
# Unified Image-Text Modeling

Zhengyuan Yang

Microsoft

# Agenda

- Image-text tasks overview; Motivation of unification
- Unified image-text models
- Summary and discussion

# Image-text Tasks Overview

## Close-set classification


Popular Image-text Tasks:
VQA, GQA, VisDial, VCR, NLVR2, image-text matching

## Open-ended text sequence


Popular Image-text Tasks:
Image captioning, paragraph captioning, storytelling, open-ended VQA

Unified Image-Text Modeling

## Box/mask localization


Popular Image-text Tasks:
Referring expression comprehension/ segmentation, phrase grounding, grounded captioning

## Pixel prediction

A donut on a white plate next to a cup of latte.

Popular Image-text Tasks:
Text-to-image synthesis, text-based image editing

# Close-set Classification

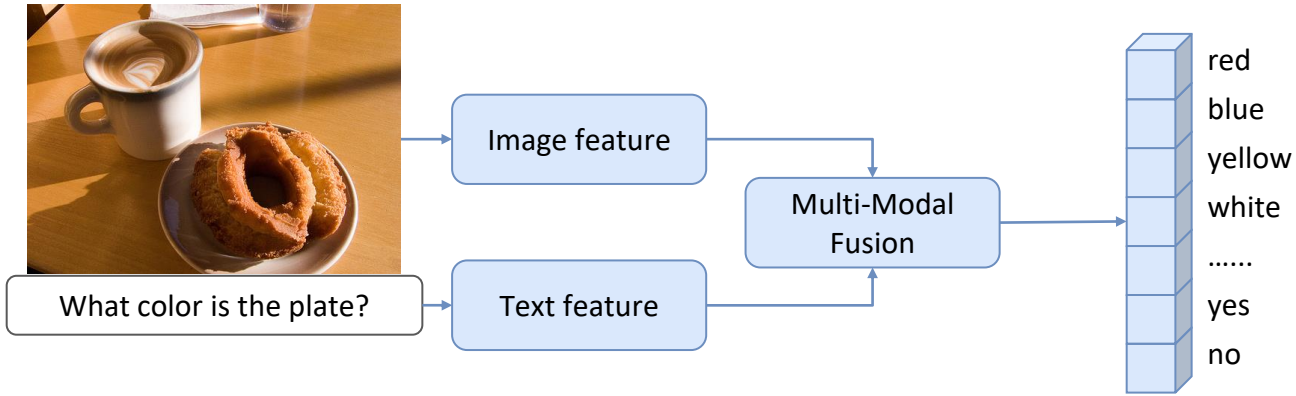
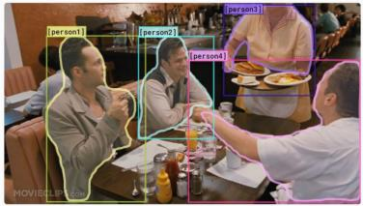




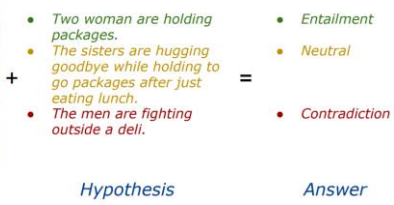
Image-text matching, NLVR2

True
False




VCR

a)
b)
c)
d)




Visual Entailment

Entailment
Neutral
Contradiction

Image credit: from the original papers: NLVR2, VCR, Visual Entailment

# Open-ended Text Sequence

Optional text input/ Empty

Image feature

Text feature

Multi-Modal Fusion

Auto-regressive decoder

A donut on a

t=0 t=1 t=2 t=3 ...

donut on coffee

integer theorem

"Open-vocab": language model tokenizer vocabulary, e.g., 30522, 50265

A donut on a white plate next to a cup of latte.

Image captioning

This image is of a family celebrating Christmas. They are all gathered around a dinner table, with a turkey and other food on it. The family is smiling and seems to be enjoying themselves. There is a Christmas tree in the background and some Christmas lights on the walls.

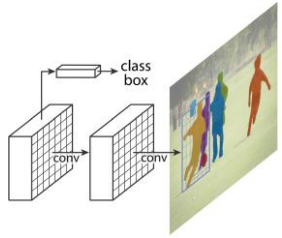
Paragraph Captioning

What color is the plate? The plate is white.

Open-ended VQA

Image credit: Yujia Xie

# Box/mask Localization

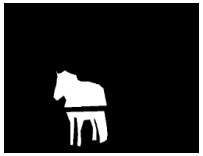



The donut on the white plate

Image feature

Text feature

Multi-Modal Fusion



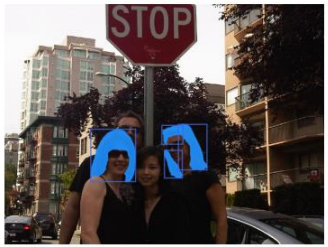
$[x_1, y_1, x_2, y_2]$:
[90.1, 83.2, 184.9, 180.4]

A man with pierced ears is wearing glasses and an ora
A man with glasses is wearing a beer can crotched ha
A man with gauges and glasses is wearing a Blitz hat
A man in an orange hat starring at something
A man wears an orange hat and glasses.

Visual grounding
(REC, phrase grounding)

glass bottles

blonde hair

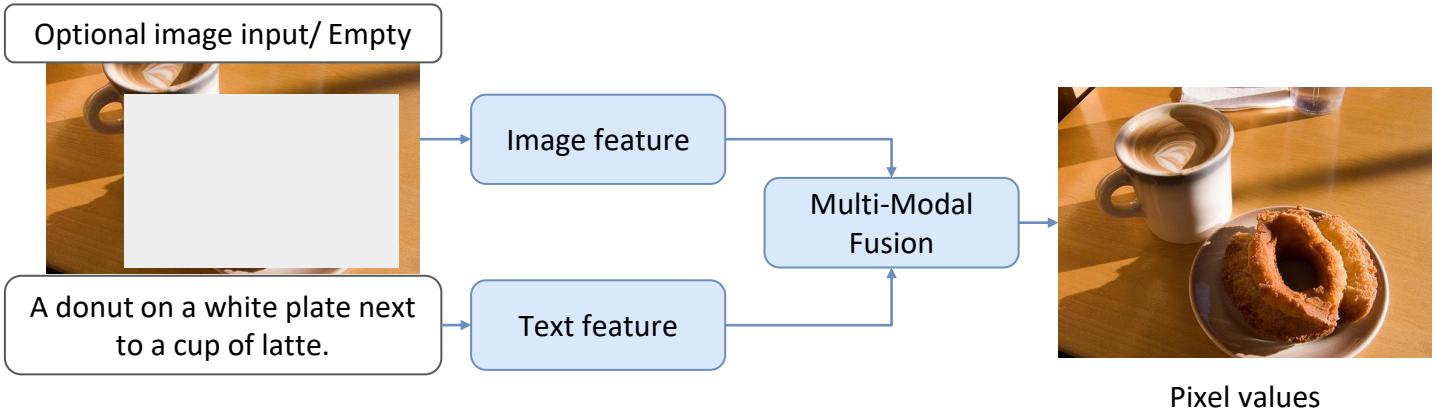



pedestrian crosswalk


Language-based segmentation
(RES)

Image credit: from the original papers: Flickr30K Entities, PhraseCut, DMS, Mask R-CNN

# Pixel Prediction




A donut on a white plate next to a cup of latte.

Text-to-image synthesis


Text-based image editing

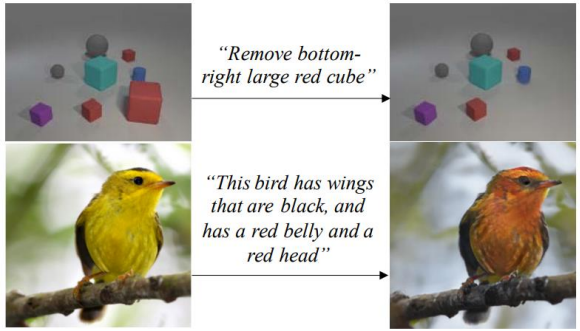
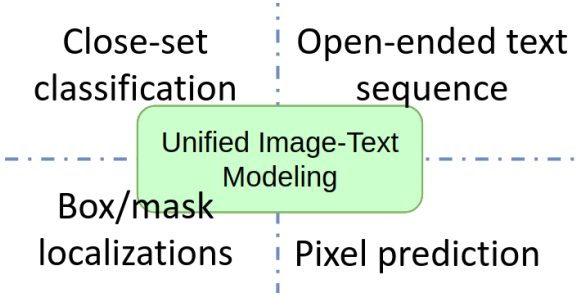


