Deep Learning for Vision and Language

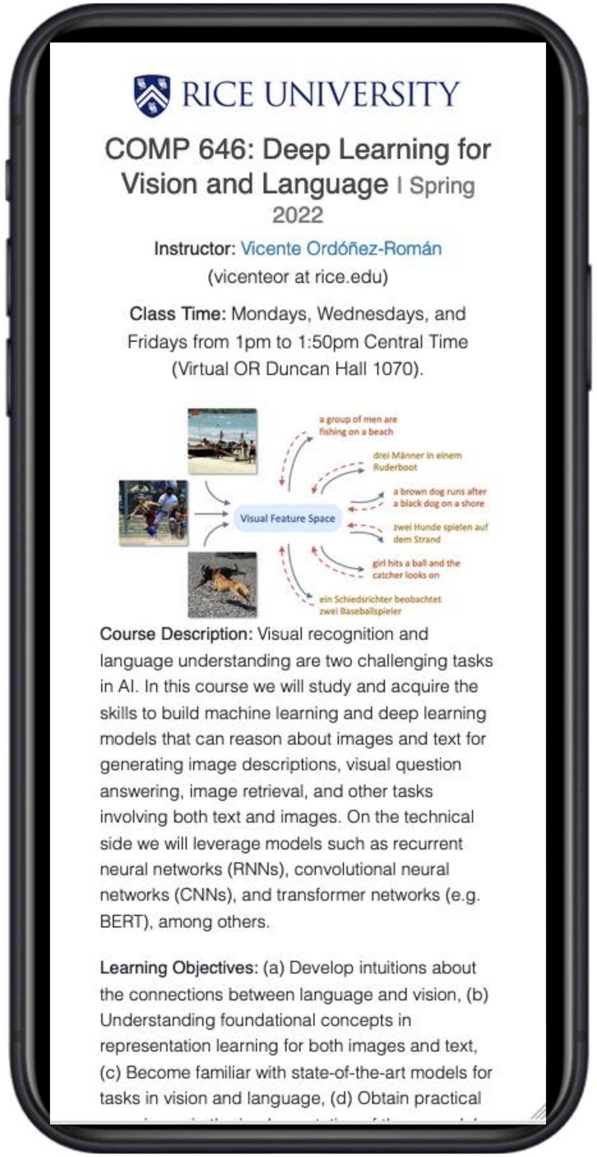
Welcome and Introduction

RICE UNIVERSITY

About the class

- COMP 646: Deep Learning for Vision and Language
- Instructor: **Vicente** Ordóñez (Vicente Ordóñez Román)
- Website: https://www.cs.rice.edu/~vo9/deep-vislang
- Location: Keck Hall 100
- Times: Tuesdays and Thursdays
  from 4pm to 5:15pm
- Office Hours: TBD (Duncan Hall 2080)
- Teaching Assistants: Jaywon Koo, Catherine He, Ruidi Chang, Jason Uwaeze
- Discussion Forum: TBD

https://www.cs.rice.edu/~vo9/deep-vislang/



RICE UNIVERSITY

COMP 646: Deep Learning for Vision and Language | Spring 2025

Instructor: Vicente Ordóñez-Román (vicenteor at rice.edu), Office Hours: TBD.
TA: Jaywon Koo (jk125 at rice.edu), Office Hours: TBD.
TA: Catherine He (ch151 at rice.edu), Office Hours: TBD.
TA: Jason Uwaeze (ju6 at rice.edu), Office Hours: TBD.
TA: Ruidi Chang (rc151 at rice.edu), Office Hours: TBD.
Class Time: Tuesdays and Thursdays from 4pm to 5:15pm Central Time. Location: Keck Hall 100.

Course Description: Visual recognition and language understanding are two fundamental tasks in the quest toward Artificial Intelligence. In this course we will study and acquire the skills to build machine learning and deep learning models that can reason about images and text for generating image descriptions, find objects in images, generating images from text, image generation and synthesis, and other general tasks involving both text and images. On the technical side we will leverage models such as convolutional neural networks (CNNs), Transformer networks (e.g. BERT, LLama, ViTs), Generative Models (e.g Latent Diffusion, DiTs, VAEs), among others. Emphasis will also be placed on re-using multimodal foundation models such as CLIP, SDXL, LLaMA-3, etc.

Learning Objectives: (a) Develop intuitions about the connections between language and vision, (b) Understand concepts in representation learning for both images and text, (c) Become familiar with state-of-the-art models for tasks in vision and language, (d) Obtain practical experience in the implementation and adaptation of these models.

Prerrequisites: There are no formal strict pre-requisities for this class. We will review basics of machine learning at the beginning of this class. Students however should have a basic knowledge of linear algebra, differential calculus, and basic statistics and probability. Moreover students are expected to have attained some level of proficiency in Python programming or be willing to learn Python programming. Students are encouraged to complete the following activity before the first lecture: [Primer on Image Processing].

Schedule

Date	Topic
Tue, Jan 14	Introduction to vision and language #welcome
Thu, Jan 16	Supervised vs unsupervised learning and linear classifiers #machine-learning
Tue, Jan 21	Stochastic Gradient Descent / Regularization / Softmax #machine-learning
Thu, Jan 23	Multi-layer Perceptrons and Backpropagation #machine-learning
Tue, Jan 28	The Convolutional Operator, Image Filtering and Convolutional Neural Networks #computer-vision
Thu, Jan 30	Convolutional Neural Network Architectures: LeNet, AlexNet, VGG, InceptionNets, ResNets #computer-vision
Tue, Feb 4	Introduction: Bag of Words, Language Models, Word Embeddings #natural-language-processing
Thu, Feb 6	Recurrent Neural Networks and Sequence-to-Sequence Models #natural-language-processing
Tue, Feb 11	#Guest-Lecture: TBD

About me -- Vicente

Associate Professor, 2021 - Present	RICE UNIVERSITY
Visiting Academic 2021 - 2025	amazon
Assistant Professor, 2016 - 2021	UNIVERSITY of VIRGINIA
Visiting Professor, 2019	Adobe Research
Visiting Researcher, 2015 - 2016	AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
MS, PhD in CS, 2009-2015	THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL; Stony Brook University

… also spent time at:

Google Microsoft ebay

What is Vision and Language?

Anything at the intersection of Computer Vision and Natural Language Processing. Systems and models that depend a little bit on both.

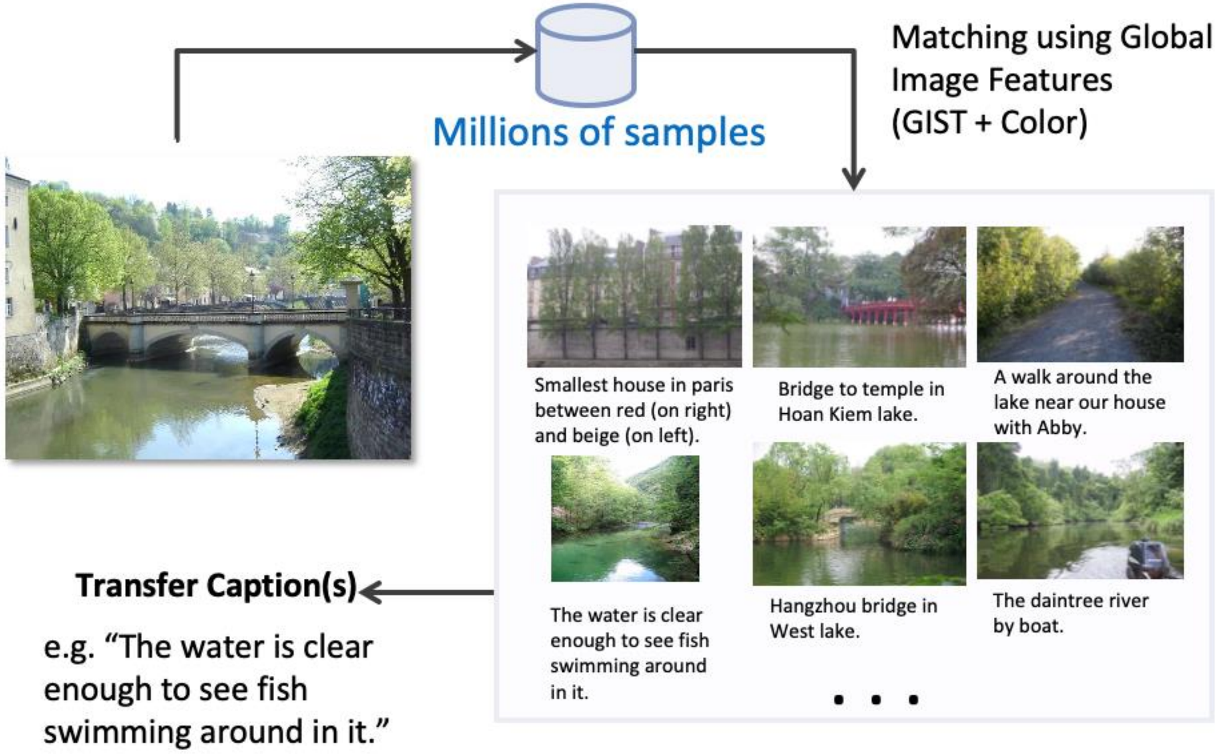
- Computer Vision: How do we teach machines to process, represent and understand images? e.g. to recognize objects in images.

