PEC
EXPLORE INNOVATE EXCEL

Vision-Language Learning using Pix2Struct & Kosmos-2

Deep Learning

GET STARTED →

Team Members

Manav Rai 22106028

Vishwas Kisaniya 22106027

Parth 22106026

PEC
EXPLORE INNOVATE EXCEL

Table Of Content

PEC
EXPLORE INNOVATE EXCEL

Problem Statement

- Visual data such as screenshots, UI layouts, and images require accurate text interpretation.
- Traditional models struggle with structured visual understanding and consistent image-to-text generation.
- The project focuses on evaluating two modern architectures (post-2020): **Pix2Struct and Kosmos-2.**
- Goal is to train both models on a real-world dataset for image-to-text tasks.
- Compare their performance in accuracy, text quality, and real-world applicability.

PEC
EXPLORE INNOVATE EXCEL

Selected Architectures

Architecture 1: Pix2Struct (2022)

- Google's vision-to-text model.
- Specially optimized for UI comprehension, OCR reasoning, diagram question answering.

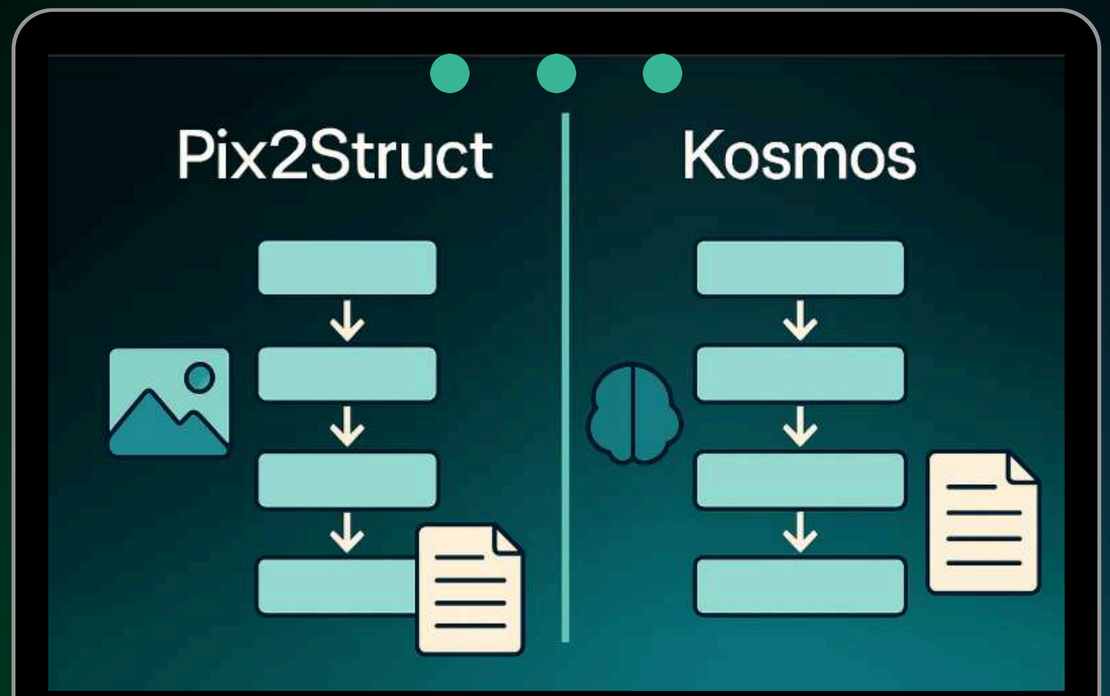
Architecture 2: Kosmos (Kosmos-2 2023)

- Uses Microsoft Multimodel Large Language Model
- Image understanding, OCR, grounding, and vision-language reasoning me expert.

ARCHITECTURES SELECTED ✓

PEC
EXPLORE INNOVATE EXCEL

Why These Architectures?



Pix2Struct

- Designed for understanding UI screens and structured layouts.
- Generates accurate text from images with low errors.
- Ideal for tasks like OCR, form reading, and diagram interpretation.

Kosmos-2

- Handles both images & text with strong multimodal reasoning.
- Performs well in captioning, OCR, and visual question answering.
- More versatile for real-world applications across different image types.