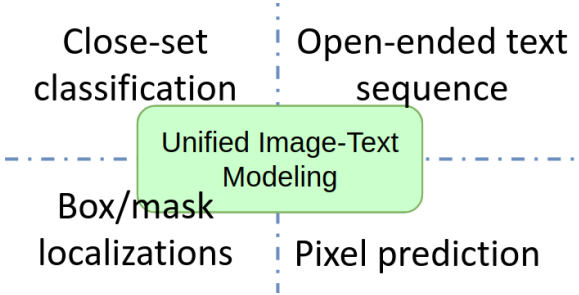

Image credit: Jing Shi

# Why Unified Image-Text Modeling

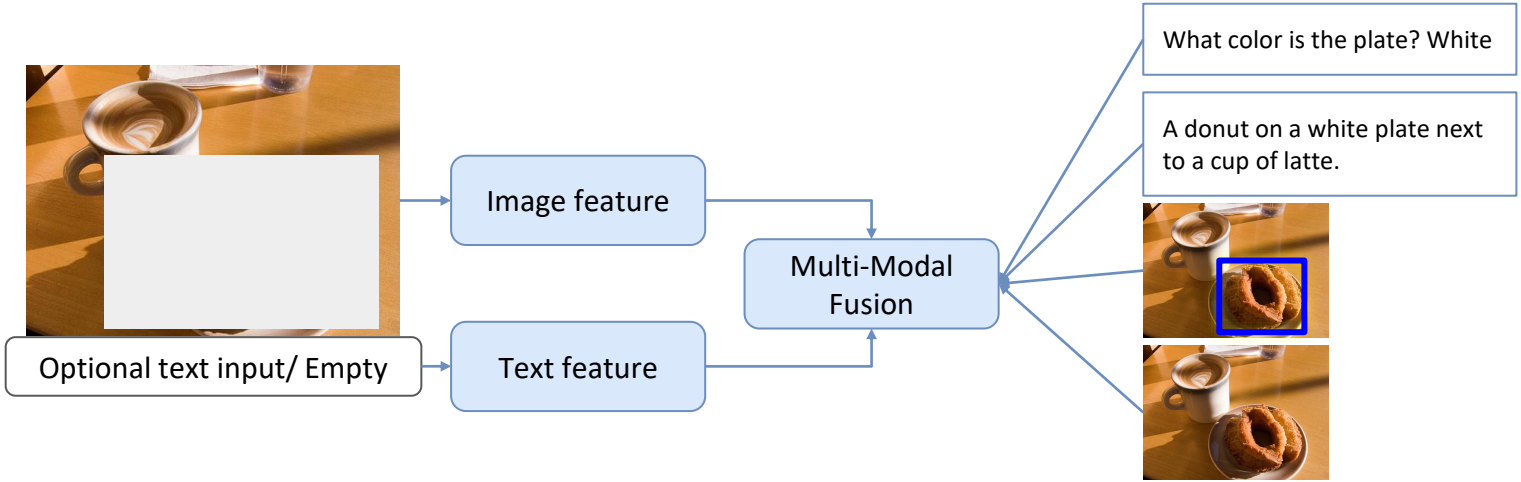
- Better performance
- New capabilities
- Task-agnostic unified systems



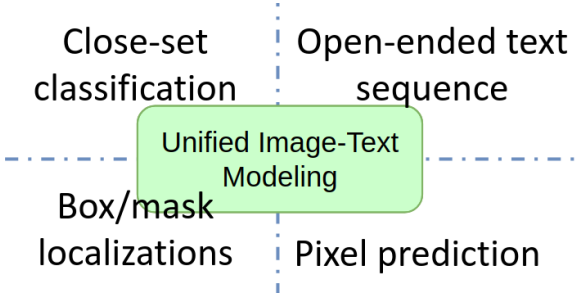
# Why Unified Image-Text Modeling



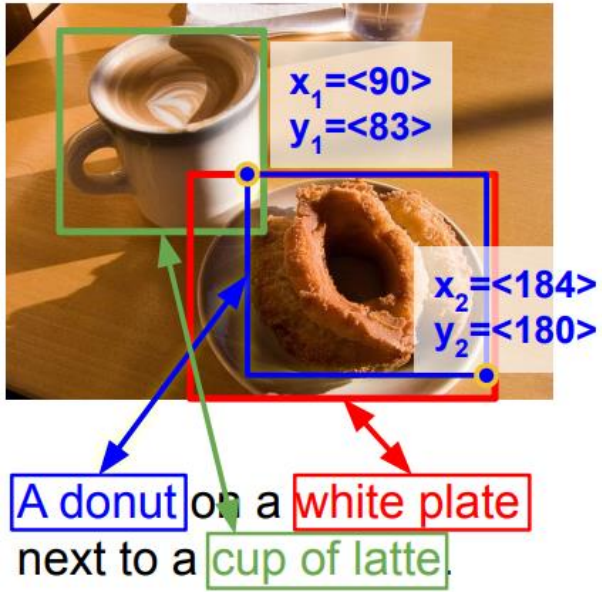

- Better performance
  - Similar abilities; Multi-task training
  - Extra data/annotations from other tasks



# Why Unified Image-Text Modeling

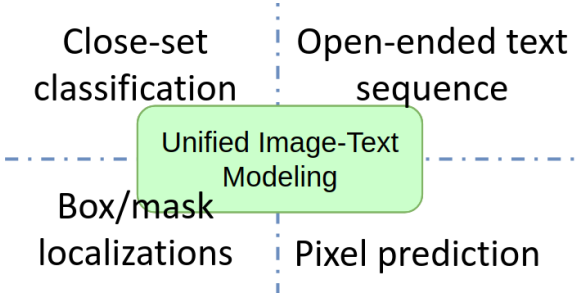



- New capabilities

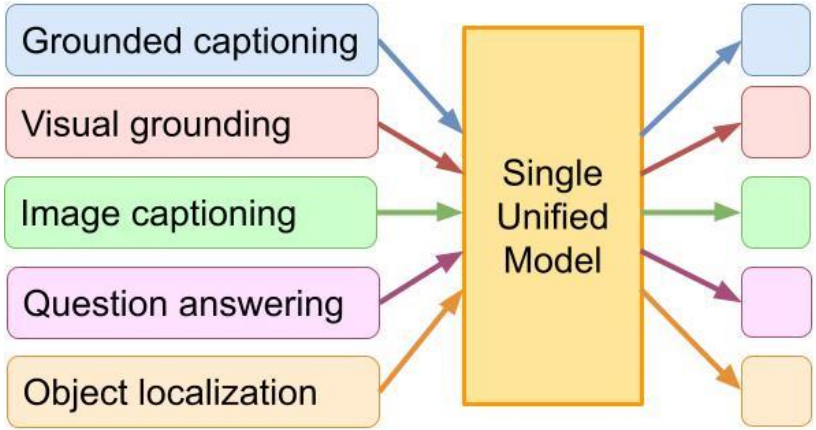


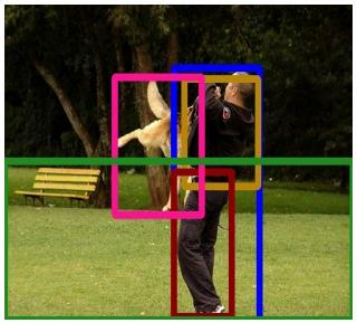

- E.g., text + box; grounded captioning
- More comprehensive and interpretable image description

# Why Unified Image-Text Modeling




- Task-agnostic unified systems
  - Ease framework design; Avoid model copies
  - Capability generalization



**A man in a black jacket and black pants is playing with a dog in a park.**

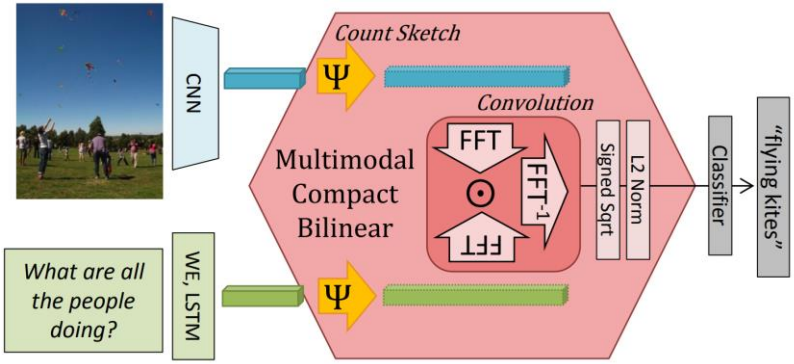
Generalizing grounded captioning to COCO

