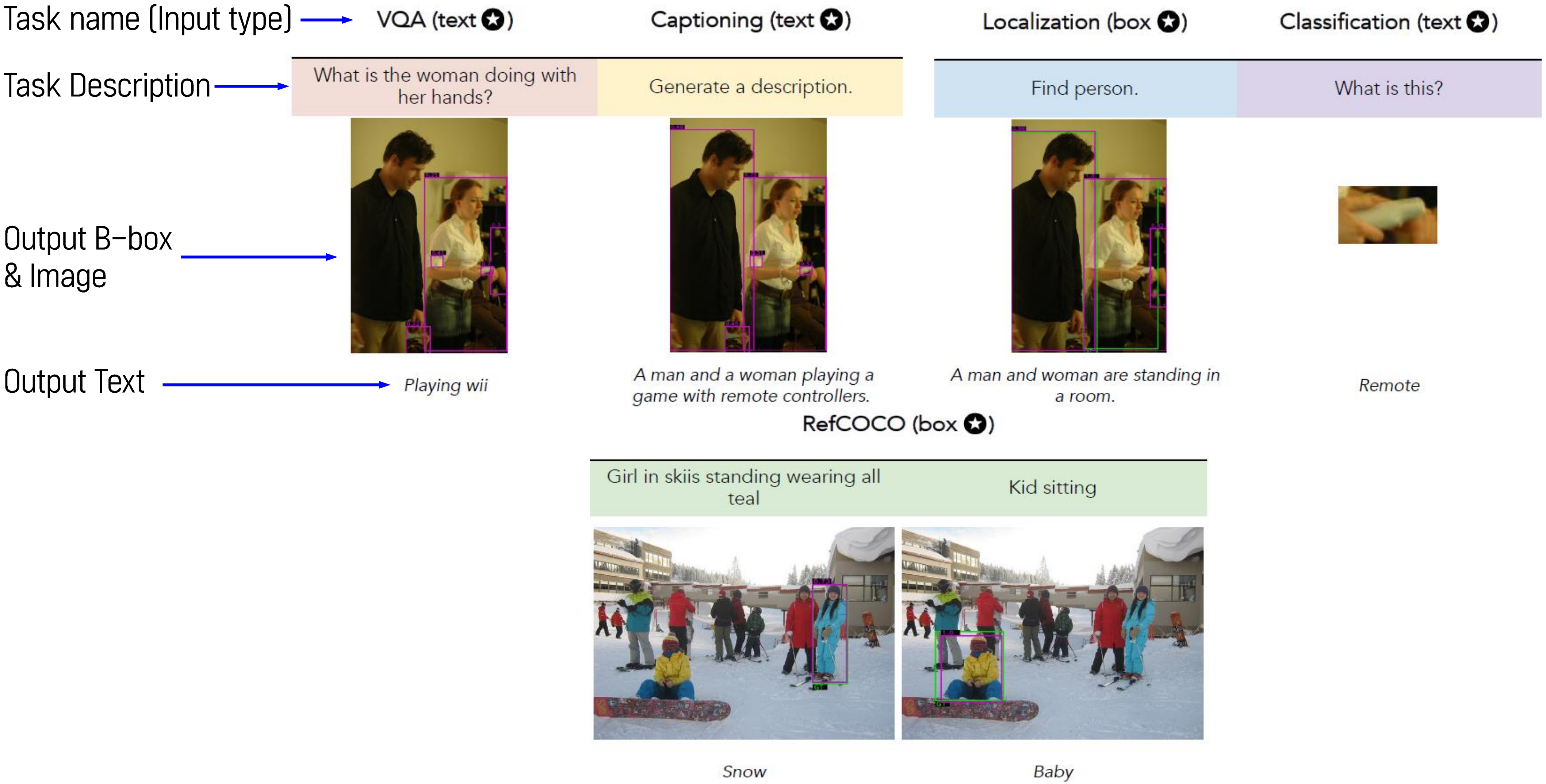
- Factors To Test
  - **RoI features** helps for VQA and Captioning
  - **Finetuning** contributes to performance across all tasks.
  - **Modality-specific Output Heads** works better in most cases compared to Task-specific output heads.

| | VQA | Cap. | Loc. | Class. |
|---|---|---|---|---|
| [a] Multitask GPV-1 | **58.8** | **0.908** | 64.7 | 75.4 |
| [b] *w/o RoI features* | 54.9 | 0.898 | **65.3** | **76.6** |
| [c] *w/o Fine-Tuning* | 56.4 | 0.883 | 63.4 | 71.5 |

**Table 3** (Top): Ablations for RoI features, Fine-tuning
**Table 4** (Bottom): Ablations for modality-specific heads

| | | VQA | | | Captioning | | | Localization | | | Classification | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Params | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* | Test | *Seen* | *Unseen* |
| [a] Head per Task | 311M | 57.67 | 58.20 | 45.86 | **0.884** | **0.922** | 0.533 | 62.05 | 65.76 | 26.13 | 74.26 | **81.93** | 0.00 |
| [b] Head per Modality | 236M | **57.73** | **58.22** | **46.91** | 0.881 | 0.915 | **0.547** | **62.53** | **66.13** | **27.75** | **74.58** | 81.76 | **5.10** |

# Example of GPV-1’s works on 5 tasks



# Contributions

- ✓ Trained to perform **any image task** that can **be performed using words or boxes**
- ✓ **Higher** (at least, comparable) **performance** than the previous specialized models
  - comparable results to specialized systems when trained on individual tasks
  - outperforms when trained jointly
- ✓ Learn new tasks sample-efficiently

# Things to be discussed

- ✓ **Slower** than specialized systems
  - GPV-1 for detection requires a separate localization inference per object category
- ✓ Still **far to go in skill-concept generalization**
  - Huge gap between Multitask GPV-1 and GPV-1 Oracle (Table 2)
- ✓ **Catastrophic forgetting** is remained
- ✓ Skills and concepts outside COCO unexplored
- ✓ Limited coverage of task
  - Currently not available for Image Manipulation or generation tasks (colorization, segmentation)
- ✓ **Non-Image inputs** (videos, point clouds…) should be handled as well.