PEC
EXPLORE INNOVATE EXCEL

Dataset Details

Architecture 1: Pix2Struct (2022)

Dataset Name: Website Screenshots (naorm/website-screenshots-blip-large)
Type: Real website UI screenshots
Why this dataset?

- Ideal for image-to-text tasks and UI screenshot captioning

Architecture 2: Kosmos (2023)

Dataset Name: VQAv2 (HuggingFace)
Type: Image + Question-Answer pairs.
Why this dataset?

- Perfect for testing visual reasoning and multimodal understanding

DATASETS USED ✓

PEC
EXPLORE INNOVATE EXCEL

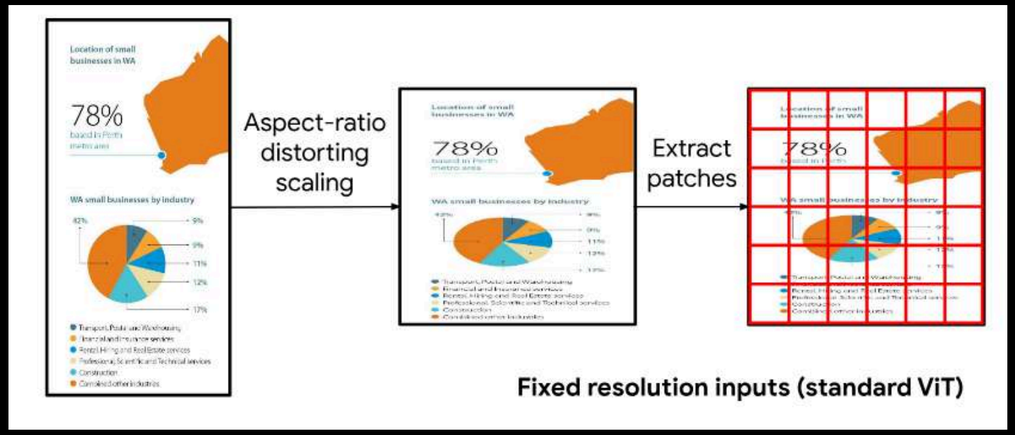
Pix2Struct

- Pix2Struct is a pretrained image-to-text model designed with the objective of screenshot parsing.
- It is a model proposed in 2023 by engineers from Google Research team.
- The model focuses on improving recognition of visually situated language, as is the case in websites and UI.
- It uses masked screenshots of web pages that is annotated with the structure of the page in an HTML-like format
- It also introduces variable resolution input representations to ViT, which overcomes one of the bigger problems with ViT based OCR models.

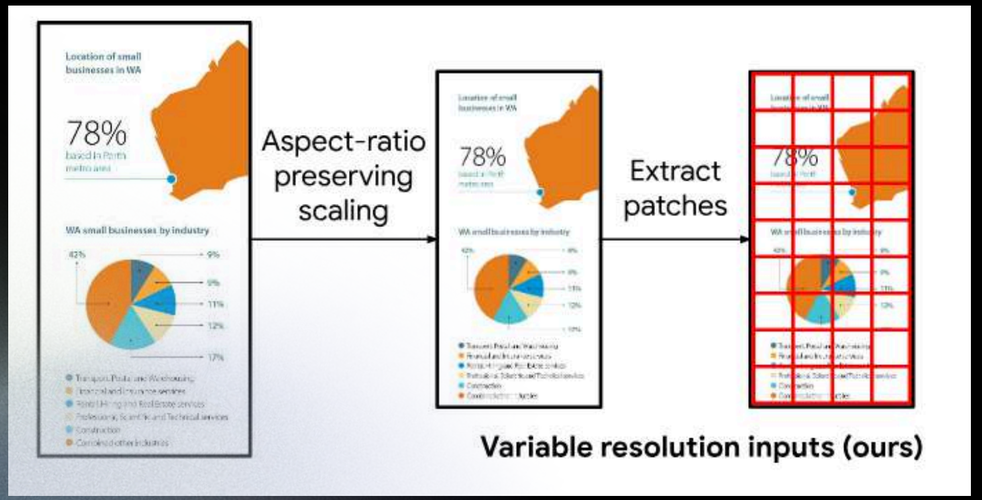
PIX2STRUCT
GOOGLE

Pix2Struct Architecture

- Pix2Struct is an image-encoder-text-decoder based on Vision Transformers (ViT).
- While most of the model is similar to a standard ViT, the input representation was changed to scale preserving aspect ratio instead of fixed resolution scaling
- The model focuses on improving recognition of visually situated language, as is the case in websites and UI.
- It uses masked screenshots of web pages that is annotated with the structure of the page in an HTML-like format