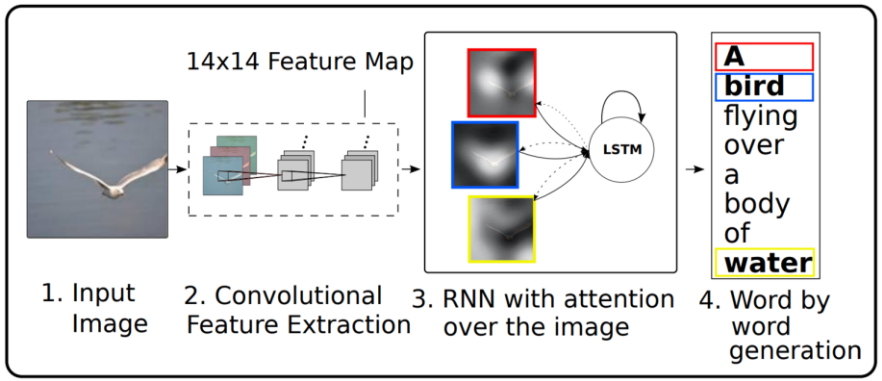
# Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
  - Classification as text generation
  - Model design and training
  - Unify text and box
  - Textualize visual outputs
- Summary and discussion

# VL Research

- Models curated for tasks
- Fast-forward to vision-language pre-training (VLP)



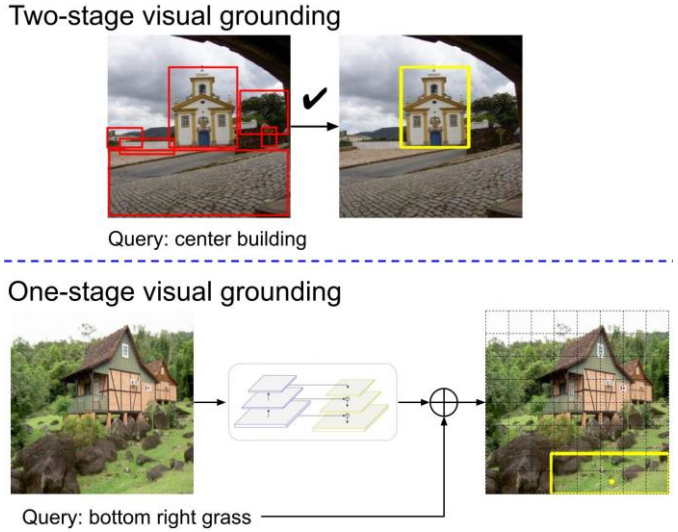
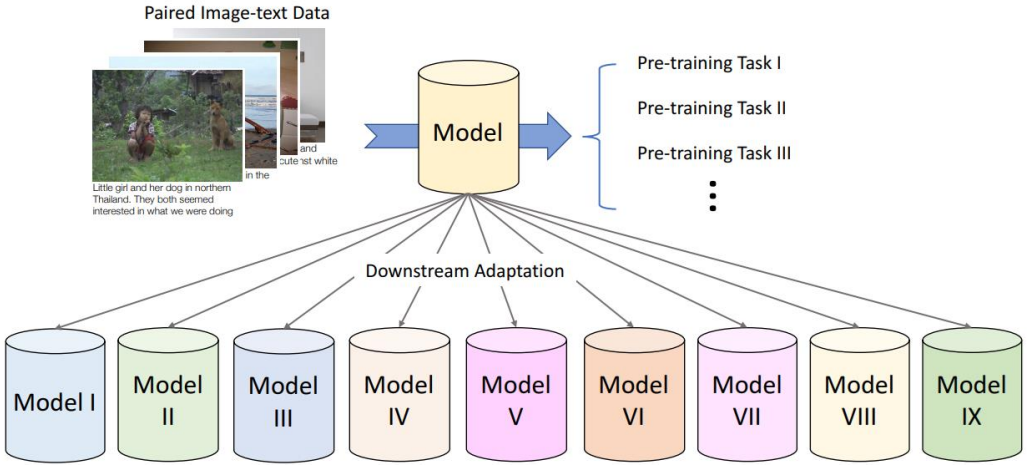




Image credit: from the original papers

# Vision-language Pre-training (VLP)

- Large-scale transformer-based self-supervised pre-training
- Reuse the same pre-training weight as initialization point
- Separate output head and finetune model copies for different downstream tasks



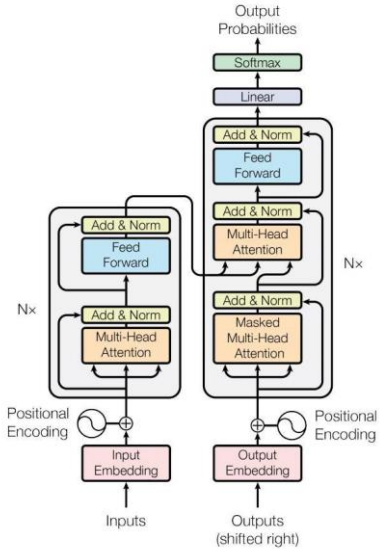
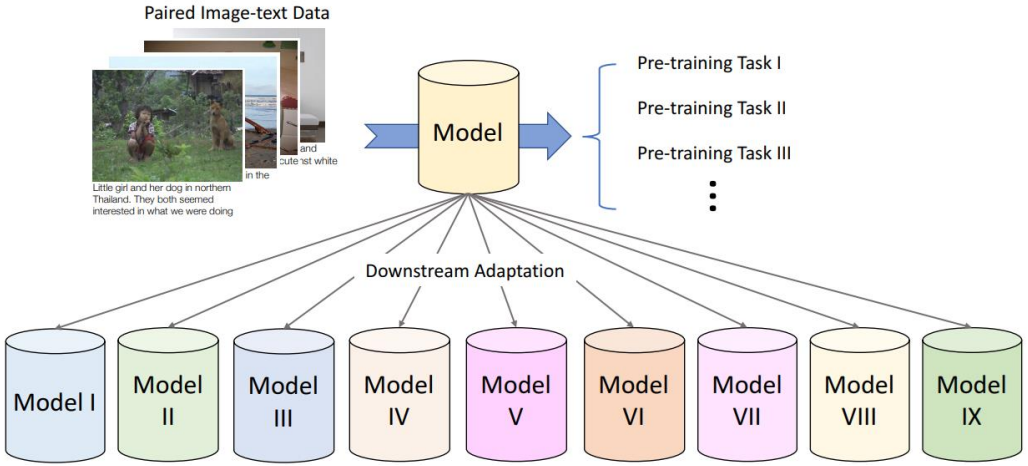




Image credit: CVPR 2020 VL Tutorial; Attention Is All You Need, NeurIps 2017

# Vision-language Pre-training (VLP)

- Large-scale transformer-based self-supervised pre-training
- Reuse the same pre-training weight as initialization point
- Separate output head and finetune model copies for different downstream tasks



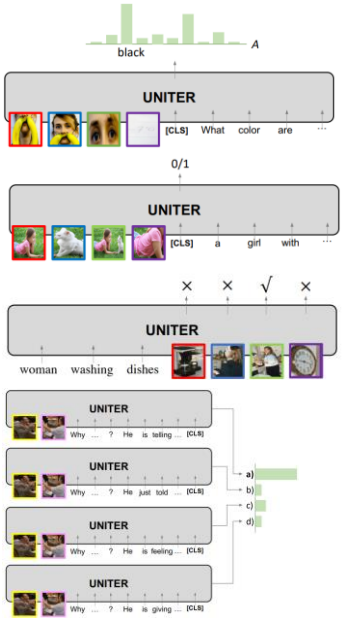
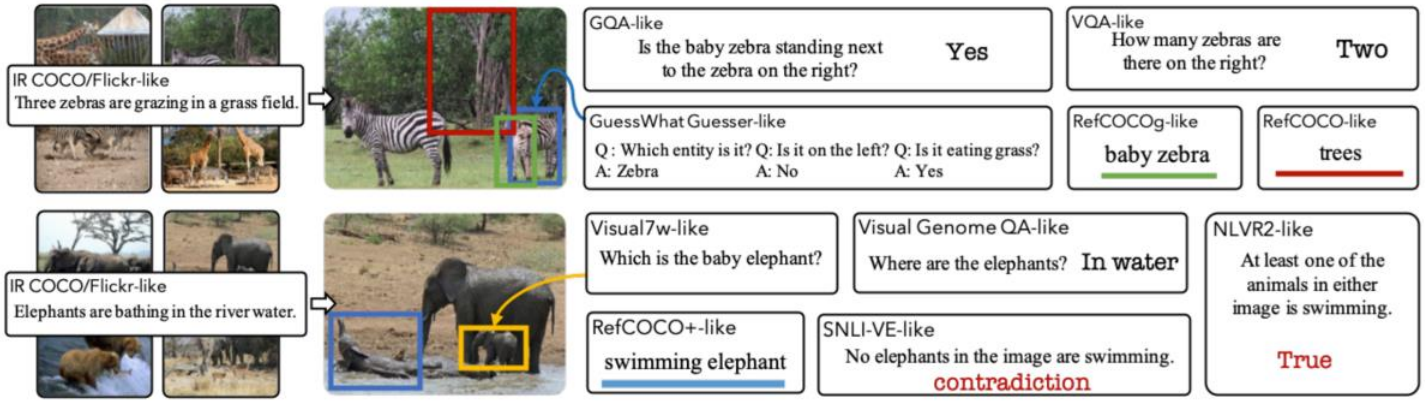