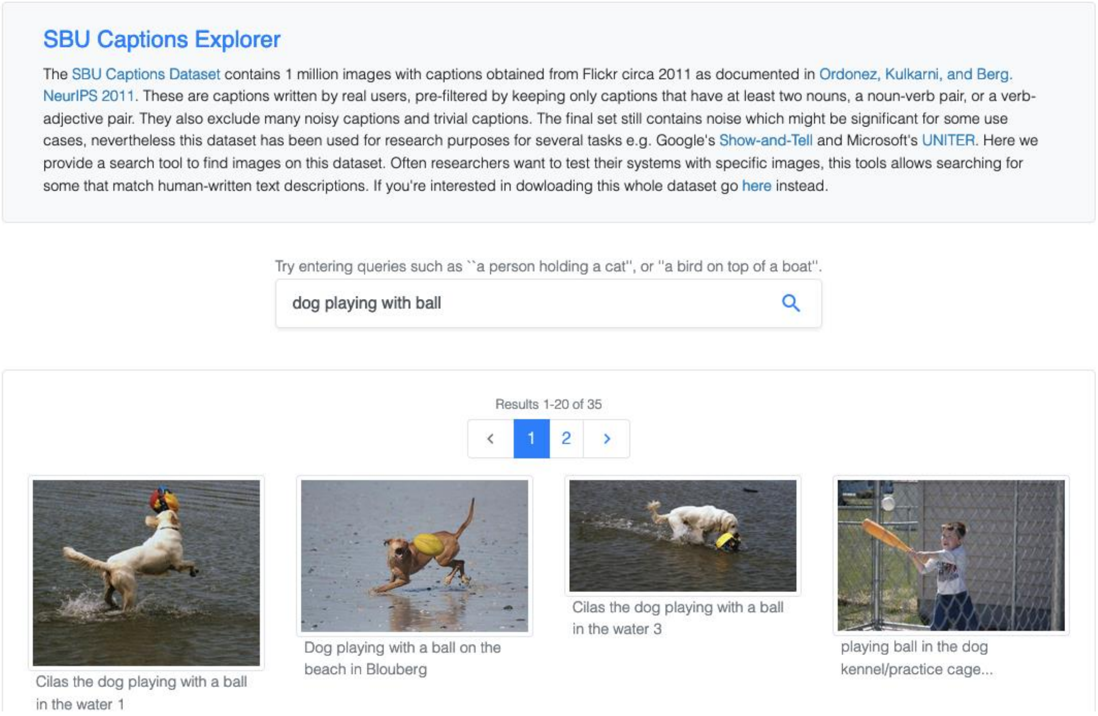
- Natural Language Processing: How do we teach machines to process, represent and understand text? e.g. to classify or generate text.

Some of our work includes…

Describing images with language

https://vislang.ai/sbu-explorer

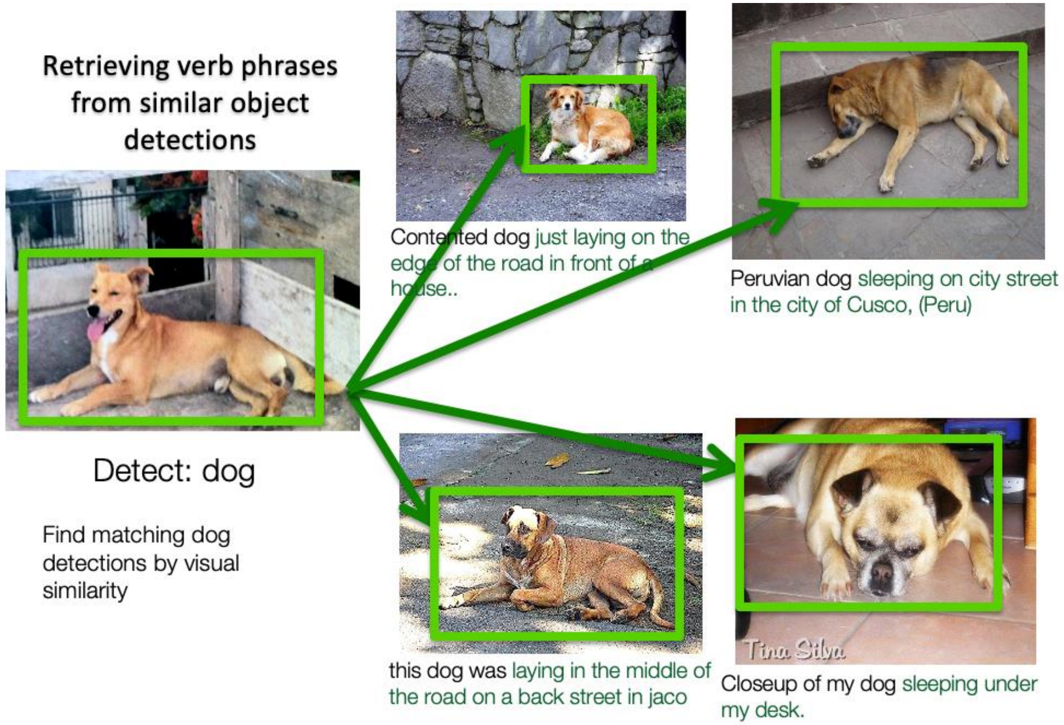




Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara L. Berg.
Advances in Neural Information Processing Systems. **NIPS 2011**. Granada, Spain. December 2011.

Some of our work includes…

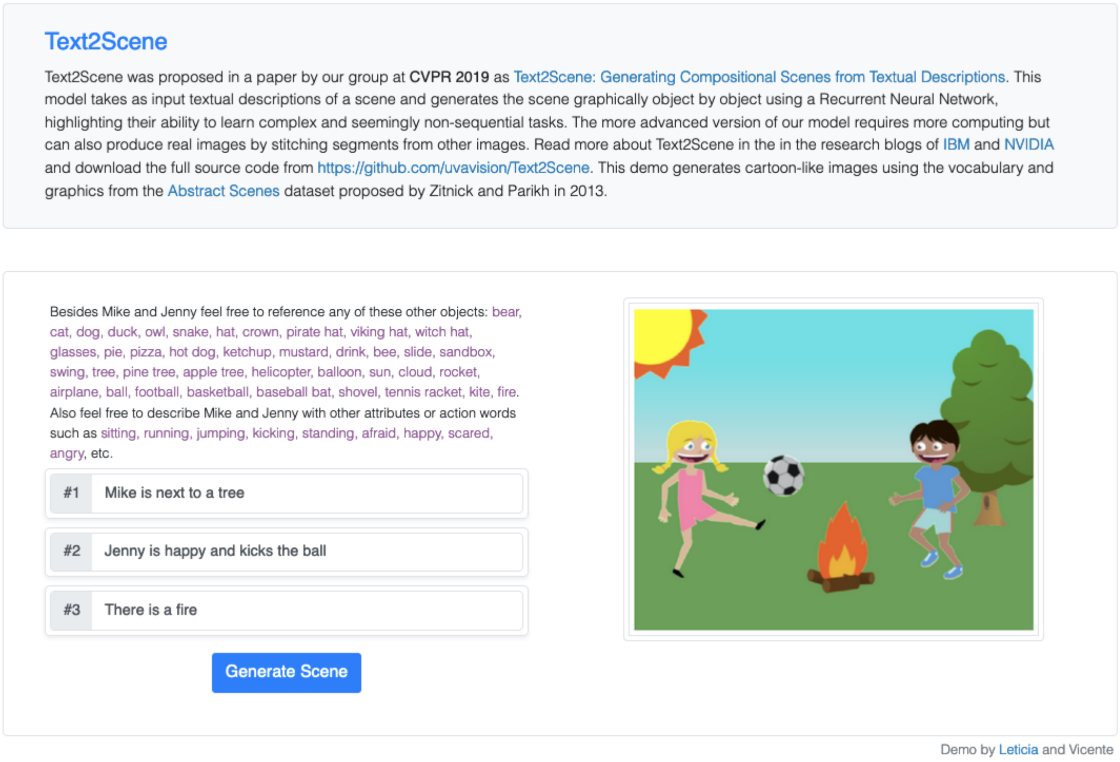
Describing images with language



Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg.
International Journal of Computer Vision. **IJCV 2015**. [August 2016 Issue]. [pdf] [link] [bibtex]

Describing language with images

https://vislang.ai/text2scene



Some of our work includes…

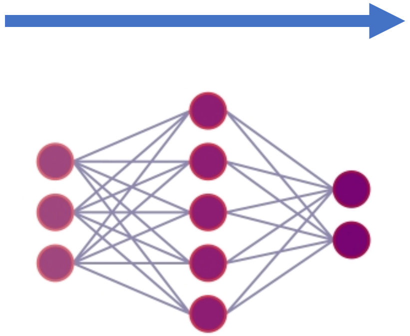
Interactive Image Retrieval

[Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries]
Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, Vicente Ordonez. Conf. on Neural Information Processing Systems. **NeurIPS 2019**. Vancouver, Canada. December 2019. [arxiv] [code] [bibtex]