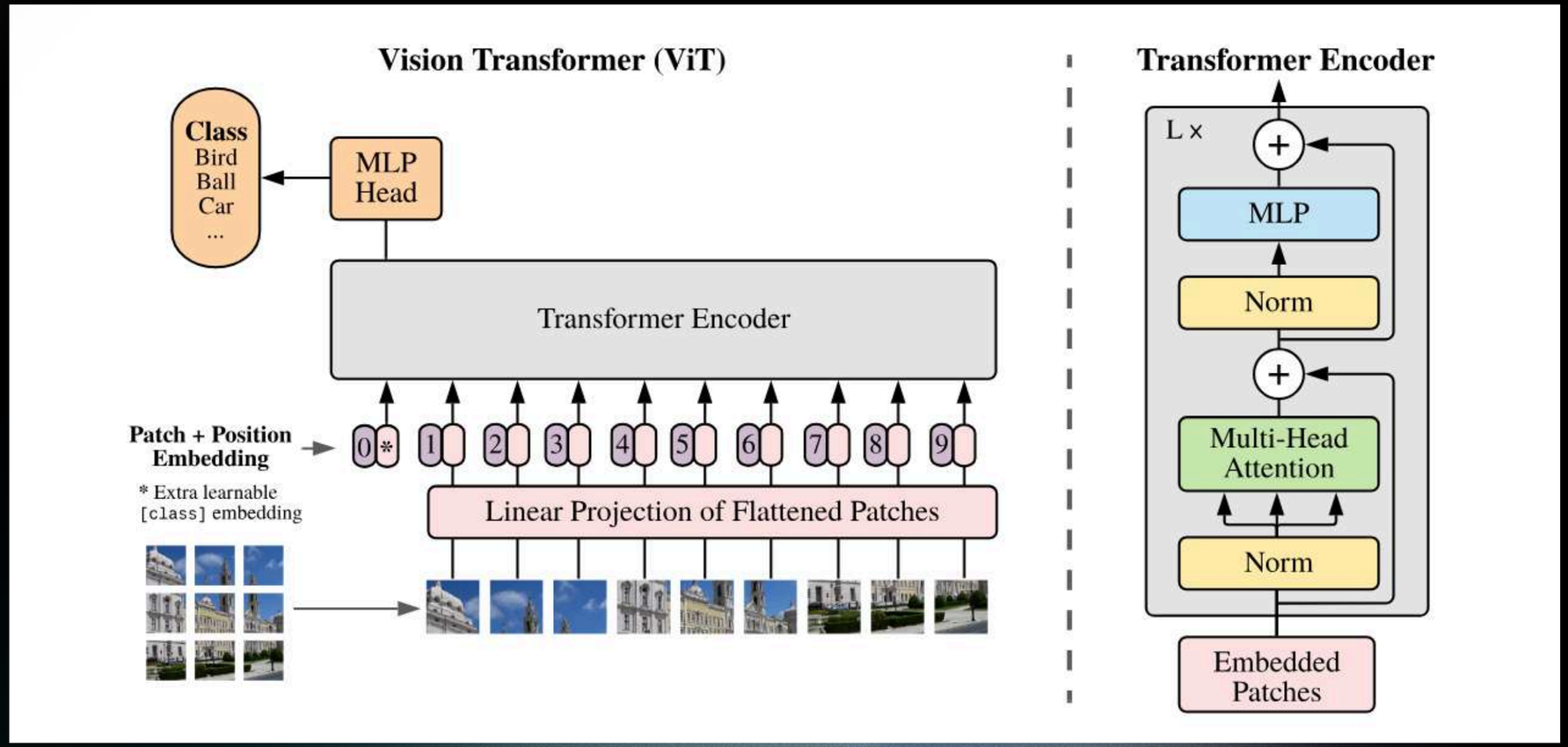


Input representation in standard ViT



Input representation in Pix2Struct

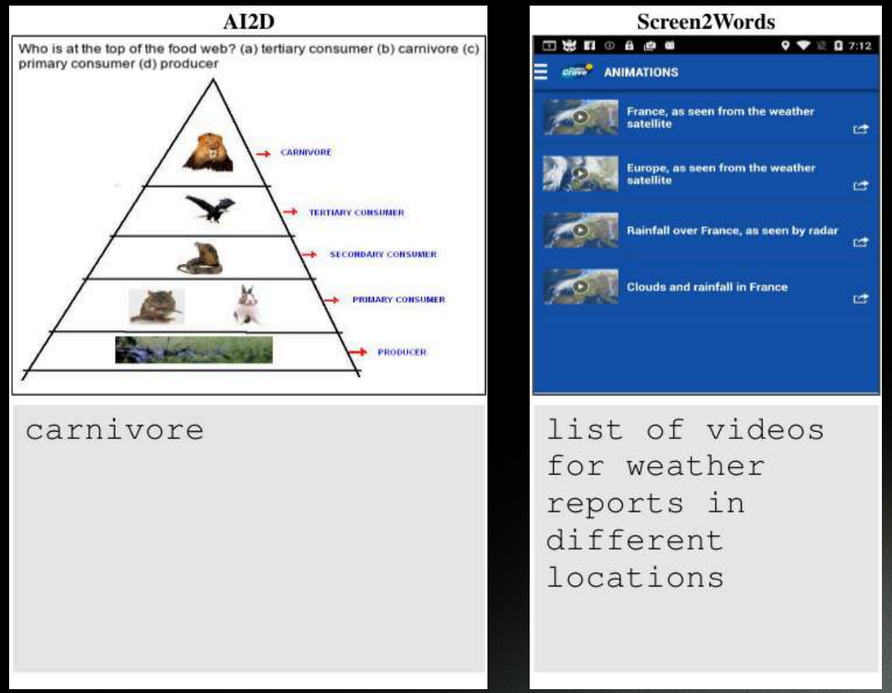
Pix2Struct Architecture



Typical ViT architecture

Pix2Struct Fine Tuning

- Finetuning Pix2Struct is straightforward matter of preprocessing the downstream data to unambiguously reflect the task in the image inputs and text outputs.
- Captioning is the most straightforward, since the input image and the output text can be directly used.
- For visual question answering, Pix2Struct opts to instead directly render the question as a header at the top of the original image. In the case of multiple choice answers, the choices also need to be rendered in the header as part of the question.



Input image and output text examples

Pix2Struct Hyperparameters

Dataset	Base			Large		
	Seq Len	Batch	Steps	Seq Len	Batch	Steps
DocVQA	4096	256	10000	3072	128	10000
InfographicVQA	6144	64	10000	3072	128	10000
AI2D	4096	32	5000	3072	32	5000
ChartQA	4096	256	10000	3072	128	10000
OCR-VQA	4096	256	10000	3072	128	10000
RefExp	4096	256	10000	3072	128	10000
Screen2Words	4096	32	10000	3072	32	10000
Widget Cap.	4096	256	5000	3072	128	5000
TextCaps	4096	256	5000	3072	128	5000