Image credit: CVPR 2020 VL Tutorial; UNITER: UNiversal Image-TExt Representation Learning, ECCV 2020

# 12-in-1

- Single model for 12 tasks (12*270M=3B -> 270M)
- Relationships among tasks; better averaged performance
- Task-specific heads and objectives

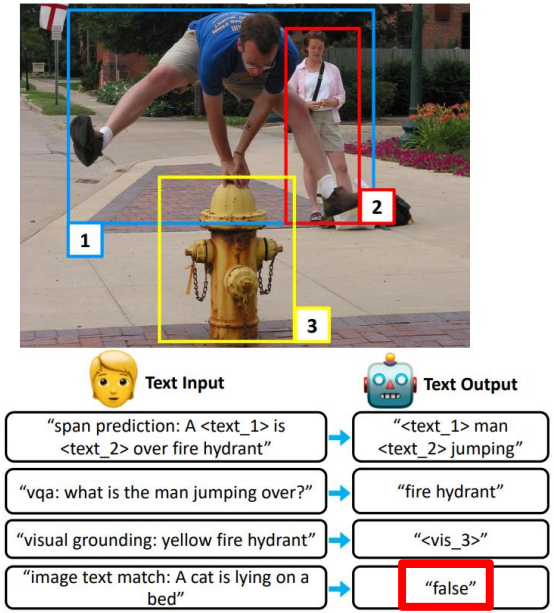



All these task require visually-grounded language understanding skills.

Image credit: 12-in-1: Multi-task vision and language representation learning, CVPR 2020

# VL-T5

- Image-text tasks as multimodal conditional text generation
- Avoid task-specific arch design and model copies



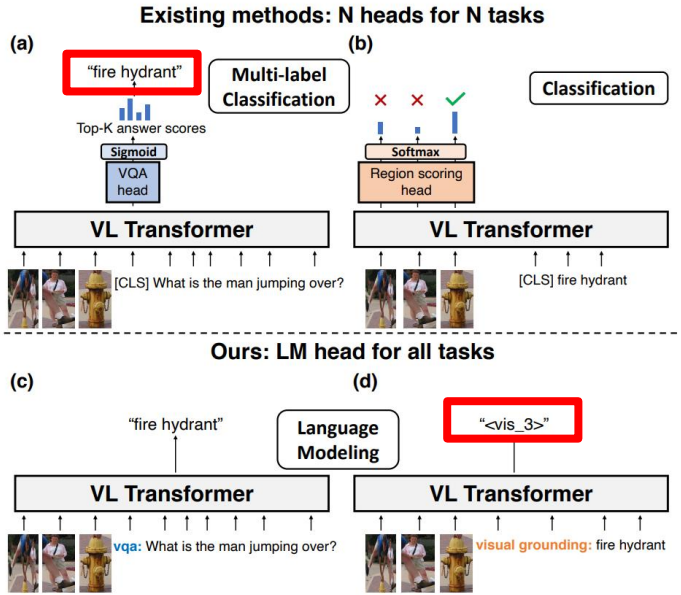




Image credit: Unifying vision-and-language tasks via text generation, ICML 2021

# VL-T5

- Image-text tasks as multimodal conditional text generation
- Avoid task-specific arch design and model copies

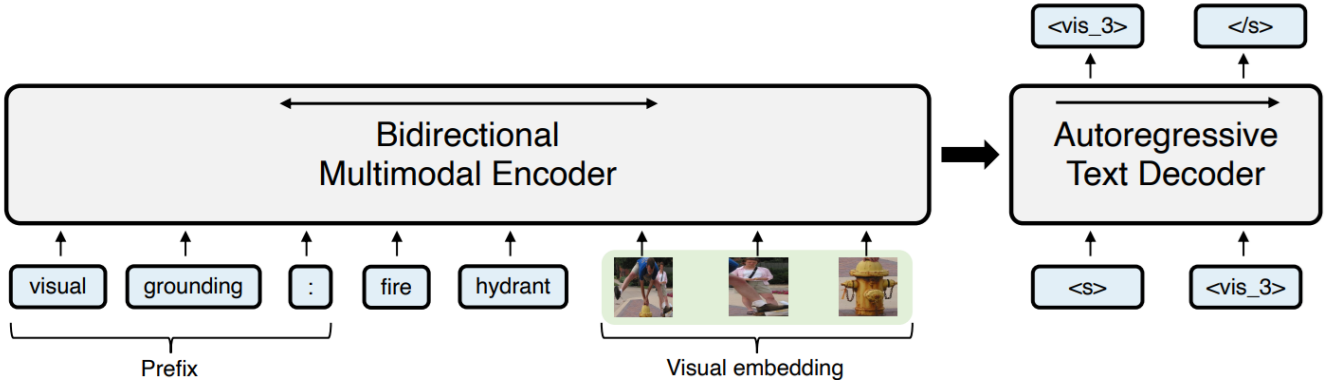



Image credit: Unifying vision-and-language tasks via text generation, ICML 2021

# Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
  - Classification as text generation
  - Model design and training
  - Unify text and box
  - Textualize visual outputs
- Summary and discussion

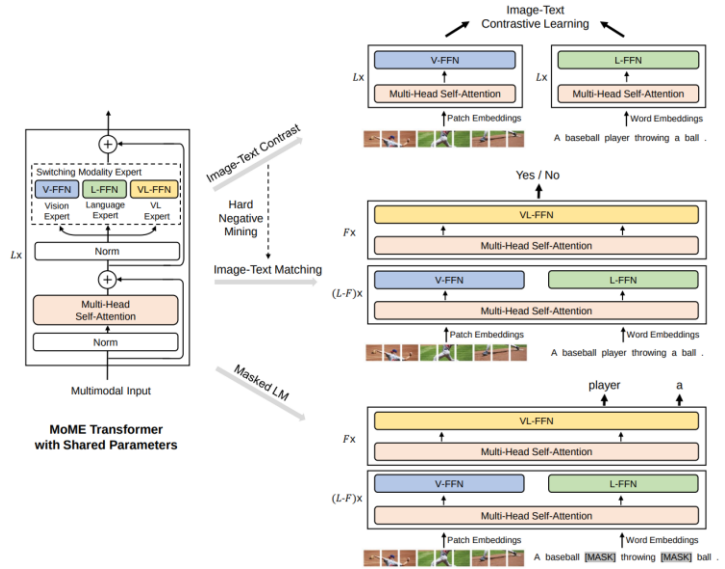
# Model design and training

- Output format unification is the first step, how to have different tasks and capabilities work well together
  - Partially-shared parameters
  - Modular network design
  - Data and training techniques

VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts
UFO: A unified transformer for vision-language representation learning
VL-BEiT: Generative Vision-Language Pretraining
FLAVA: A Foundational Language And Vision Alignment Model
Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text
UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning
UNIMO-2: End-to-End Unified Vision-Language Grounded Learning
etc.