Visual Grounding: Identifying Any Object in Images

Input Image + Text

A picture of a cathedral next to a park

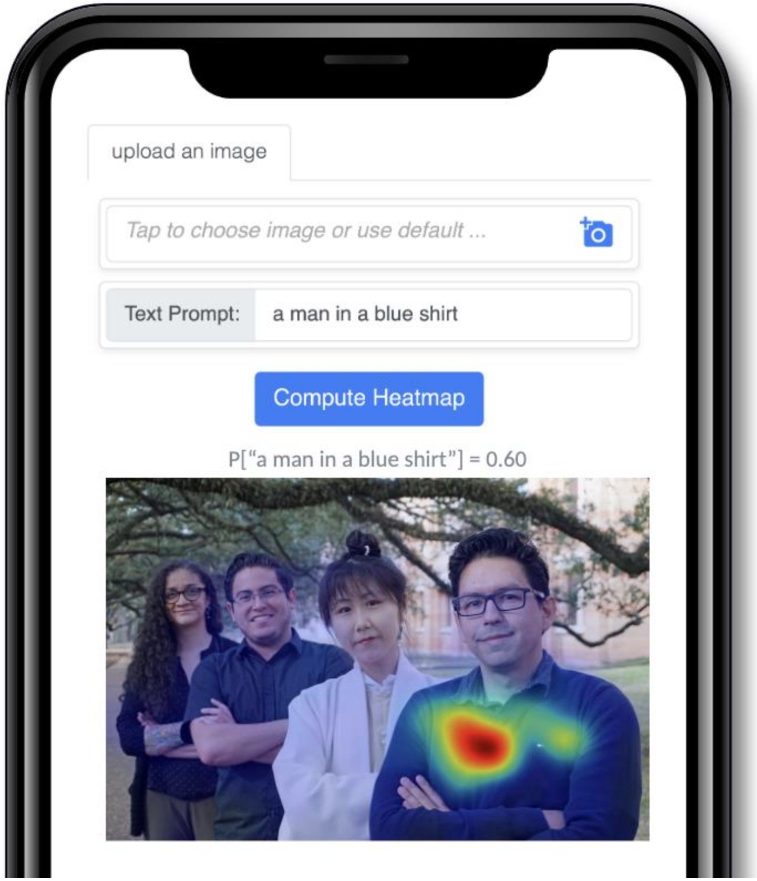


w/ Attention Mask Consistency (Ours)

A picture of a cathedral next to a park

Try our Demo

https://vislang.ai/amc



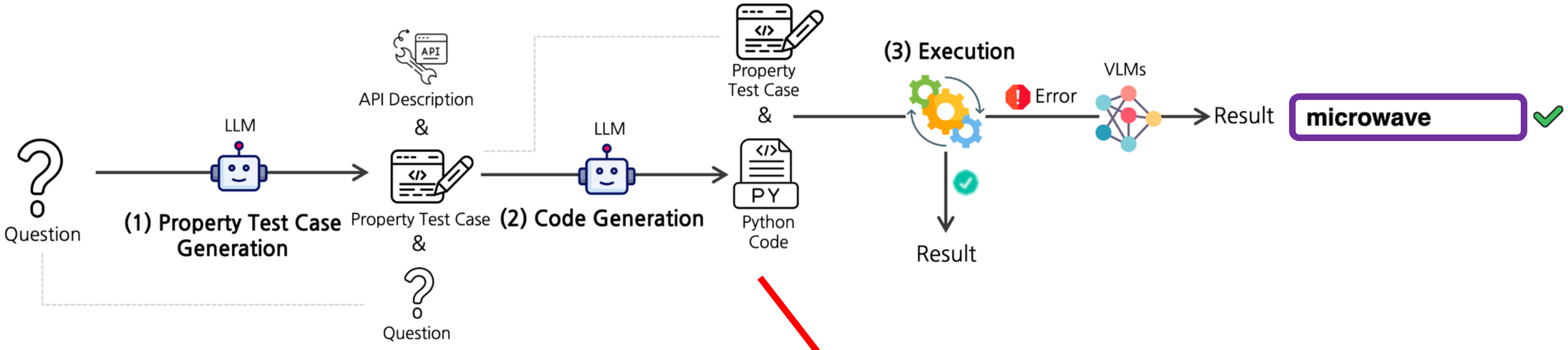
[Improving Visual Grounding by Encouraging Consistent Gradient-based Explanations.](#)
Ziyan Yang, Kushal Kafle, Franck Dernoncourt, Vicente Ordonez. Conf. on Computer Vision and Pattern Recognition. **CVPR 2023**. Vancouver, Canada.

[Improved Visual Grounding through Self-Consistent Explanations](#)
Ruozhen He, Paola Cascante-Bonilla, Ziyan Yang, Alexander C. Berg, Vicente Ordonez
Conf. on Computer Vision and Pattern Recognition. **CVPR 2024**. Seattle, WA.

Probing LLMs for Visual Reasoning

[PropTest: Automatic Property Testing for Improved Visual Programming](#)
Jaywon Koo, Ziyan Yang, Paola Cascante-Bonilla, Baishakhi Ray, Vicente Ordonez. Conf. on Empirical Methods in Natural Language Processing. **EMNLP 2024** (Findings).

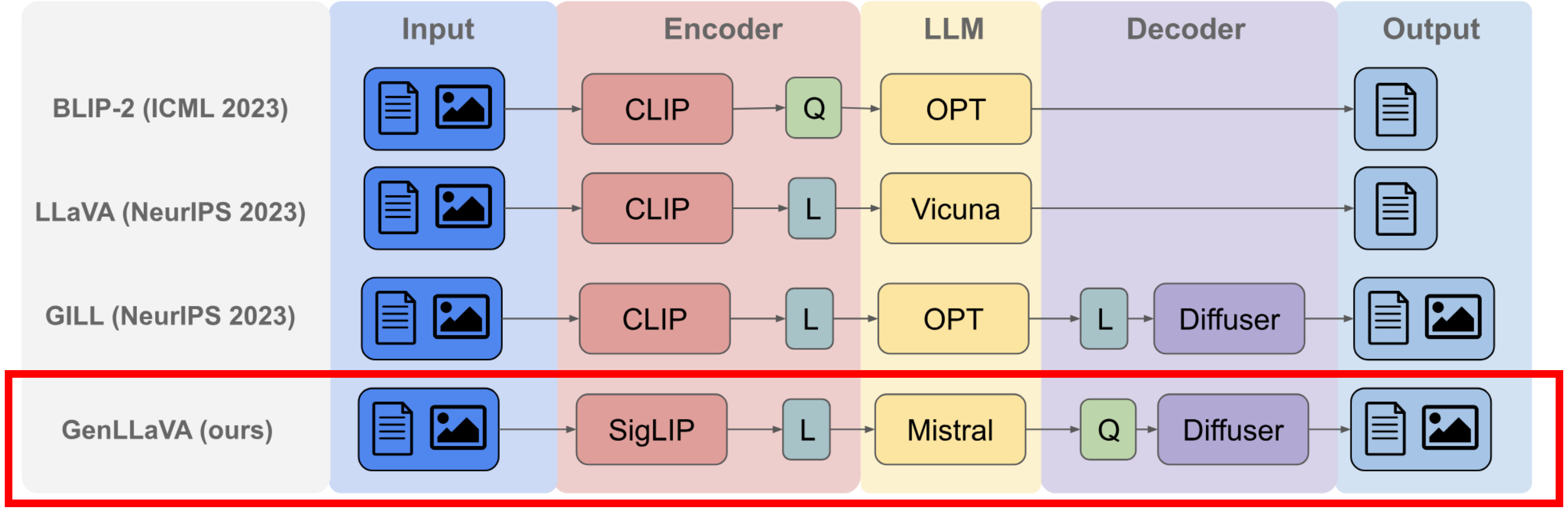
Question: What appliance is above the bananas?



Generated Code

```
def execute_command(image, my_fig, time_wait_between_lines, syntax):
    image_patch = ImagePatch(image)
    banana_patches = image_patch.find("banana")
    if len(banana_patches) == 0:
        return image_patch.simple_query("What appliance is above the banana?")
    for banana_patch in banana_patches:
        appliance_patches = image_patch.find("appliance")
        for appliance_patch in appliance_patches:
            if appliance_patch.vertical_center > banana_patch.vertical_center:
                # crop would include the appliance so keep it in the query
                return appliance_patch.simple_query("What is the name of the appliance?")
    # If no appliance is above the banana, pick the first appliance
    return appliance_patches[0].simple_query ("What is the name of the appliance?")
```

Combining LLMs with Vision Models: GenLLaVA



[Generative Visual Instruction Tuning]
Jefferson Hernandez, Ruben Villegas, Vicente Ordonez. arXiv:2406.11262
June 2024.

Generative AI: Text-to-Image and Text-to-Video

A tiger in a lab coat with a 1980s Miami vibe, turning a well oiled science content machine, digital art.

[ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation](https://elasticdiffusion.github.io/)
Moayed Haji Ali, Guha Balakrishnan, Vicente Ordonez Conf. on Computer Vision and Pattern Recognition. **CVPR 2024**. Seattle, WA.