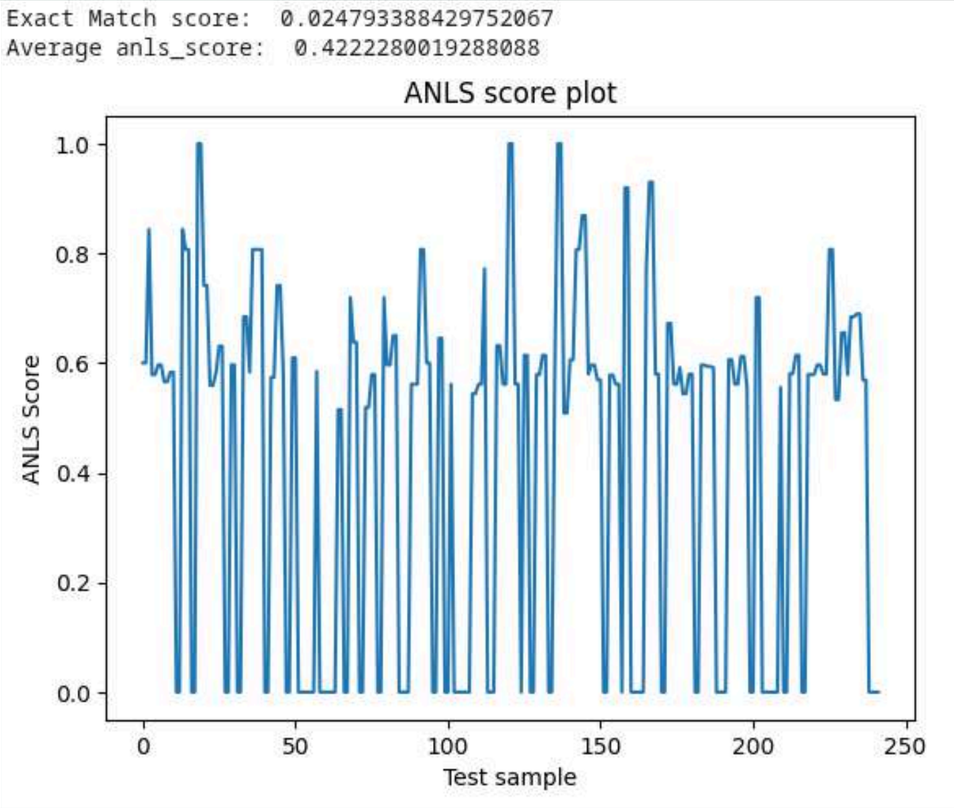
Pix2Struct Fine Tuning Hyperparameters

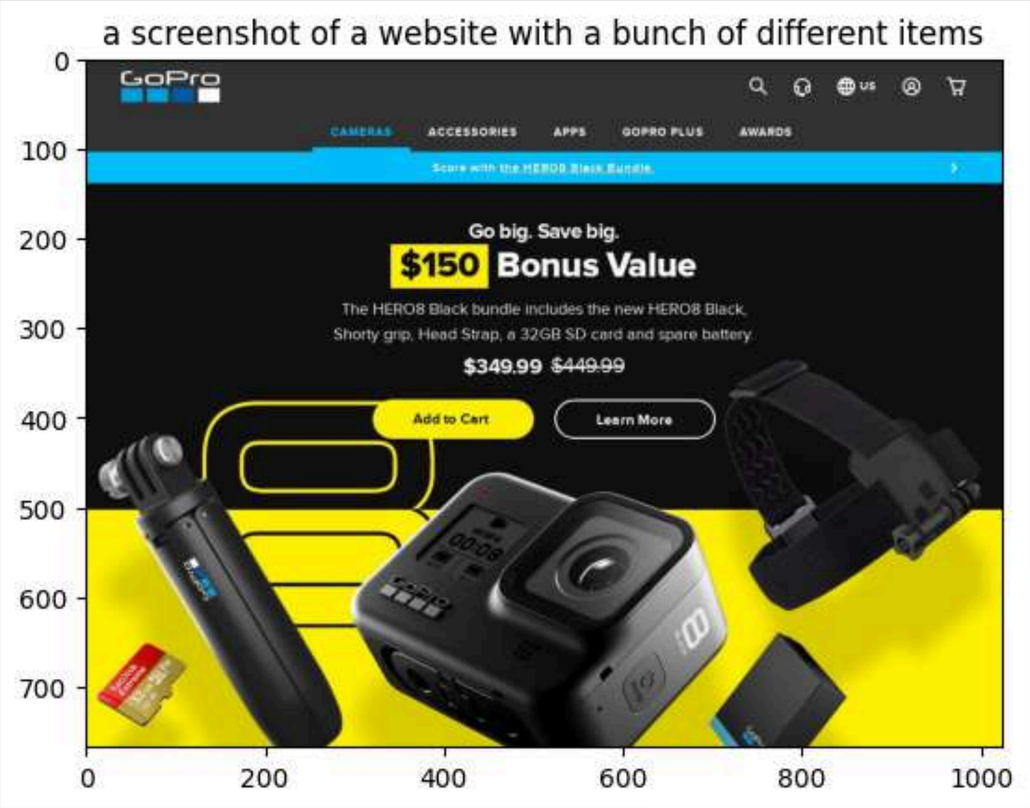
- MAX_PATCHES = 1024
- Epochs = 2
- Optimizer = AdamW
  Learning Rate = 1e-5

- MAX_PATCHES is used to define maximum amount of aspect-ratio preserving patches. The model is very sensitive to this parameter.
- 2 Epochs were used as the loss becomes zero in 2 epochs.
- AdamW with 1e-5 is useful as the model is also trained using an optimizer with a decaying learning rate.