# Partially-shared Parameters

- Mixture of modality experts
- Task-specific parameters



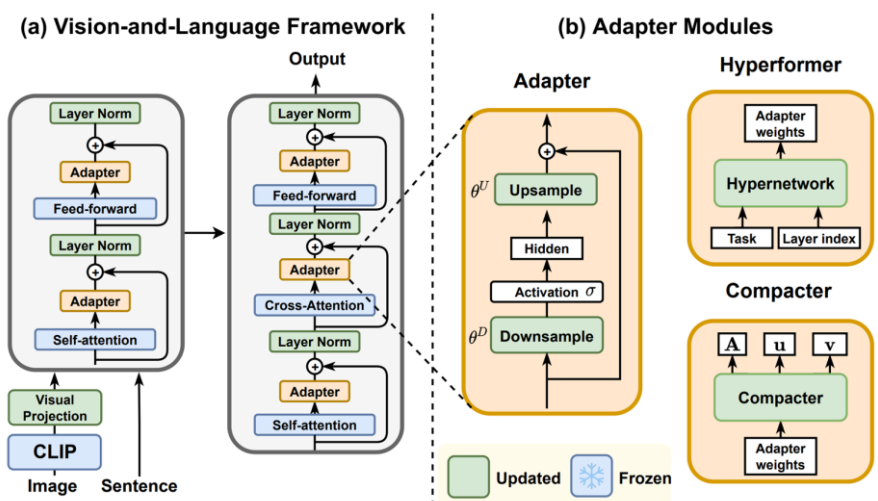




Image credit: VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts
VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks

# Modular Network Design

- Unimodal encoders
  for single-modality tasks
- Reuse adjusted submodules
  for different tasks

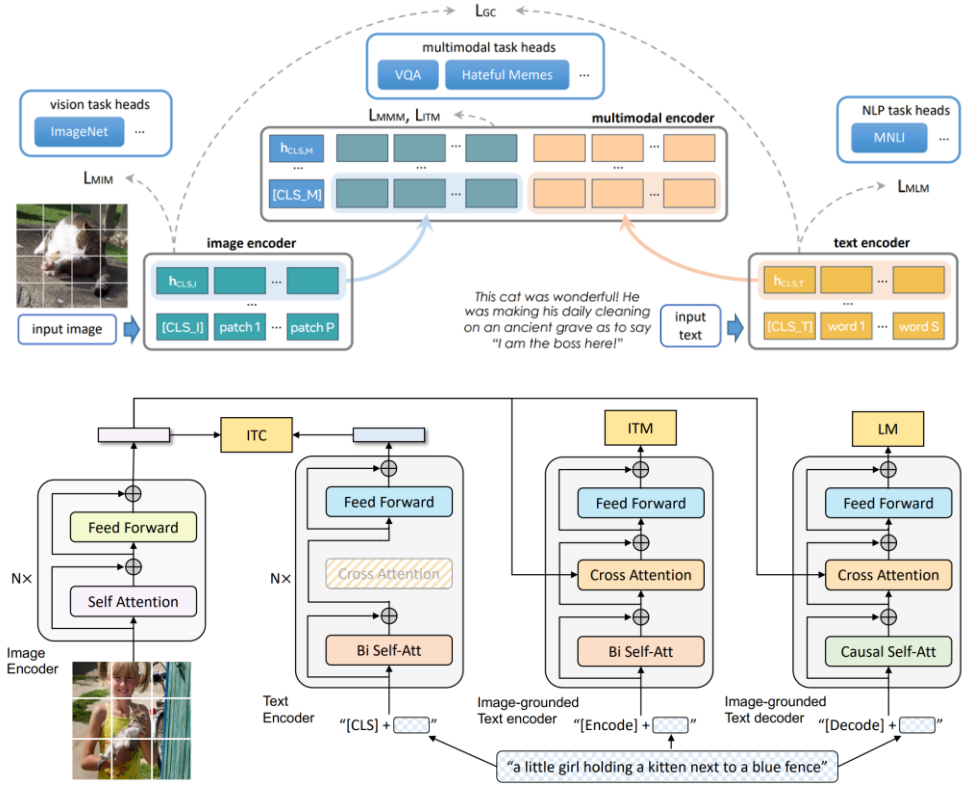
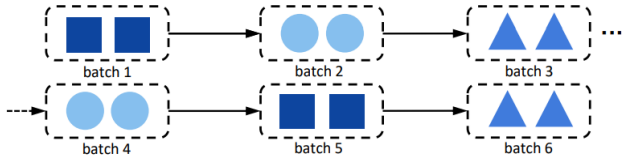



Image credit: FLAVA: A Foundational Language And Vision Alignment Model
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

# Data and Training Techniques

- Training corpus, batch construction
- Optimizing and loss design




$$\sum_{m=1}^{M} \lambda_m \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left[ -\sum_{\ell=1}^{L} \log p(y_\ell | y_{<\ell}, x_{\leq\ell}) \right]$$

$$G_{\text{txt}} = \frac{\partial \mathcal{L}_{\text{txt}}}{\partial \theta}, \quad G_{\text{img}} = \frac{\partial \mathcal{L}_{\text{img}}}{\partial \theta}$$

$$G_{\text{global}} = M \odot G_{\text{txt}} + (1 - M) \odot G_{\text{img}}$$

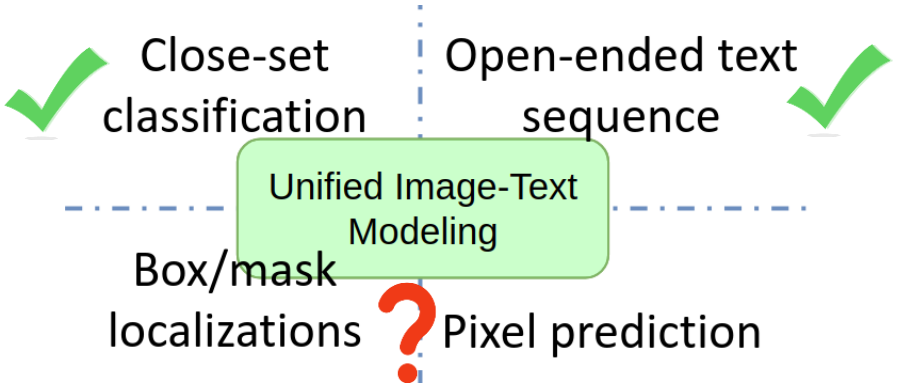
Image credit: ZeroVL: A Strong Baseline for Aligning Vision-Language Representations with Limited Resources
Flamingo: a Visual Language Model for Few-Shot Learning
Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text

# Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
  - Classification as text generation
  - Model design and training
  - Unify text and box
  - Textualize visual outputs
- Summary and discussion

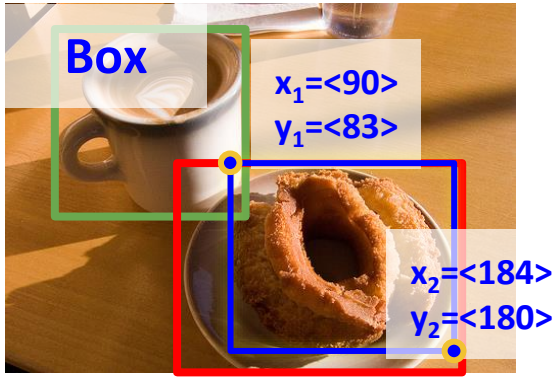
# Good for Text Outputs, Others?

- Back to unifying I/O formats
- Output formats:
  - ✓ Classification, text sequence
  - ? Box/mask
  - ? Pixel value
- Text+box as a case study

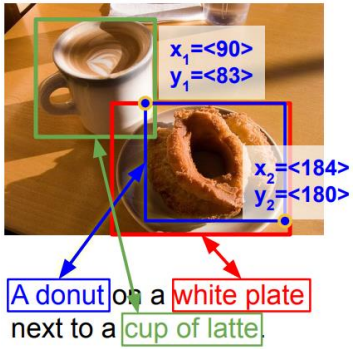


# Unifying Text and Box Outputs


- A donut on a white plate next to a cup of latte.

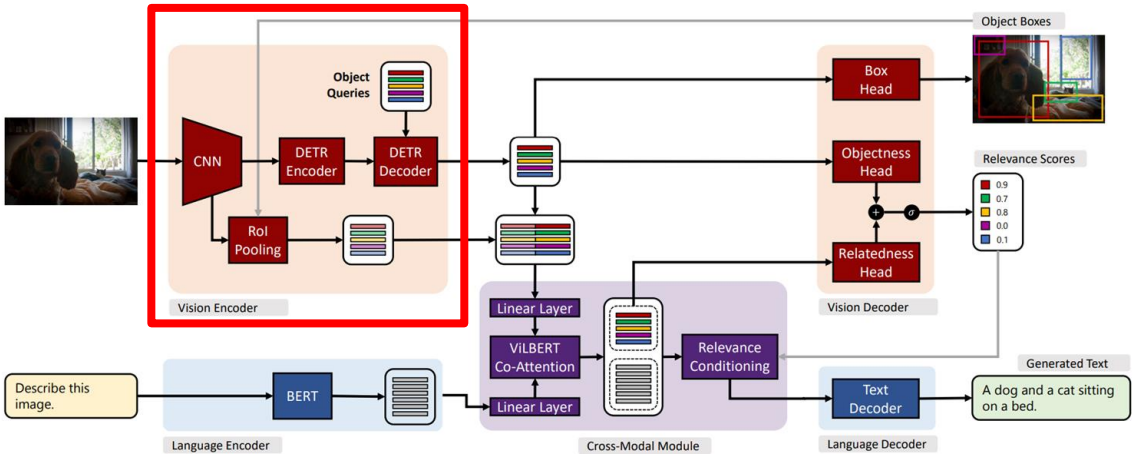


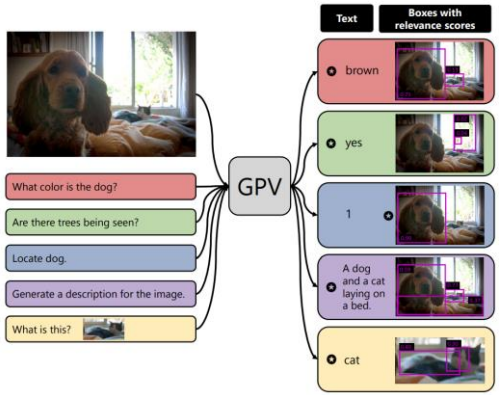

- OD/grounding: white plate, donut, coffee mug




1. Support both outputs
2. Word-box alignments

# GPV-1




- Text and box outputs: detector for image -> regions
- Word-box alignments: region index prediction
  - Related to the modeling in region-based VL models, but with detector E2E finetuned