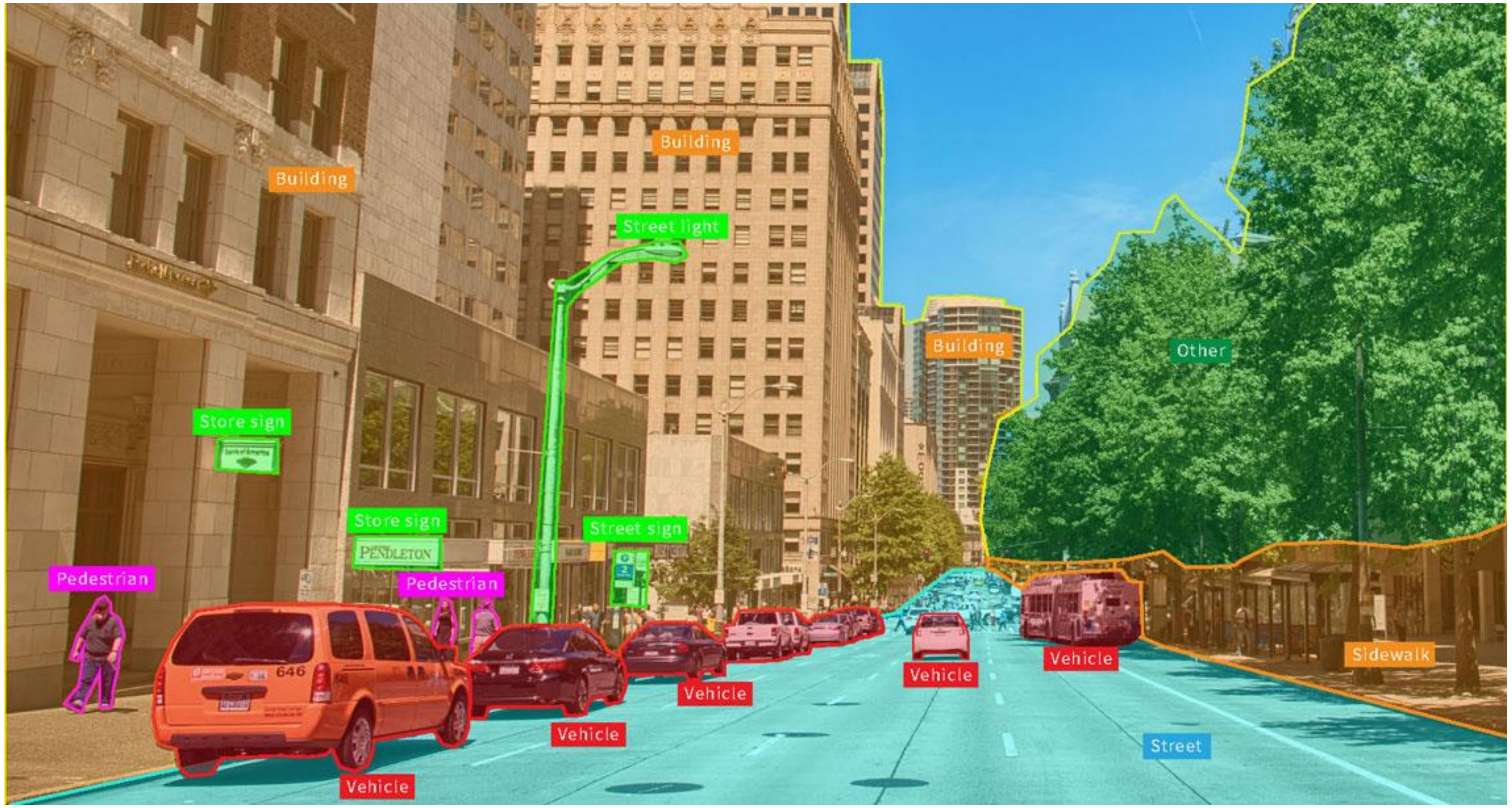
https://elasticdiffusion.github.io/

Why Vision and Language Together?

- What makes us intelligent?

Why Vision and Language Together?

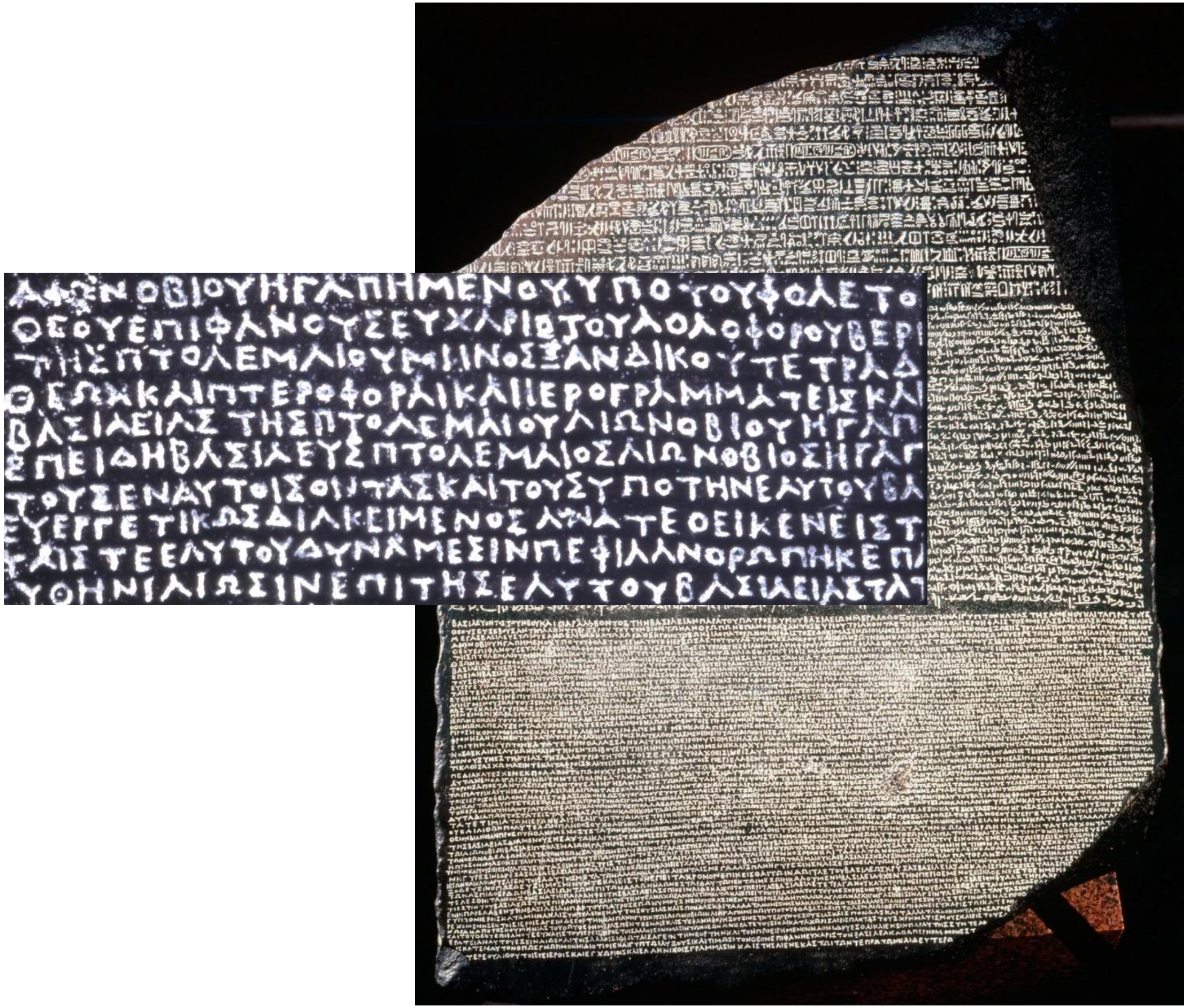
- What makes us intelligent?
- Vision is not just sensing – but interpreting what our eyes capture



https://appen.com/blog/computer-vision-vs-machine-vision/

Why Vision and Language Together?

- What makes us intelligent?
- Vision is not just sensing – but interpreting what our eyes capture
- Language is not just a sequence of symbols – but interpreting what do they mean – think of a foreign language to you



Can we learn language through pictures?

https://www.hameraypublishing.com/blogs/all/teaching-kids-about-the-structure-of-the-spanish-language

Vision and Language in Practice

- Searching products using language can be hard – e.g. I want to find a "rustic vintage curio with dark cherry finishes"

Vision and Language in Practice

- Robotics: Instruction Following

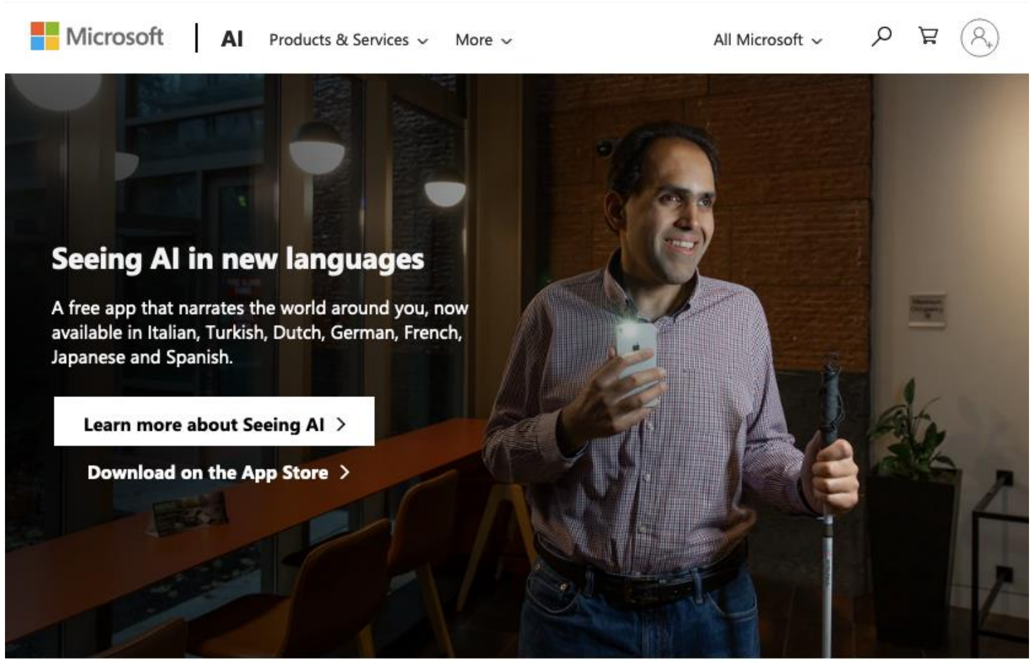
Amazon launches home robot Astro and giant Alexa display

Robot that can check on loved ones and pets is one of plethora of devices announced at big launch event

Astro is Amazon's first attempt at a home robot designed to be a roving smart platform for Alexa, video calling and many other services. Photograph: Amazon

Vision and Language in Practice

- Assistive Technologies



Complete multiple tasks with one app

Switch between channels to tune the description of what's in front of the camera.