Pix2Struct Results

Exact Match score: 0.024793388429752067
Average anls_score: 0.4222280019288088

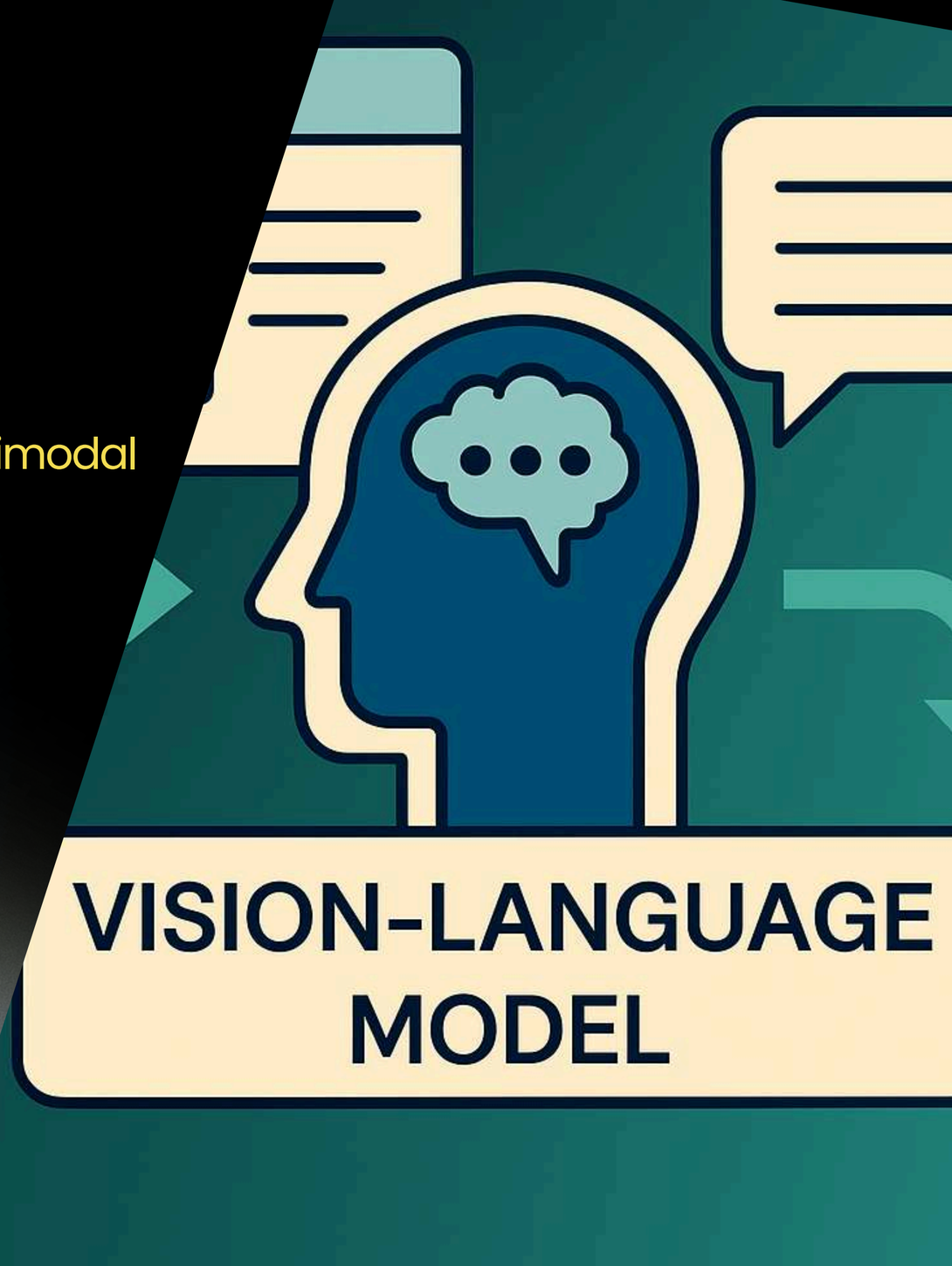




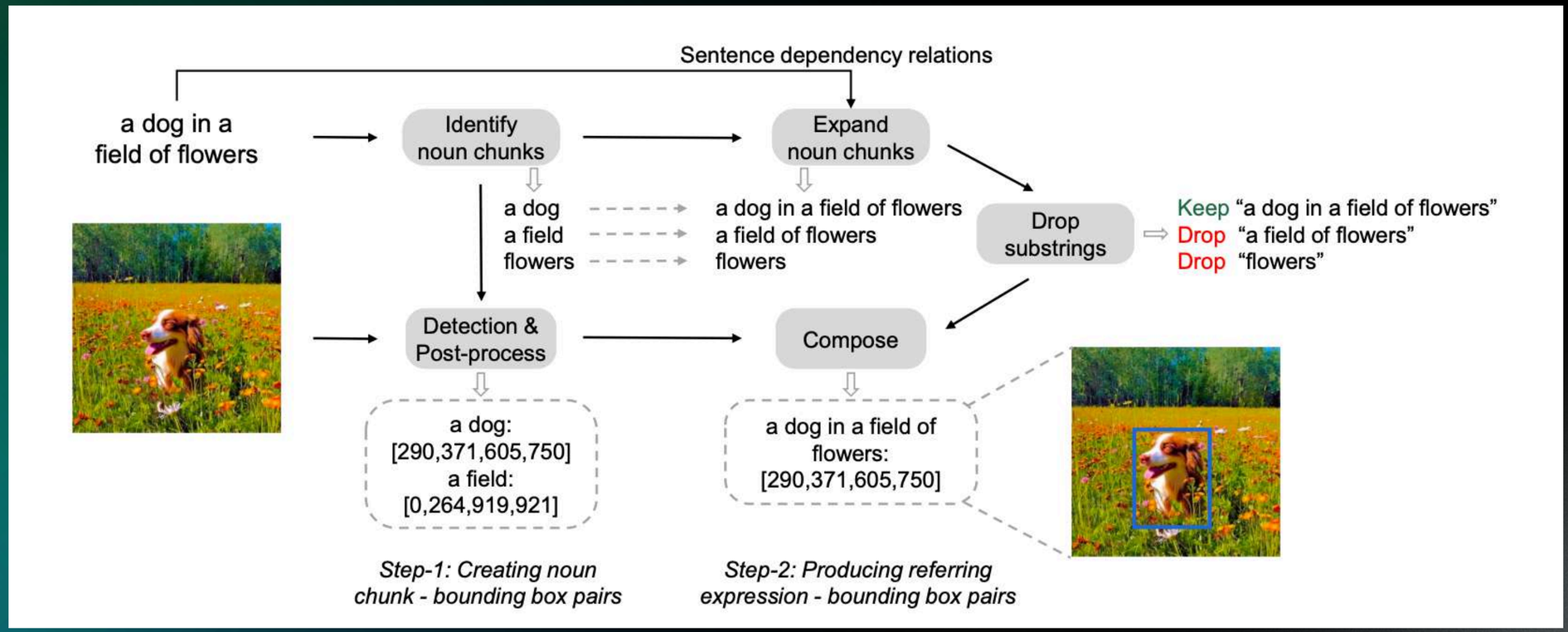
PEC
EXPLORE INNOVATE EXCEL

Kosmos-2

- It is designed to understand and generate grounded multimodal content by integrating text and visual information.
- KOSMOS-2 is a multimodal large language model (MLLM) proposed in 2023 by researchers from Microsoft Research.
- It extends KOSMOS-1 by adding grounding and referring capabilities — allowing the model to connect text spans with image regions through discrete location tokens.
- The model is trained using the same causal language modeling objective as KOSMOS-1, but on newly introduced grounded image–text data (GRIT corpus) to enable spatial understanding.