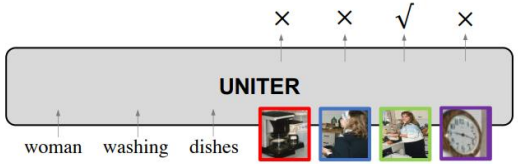
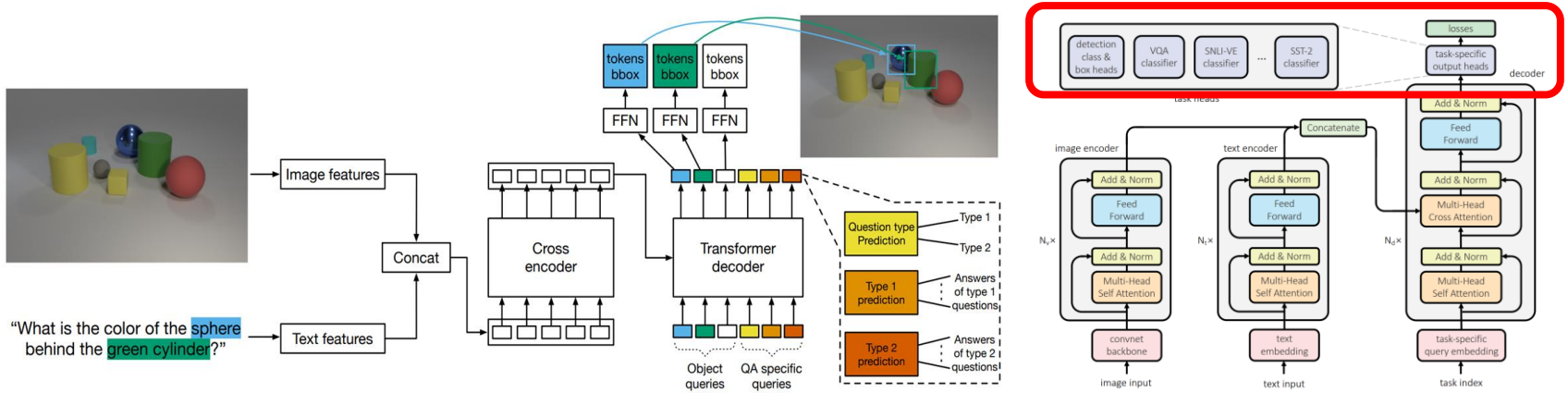



Image credit: Towards General Purpose Vision Systems, CVPR 2022

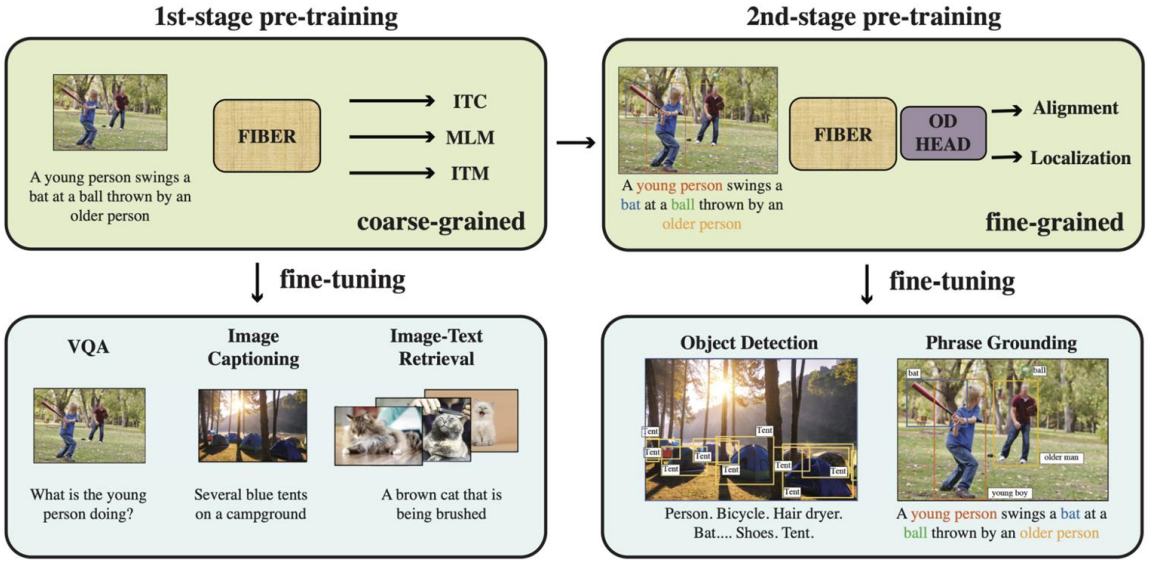
# MDETR, UniT




- Avoid explicit detection module?
- Text and box outputs:
  (box): coordinate regression; (text): heads for classification output
- Word-box alignments: input word index, or OD vocab

Image credit: Mdetr-modulated detection for end-to-end multi-modal understanding, ICCV 2021
UniT: Multimodal Multitask Learning with a Unified Transformer, ICCV 2021

# FIBER




- Challenge: resolution, computing cost trade-offs
- Coarse-to-fine two-stage vision-language pre-training
- Text and box outputs; Fusion in-the-backbone

Image credit: Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone

# Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
  - Classification as text generation
  - Model design and training
  - Unify text and box
  - Textualize visual outputs
- Summary and discussion

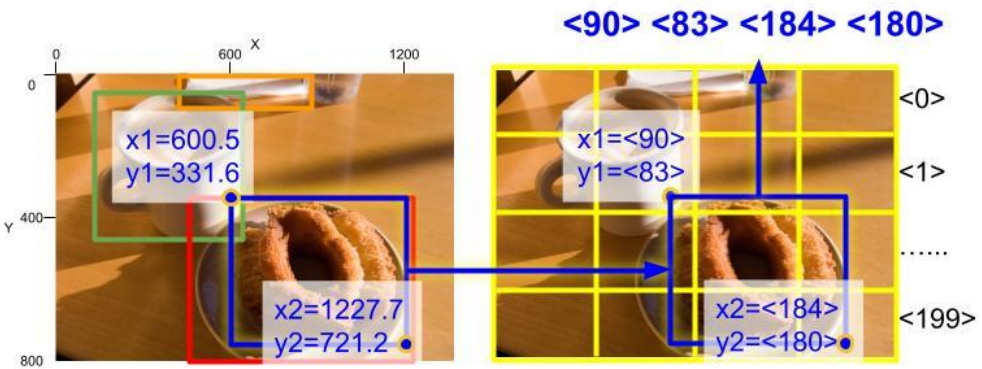
# Textualize Visual Outputs

- Visual (box, mask, pixel) outputs and text outputs (cls, text) are often modeled differently and require different modules:
  - Object detector (OD)
  - Coordinate regression head
- A single model that unifies text and box(visual) outputs?