Scene

An experimental feature to describe the scene around you

Color

Describes the perceived color

Generative for Advertising and Marketing

Amazon Ads introduces AI-powered video generation and live image capabilities to help brands deliver compelling creative for customers

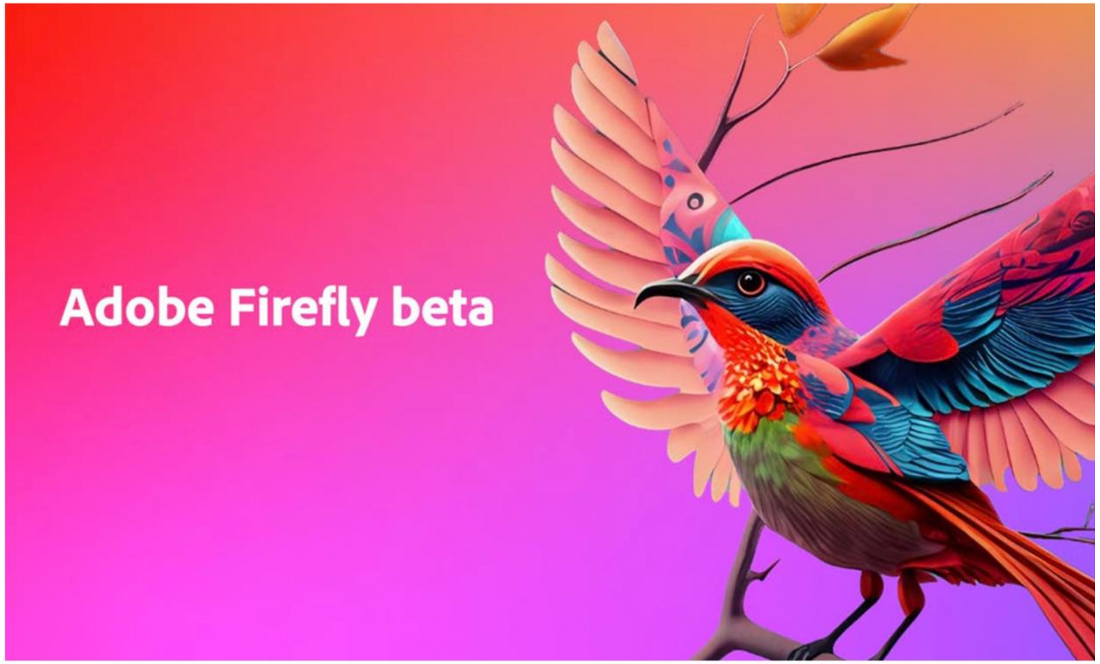
September 19, 2024 | By Matt Miller, Senior Copywriter

Image and Media Editing and Generation

Adobe Research is helping shape the future of generative AI for creative expression with Firefly

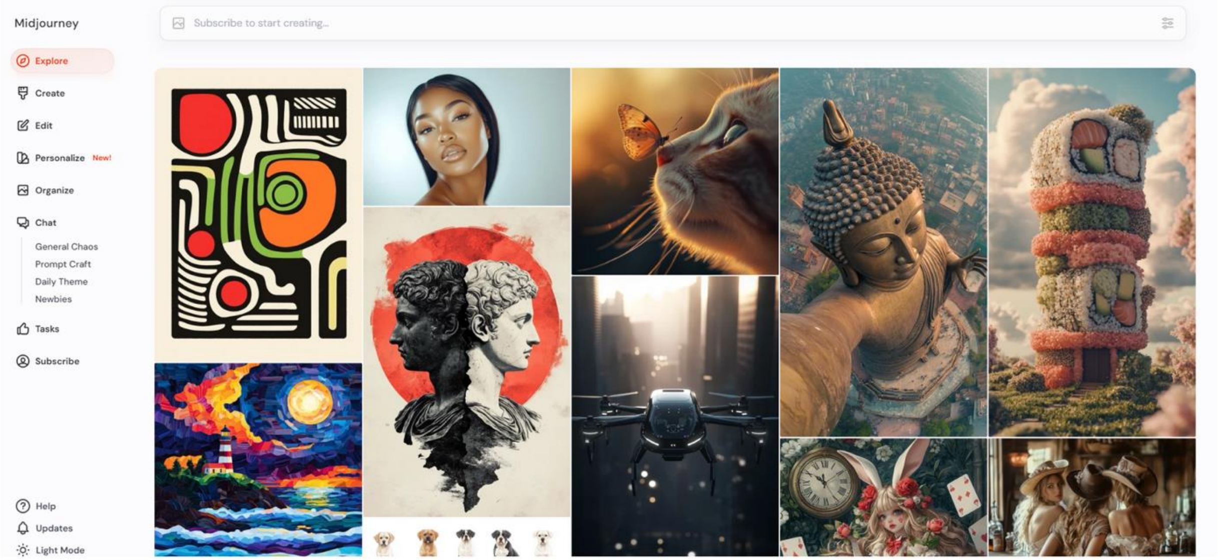
May 30, 2023

Tags: AI & Machine Learning, Computer Vision, Imaging & Video, Graphics (2D & 3D)



Adobe has launched Firefly, a new family of generative AI models for creative expression. The technology is now available in and in addition, Firefly is now powering a new tool in the Photoshop desktop beta app, Generative Fill.

With Firefly, users can bring a creative vision to life by describing it in their own words. They can instantly iterate though vari amplifying their creativity and their imaginations. Work by Adobe Research has helped to power the technology inside Firefly, and the research team is currently developing features for Firefly's next phase.



Midjourney V7: What We Know So Far

A Complete Rebuild

Midjourney V7 is more than just an update; it's a complete overhaul of the AI system. The developers are rebuilding the model from scratch, using a new architecture and fresh datasets. This suggests significant changes are coming, potentially leading to a major leap in image quality and the AI's ability to understand your creative vision.

What will we cover in this class?

In terms of tools

Duration	Topics
3 weeks	Introduction to ML / Vision / NLP; Neural Networks (NNs) / Deep Learning.
2 weeks	Convolutional Neural Networks (CNNs)
2 weeks	Recurrent Neural Networks (RNNs, LSTMs, GRUs)
8 weeks	Transformers (e.g. BERT, GPT, FLAN-T5, LLaMA, etc); Diffusion Models (e.g. Stable Diffusion, ControlNet); State-of-the-art and Recent Developments

What will we cover in this class?

In terms of topics

- Image Captioning
- Referring Expression Comprehension
- Visually-grounded Question Answering
- Learning from Text and Images
- Retrieving Images from Natural Language Queries
- Generating Images from Text
- Multimodal Translation using both Images and Text
- Vision-Language Navigation
- Biases in Vision and Language Tasks
- Possibly more topics…

https://www.cs.rice.edu/~vo9/deep-vislang/

Pre-requisites

- No formal pre-requisites but…
- You need to know how to program with Python or be VERY motivated to learn as you go. Definitely know how to program at a college graduate level.
- You will benefit from knowing some Machine Learning or be VERY motivated to do some self-learning as you go.
- You need to be proficient on basic calculus, linear algebra, and statistics. Nothing advanced but the right basic terminology and concepts are needed. (matrices, vectors, vector spaces, chain rule of calculus, derivatives, gradients, maximum likelihood estimation, least squares regression)

Grading for this class: COMP 646

- Assignments: 30pts (3 assignments: 10pts + 10pts + 10pts)
- **Class Project: 60pts**
- Quiz: 10pts

Total: 100pts

- Grade cutoffs: TBD but no harsher than those indicated in our syllabus.

Class Project Timeline

- Class Project: 60pts
  - You can form a group: 3 students maximum per group
  - You can also work solo – 1 student groups.
- In ~4 weeks: Submit as a group a project proposal (1 page PDF)
- In ~6 weeks: Submit as a group a final project proposal (1 page PDF)
- In ~10 weeks: Submit a project progress report (2 page PDF)
- End of semester: Submit the following:
  - Project report PDF (4 pages)
  - Slides + Presentation (Video / Demo)
  - Source code + ideally an online demo (if appropriate)

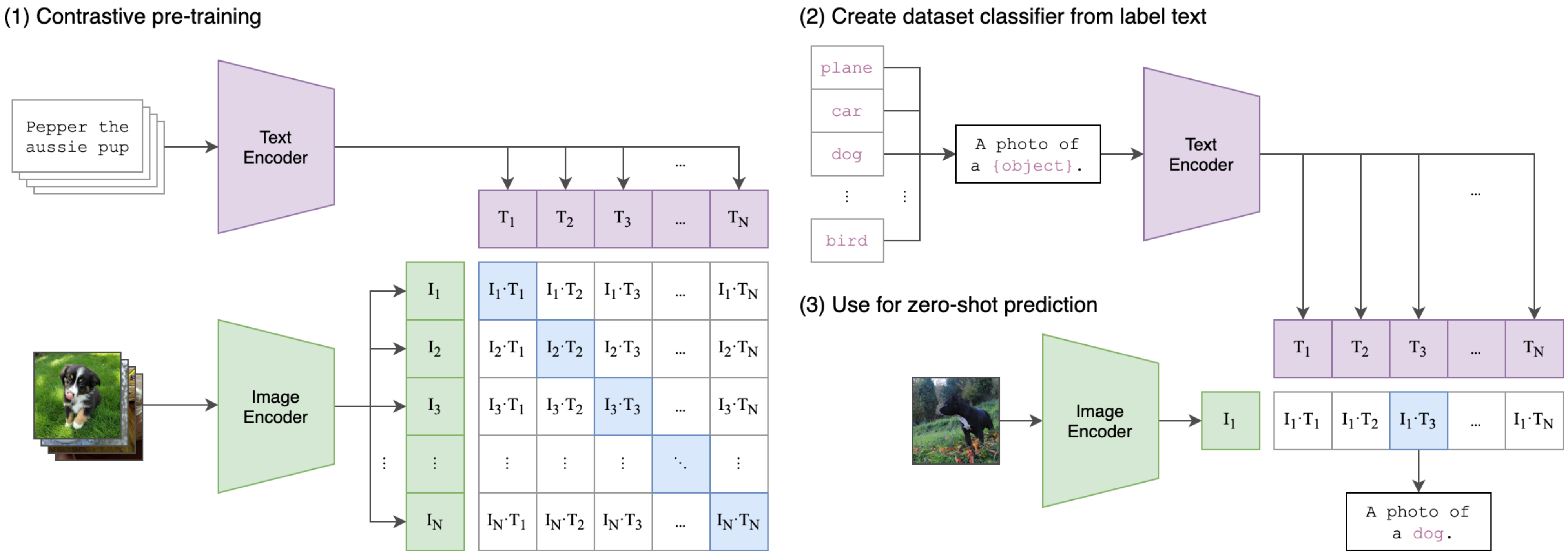
Project Requirements

- The project has to be at least as challenging as the most challenging assignment on this class. Do not work on something too trivial or too demanding for a class project. I can provide guidance if you submit on time a high quality project proposal.
- All projects need to involve both vision and language capabilities. If you are interested in a project that only involves image classification, object detection, etc. This is not the right class. You should take:
  - COMP 447: Computer Vision (Guha Balakrishnan, Spring 2025)
- If you are interested in a project that only involves text, e.g. sentiment analysis, building a chatbot, etc. This is not the right class. You should take:
  - COMP 652: Natural Language Processing (Hanjie Chen, Spring 2025)

You are encouraged to take advantage of recent open foundation models for your project.

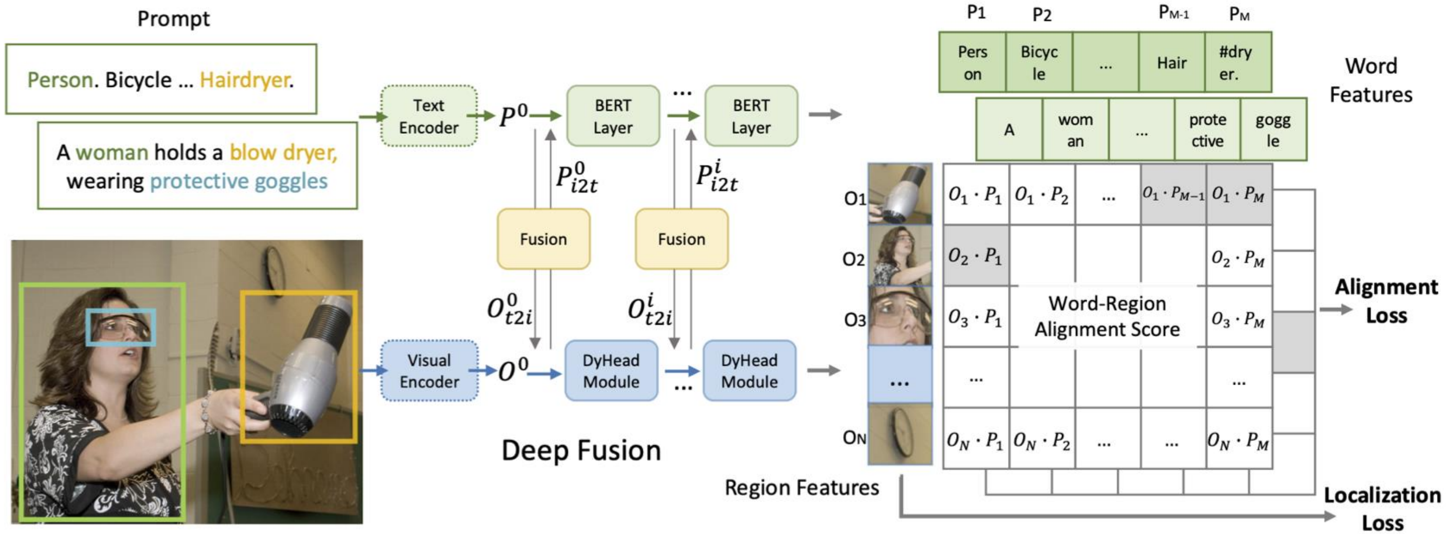
- CLIP by OpenAI or SigLIP or OpenCLIP or MetaCLIP (Images + Text)
- InternVL (Images + Text) or InternVideo (Images + Video)
- FLAN-T5 by Google (Text)
- Llama-2 or Llama-3 by Meta or Mistral-7B by MistralAI (Text)
- GLIP by Microsoft, or OwL-ViT (Images + Text)
- Whisper by OpenAI (Speech to Text)
- StableDiffusion v1.5 (or SDXL) models (Text to Images)
- LLaVa v1.5 (Multimodal LLM)