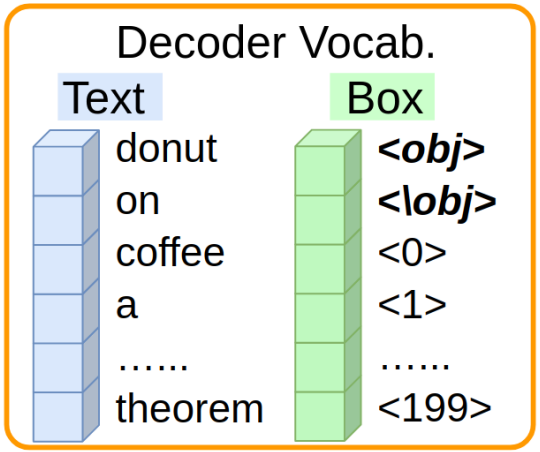
| Representative Models | Visual Modeling | Text Output | Box Output | Word-box Align |
|---|---|---|---|---|
| ViLBERT [42], OSCAR [38], UNITER [12], VinVL [73], *etc.* [37,59,35,58,78,43] | Offline OD | Task-specific Heads | Region Index | ✗ |
| PixelBERT [29], SOHO [28], ViLT [33], SimVLM [64], *etc.* [56,36,68,19] | Image Patches | Task-specific Heads | ✗ | ✗ |
| VL-T5 [13] | Offline OD | Single Output Seq. | Region Index | Extra Prediction |
| GPV [23] | Online OD | Single Output Seq. | Region Index | Extra Prediction |
| MDERT [31] | Image Patches | Task-specific Heads | Box Coordinate | ✗ |
| UniT [26] | Image Patches | Task-specific Heads | Box Coordinate | ✗ |
| UniTAB (Ours) | Image Patches | Single Output Seq. | Box Coordinate | Inline Indicated |

# UNICORN

- Textualize bounding box for object detection [1]
- Text and box outputs: Unified text+box decoding vocabulary
- Word-box alignments: in-line in output sequence



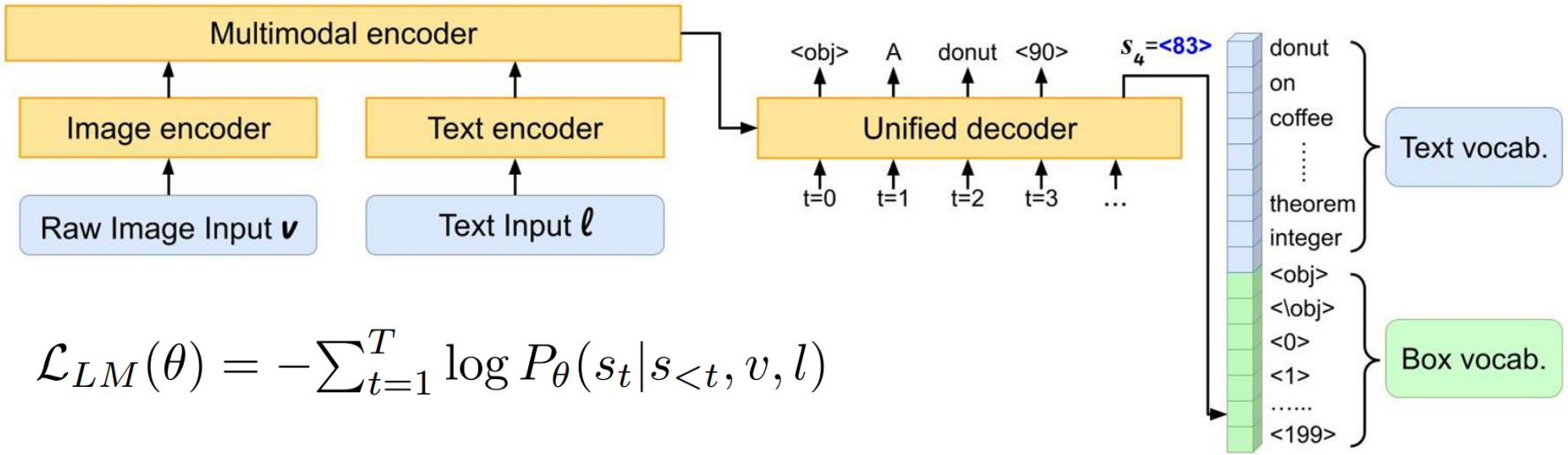

<obj> A donut <90> <83> <184> <180> <\obj> on a <obj> white plate <78> <84> <186> <199> <\obj> next to a <obj> cup of latte <10> <7> <94> <113> <\obj>.




UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling

[1] Pix2seq: A language modeling framework for object detection, ICLR 2022

# Model and Training




$$\mathcal{L}_{LM}(\theta) = -\sum_{t=1}^{T} \log P_\theta(s_t | s_{<t}, v, l)$$

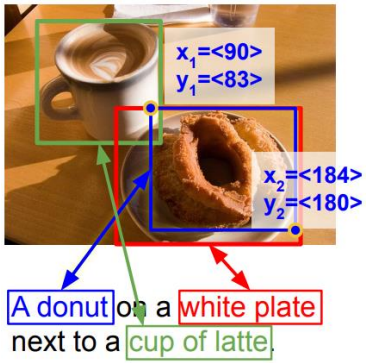
- Encoder-decoder architecture
- Single LM objective

<obj> A donut <90> <83> <184> <180> <\obj> on a <obj> white plate <78> <84> <186> <199> <\obj> next to a <obj> cup of latte <10> <7> <94> <113> <\obj>.

Image credit: UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling

# Unifying Different VL Tasks

- Textualized outputs: text, box, alignment
- Multi-task finetuning, capability generalization



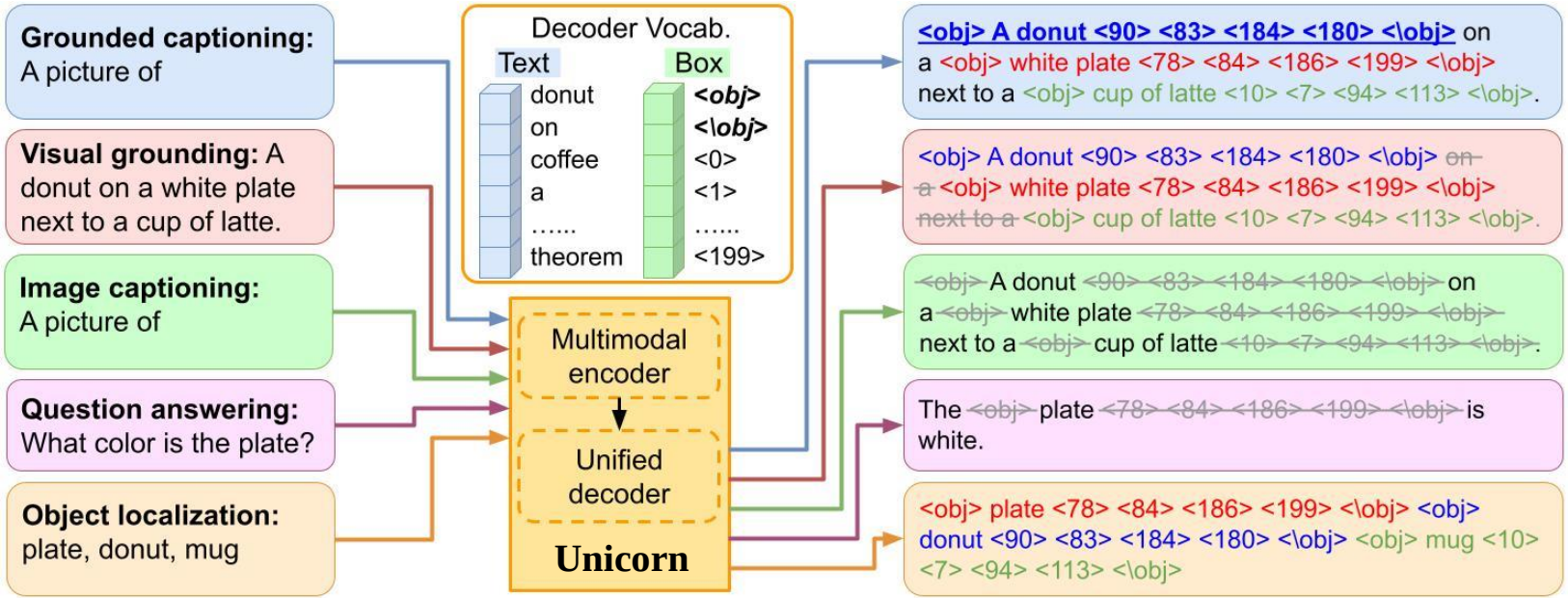
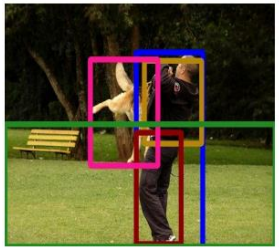




Image credit: UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling

# Capability Generalization

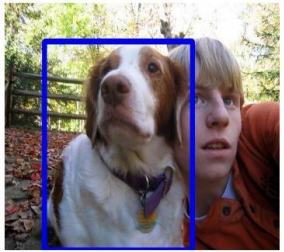
- MSCOCO; Grounded description
- For eval: text, box, alignment
- Metrics: captioning



**Input text:** A picture of
**Output seq.:** <obj> **A man** <97> <40> <146> <199> <\obj> **in** <obj> **a black jacket** <103> <47> <145> <115> <\obj> **and** <obj> **black pants** <97> <106> <130> <197> <\obj> **is playing with** <obj> **a dog** <63> <46> <113> <133> <\obj> **in** <obj> **a park** <0> <99> <199> <199> <\obj> .