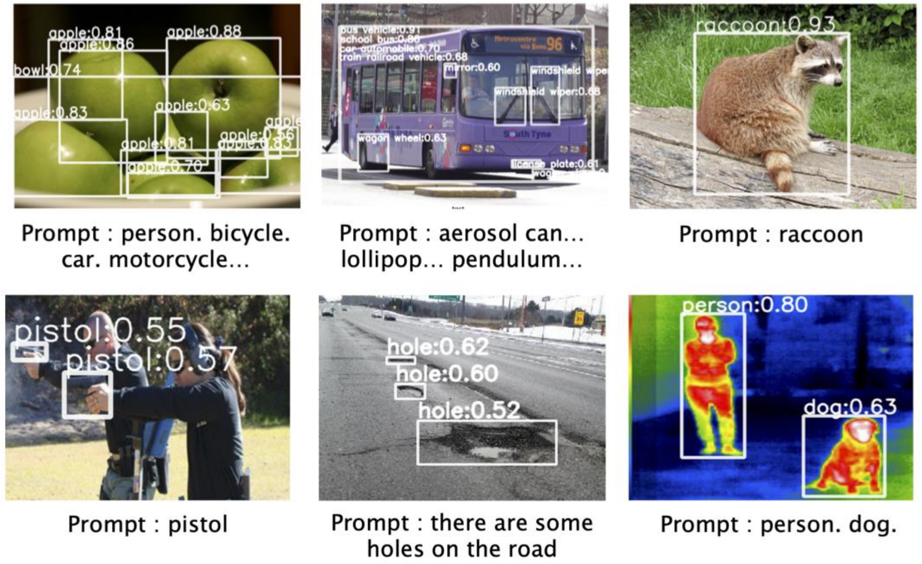
CLIP



https://github.com/openai/CLIP

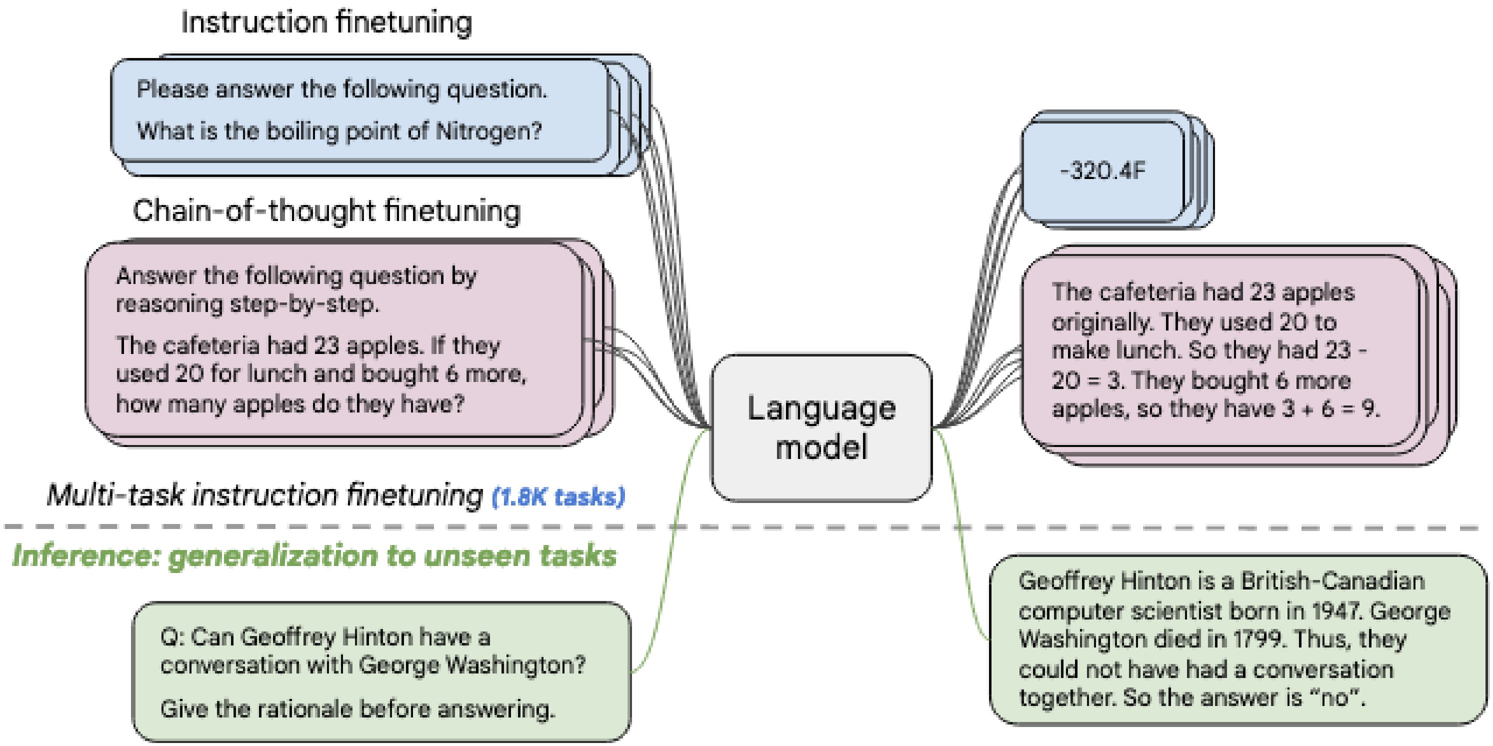
GLIP





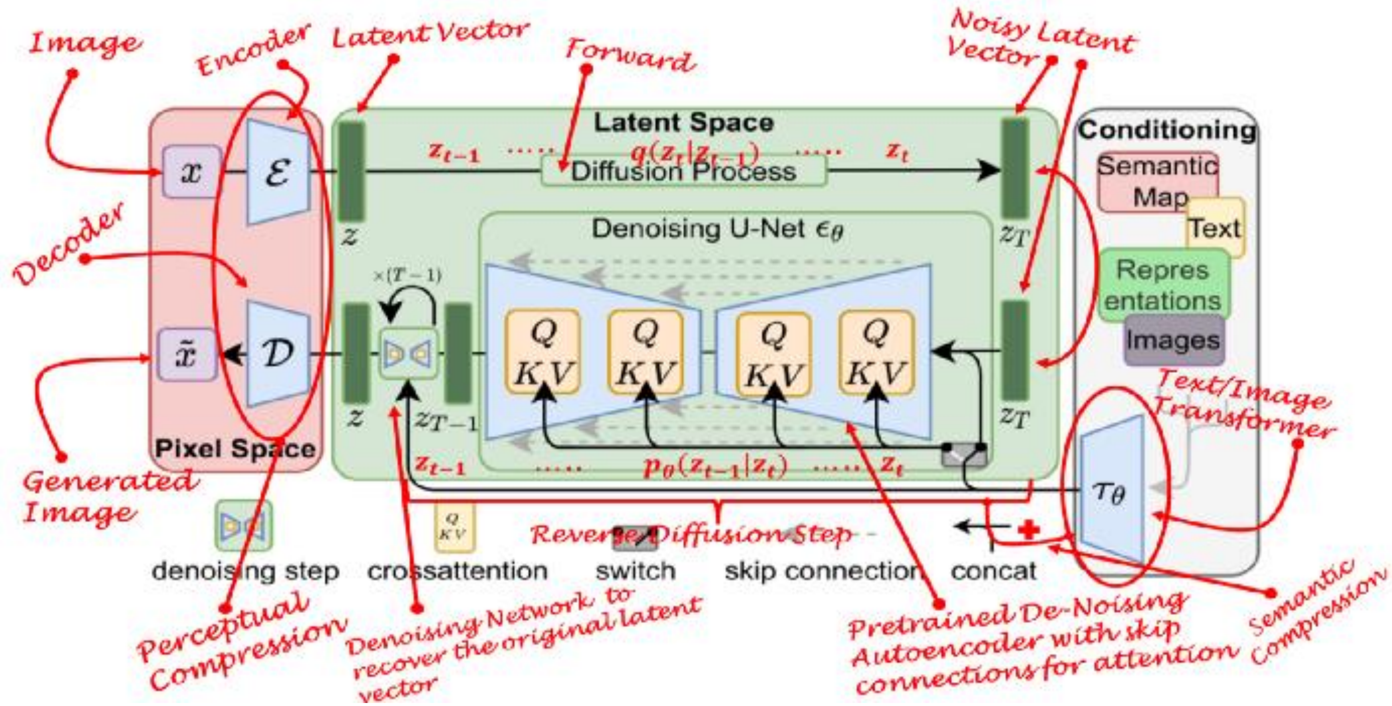
https://github.com/microsoft/GLIP

FLAN T5



https://huggingface.co/docs/transformers/model_doc/flan-t5

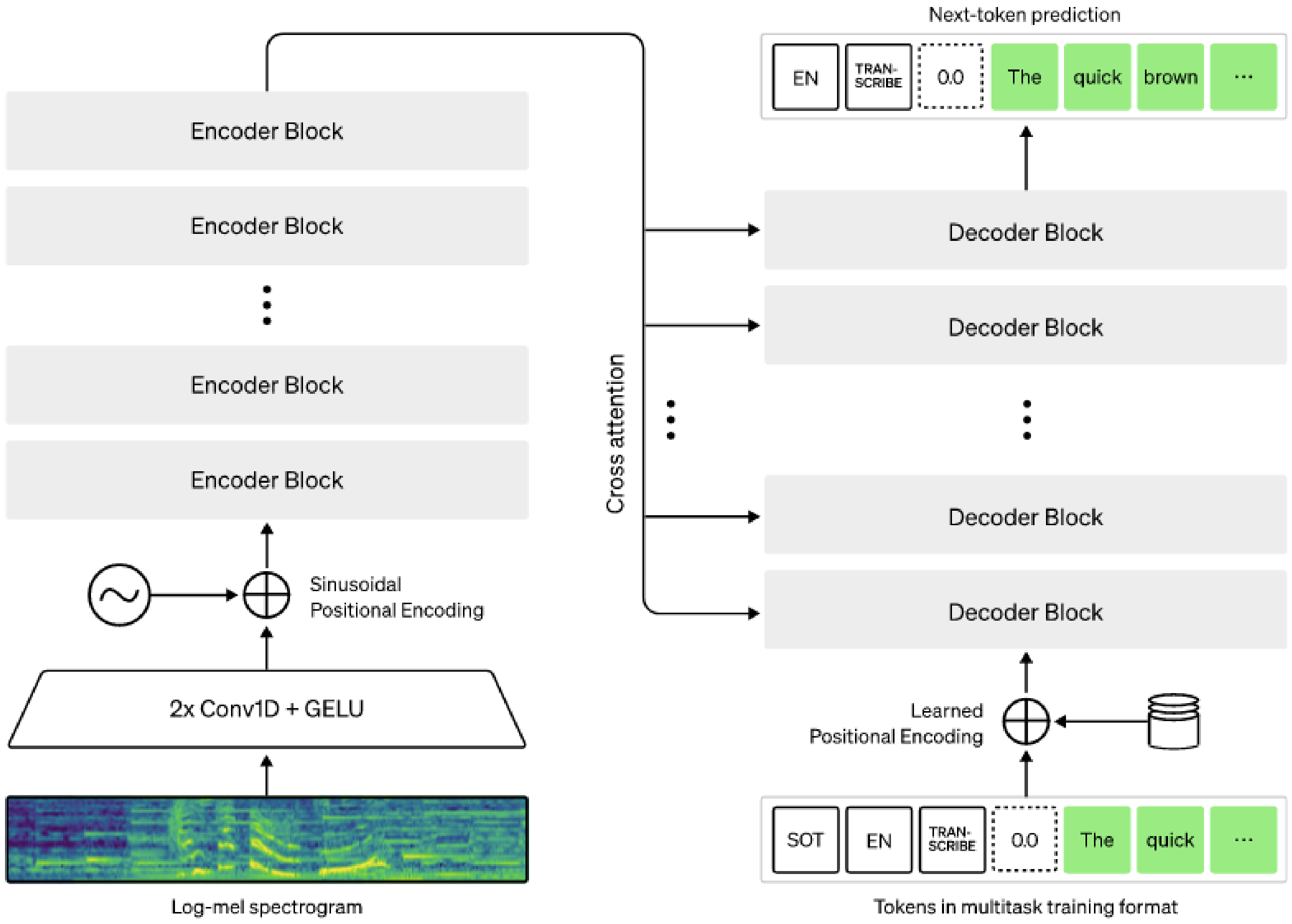
Stable Diffusion v2



https://towardsdatascience.com/what-are-stable-diffusion-models-and-why-are-they-a-step-forward-for-image-generation-aa1182801d46

https://huggingface.co/stabilityai/stable-diffusion-2

Whisper



https://github.com/openai/whisper

We will be using

PyTorch

https://pytorch.org/

HUGGING FACE

https://huggingface.co/

But you’re free to use any other framework especially for your projects: e.g. Tensorflow, Apache MXNet, JAX

We will also be using…

colab

https://colab.research.google.com/

Choose the Colab plan that's right for you

Whether you're a student, a hobbyist, or a ML researcher, Colab has you covered

Colab is always free of charge to use, but as your computing needs grow there are paid options to meet them.

Restrictions apply, learn more here

Pay As You Go

$9.99 for 100 Compute Units

$49.99 for 500 Compute Units

You currently have 151.52 compute units.

Compute units expire after 90 days. Purchase more as you need them.

- No subscription required.
  Only pay for what you use.
- Faster GPUs
  Upgrade to more powerful GPUs.

Colab Pro

$9.99 per month

Current plan

- 100 compute units per month
  Compute units expire after 90 days. Purchase more as you need them.
- Faster GPUs
  Upgrade to more powerful GPUs.
- More memory
  Access our highest memory machines.
- Terminal
  Ability to use a terminal with the connected VM.

Colab Pro+

$49.99 per month

All of the benefits of Pro, plus:

- An additional 400 compute units for a total of 500 per month.
  Compute units expire after 90 days. Purchase more as you need them.
- Faster GPUs
  Priority access to upgrade to more powerful premium GPUs.
- Background execution
  With compute units, your actively running notebook will continue running for up to 24hrs, even if you close your browser.

Colab Enterprise

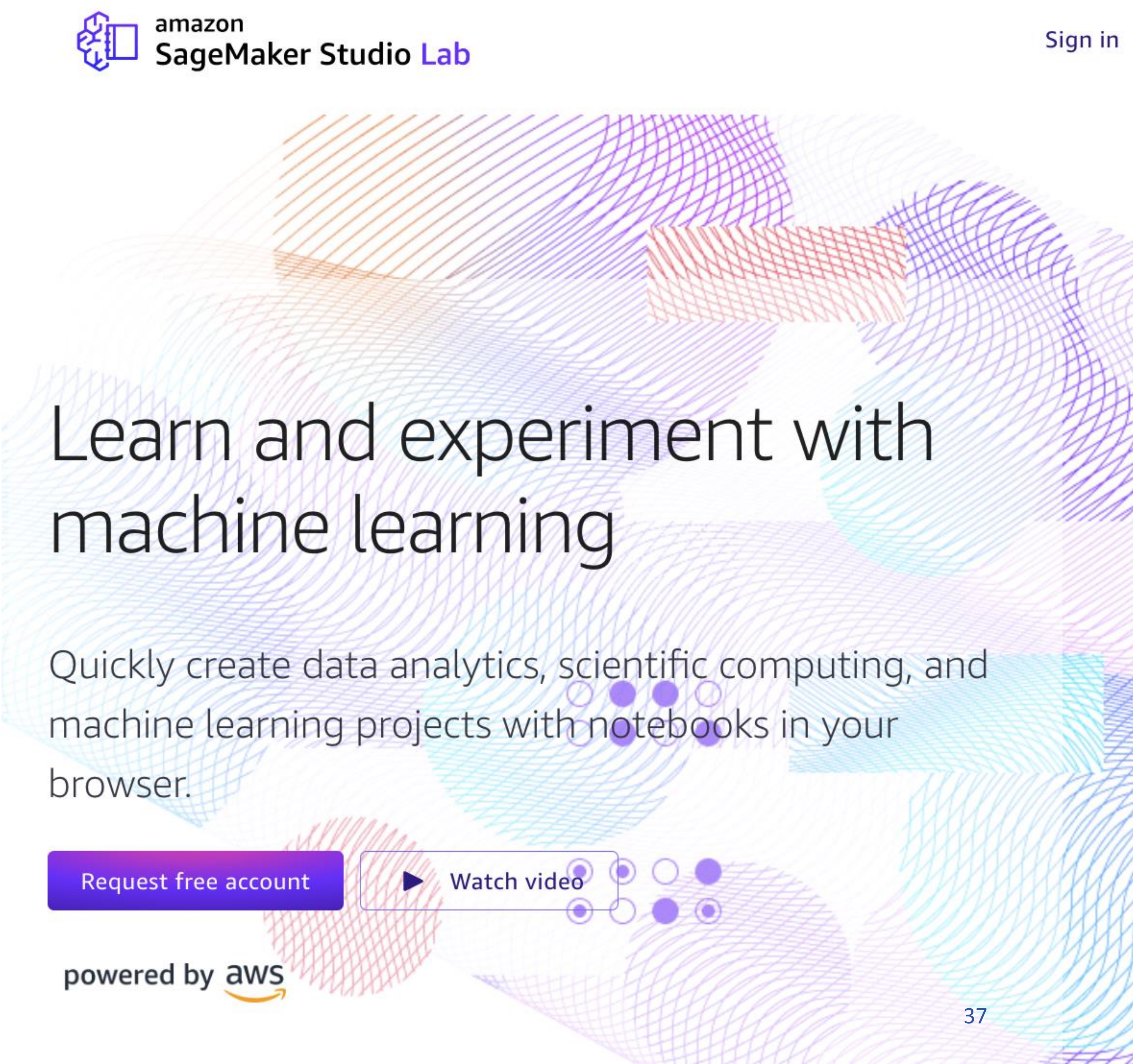
Pay for what you use

- Integrated
  Tightly integrated with Google Cloud services like BigQuery and Vertex AI.
- Enterprise notebook storage
  Replace your usage of Google Drive notebooks with GCP notebooks, stored and shared within your cloud console.
- Productive
  Generative AI powered code completion and generation.

You will benefit if you have / but not required

NVIDIA Ampere A100	$17,000
NVIDIA Tesla v100	$7,000
NVIDIA RTX 4090	$1600
NVIDIA GTX 1080 Ti	$700

Also try using:



https://aws.amazon.com/sagemaker/studio-lab/

Demos

https://vislang.ai/genderless

vislang | University of Virginia

home people demos publications

Genderless

Our group has produced several models and diagnostic methods for addressing gender bias in natural language processing and computer vision. Here we leverage our **ICCV 2019** paper: Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In this paper we proposed a method to adversarially remove as much as possible from an image any features that could be predictive of whether a person will use a gendered word to describe it. We used a large dataset of images with captions and selected images that had references in the text such as "man" or "woman" and trained a model that can recognize the objects in the image but has as much difficulty as possible in predicting gender. When we applied this transformations to the image space, we can examine what the model is trying to do. Try your own images below and see what it does.

upload an image paste image URL

Tap here to choose an image...

Original Image

Genderless Image

Demo by Lindsey and Vicente

Demos

https://vislang.ai/text2scene

vislang | University of Virginia

home people demos publications

Text2Scene

Text2Scene was proposed in a paper by our group at **CVPR 2019** as Text2Scene: Generating Compositional Scenes from Textual Descriptions. This model takes as input textual descriptions of a scene and generates the scene graphically object by object using a Recurrent Neural Network, highlighting their ability to learn complex and seemingly non-sequential tasks. The more advanced version of our model requires more computing but can also produce real images by stitching segments from other images. Read more about Text2Scene in the in the research blogs of IBM and NVIDIA and download the full source code from https://github.com/uvavision/Text2Scene. This demo generates cartoon-like images using the vocabulary and graphics from the Abstract Scenes dataset proposed by Zitnick and Parikh in 2013.

Besides Mike and Jenny feel free to reference any of these other objects: bear, cat, dog, duck, owl, snake, hat, crown, pirate hat, viking hat, witch hat, glasses, pie, pizza, hot dog, ketchup, mustard, drink, bee, slide, sandbox, swing, tree, pine tree, apple tree, helicopter, balloon, sun, cloud, rocket, airplane, ball, football, basketball, baseball bat, shovel, tennis racket, kite, fire. Also feel free to describe Mike and Jenny with other attributes or action words such as sitting, running, jumping, kicking, standing, afraid, happy, scared, angry, etc.

#1 Mike is next to a tree

#2 Jenny is happy and kicks the ball

#3 There is a fire

Generate Scene

Demo by Leticia and Vicente

AMC Visual Grounding

https://vislang.ai/amc

upload an image

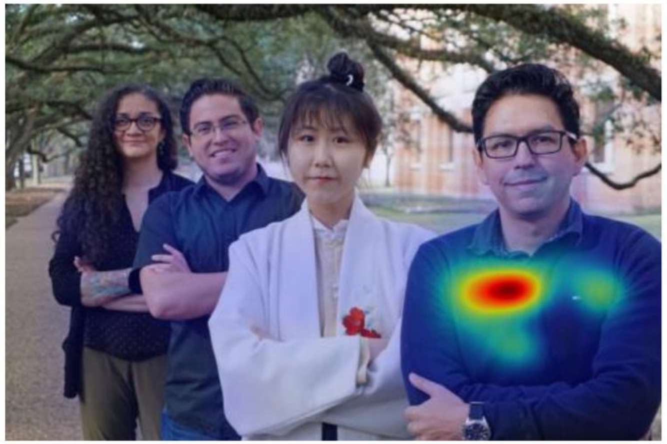
Tap to choose image or use default ...

Text Prompt: a man in a blue shirt

Compute Heatmap

Original Image

P("a man in blue shirt") = 0.65



Demo by Ziyan and Vicente

For Next Class…

- Intro to Machine Learning

- You need to complete the following two activities:

Completing this [Primer on Image Processing], and optionally, the tutorial and assignment on [Image Classification] from my old Deep Learning for Visual Recognition class.

Questions?