**Text:** A man in a black jacket and black pants is playing with a dog in a park .
**Box:** Not Used

- ImageNet; Object localization
- For eval: text, box, alignment
- Metrics: accuracy

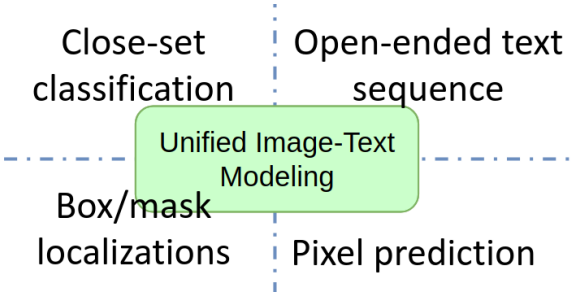


**Input text:** Brittany spaniel
**Output seq.:** <obj> **Brittany spaniel** <29> <31> <136> <199> <\obj>

**Text:** Not Used
**Box:** **Brittany spaniel**

Image credit: UNICORN. Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling

# Textualize Visual Outputs




- OFA
- Image tokens for pixel outputs

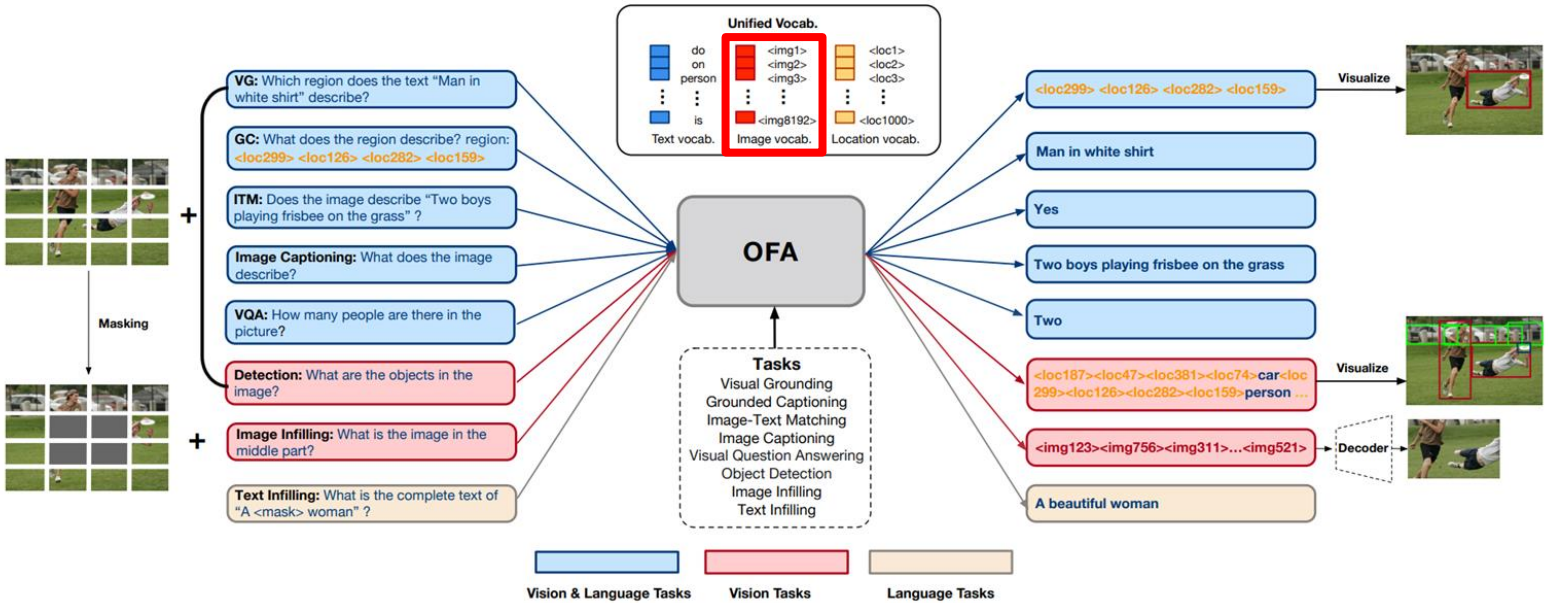



Image credit: OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, ICML 2022

# Textualize Visual Outputs

- Pix2Seq-V2
- Masks as polygon, Keypoints

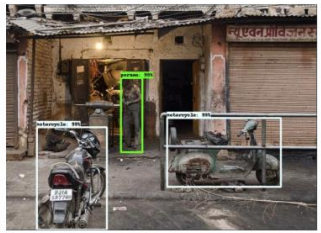
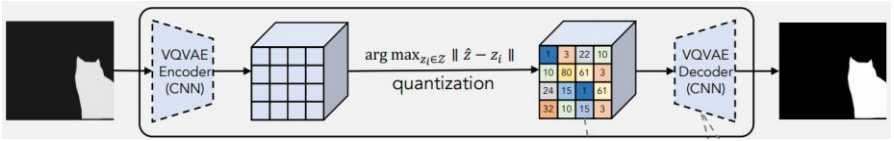
| Task prompt | Task output | Output visualization |
| --- | --- | --- |
| [Detect] | $y_{min}$=327 $x_{min}$=370 $y_{max}$=653 $x_{max}$=444 person ..... |  |
| [Segment] $y_{min}$=503 $x_{min}$=518 $y_{max}$=805 $y_{max}$=892 Motocycle | $y_0$=553 $x_0$=599 $y_1$=788 $y_1$=664 ..... |  |
| [Keypoint] $y_{min}$=327 $x_{min}$=370 $y_{max}$=653 $x_{max}$=444 person ..... | Nose $y_{min}$=1 $x_{min}$=57 left eye ..... |  |
| [Describe] | A person working in mechanical shop with two mopeds outside. | |

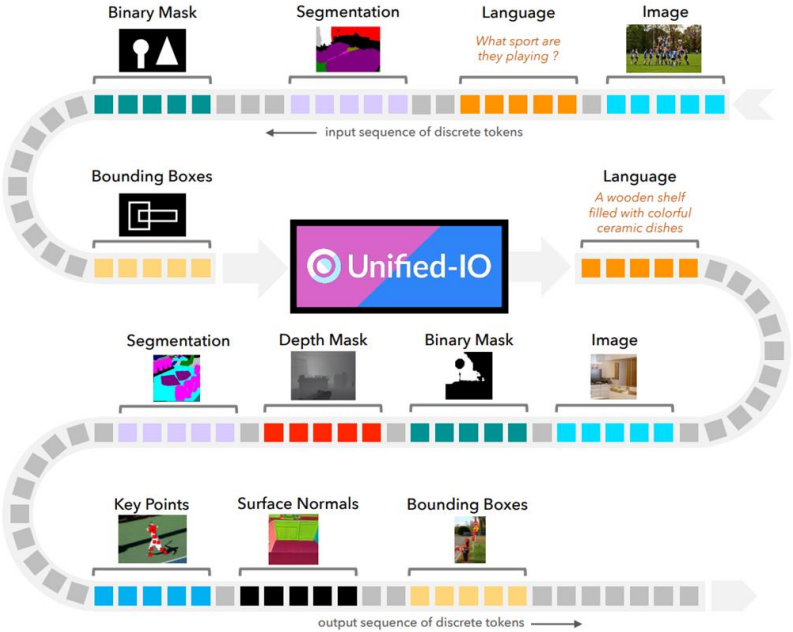
Image credit: A Unified Sequence Interface for Vision Tasks

# Textualize Visual Outputs




- Unified-IO

all vision tasks (seg., depth, surface, etc.) that require dense prediction as (conditional) image generation




**Tasks**

Image Classification
Object Detection
Semantic Segmentation
Depth Estimation
Surface Normal Estimation
Segment-based Image Generation
Image Inpainting
Pose Estimation
Relationship Detection
Image Captioning
Visual QA
Referring Expressions
Situation Recognition
Text-based Image Generation
Visual Commonsense
Classification in context
Region Captioning
GLUE Benchmark tasks
Reading comprehension
Natural Language Inference
Grounded Commonsense Inference

Image credit: Unified-IO: Sequential Modeling for Generally Applicable Vision Models

# Agenda

- Image-text tasks overview; Motivation
- Unified image-text models
- Summary and discussion

## Take-away Messages

- Unified image-text modeling from the view of I/O format
- Textualized visual outputs for unified image-text modeling
- Format unification is the first step, improving unified models
- Grand vision of general-purpose visual understanding

## Challenges and Future Directions

- How to better show the advantage of unified models
  - Relationship among tasks; Gain from MTL
  - New capability showcase; Generalization setups
- How to better train the unified models
  - Balance the degree of unification
  - Better ways of format unification
- Foundation models at what granularity

Thank you!
Any Questions?