Kosmos-2 Architecture



Kosmos-2 Hyperparameters

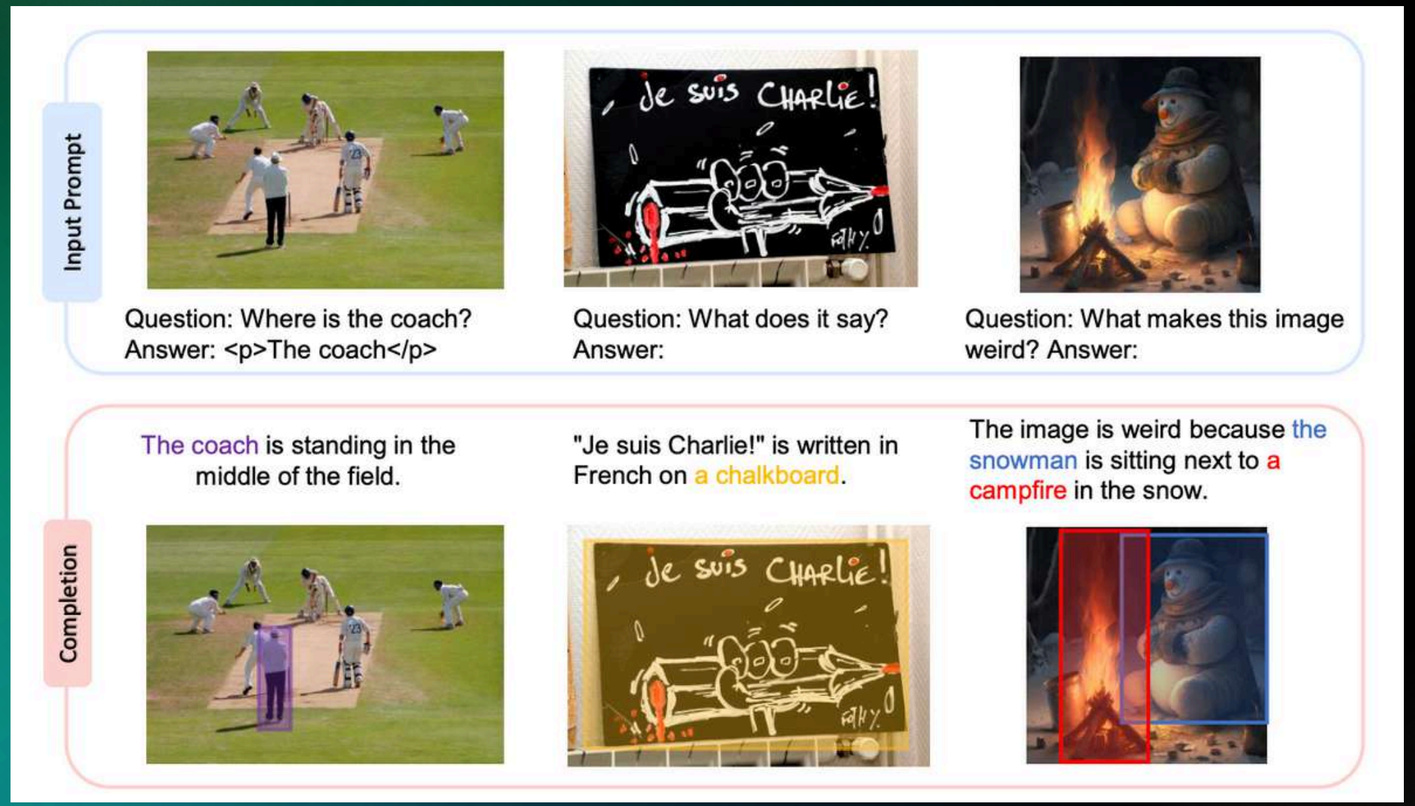
Hyperparameters	
Image embedding number	64
Location tokens	1,024
Training steps	60,000
Warmup steps	375
Optimizer	AdamW
Learning rate	2e-4
Learning rate decay	Linear
Adam β	(0.9, 0.98)
Weight decay	0.01
Batch size of text corpora	93
Batch size of original image-caption pairs	1,117
Batch size of grounded image-text pairs	1,117
Batch size of interleaved data	47

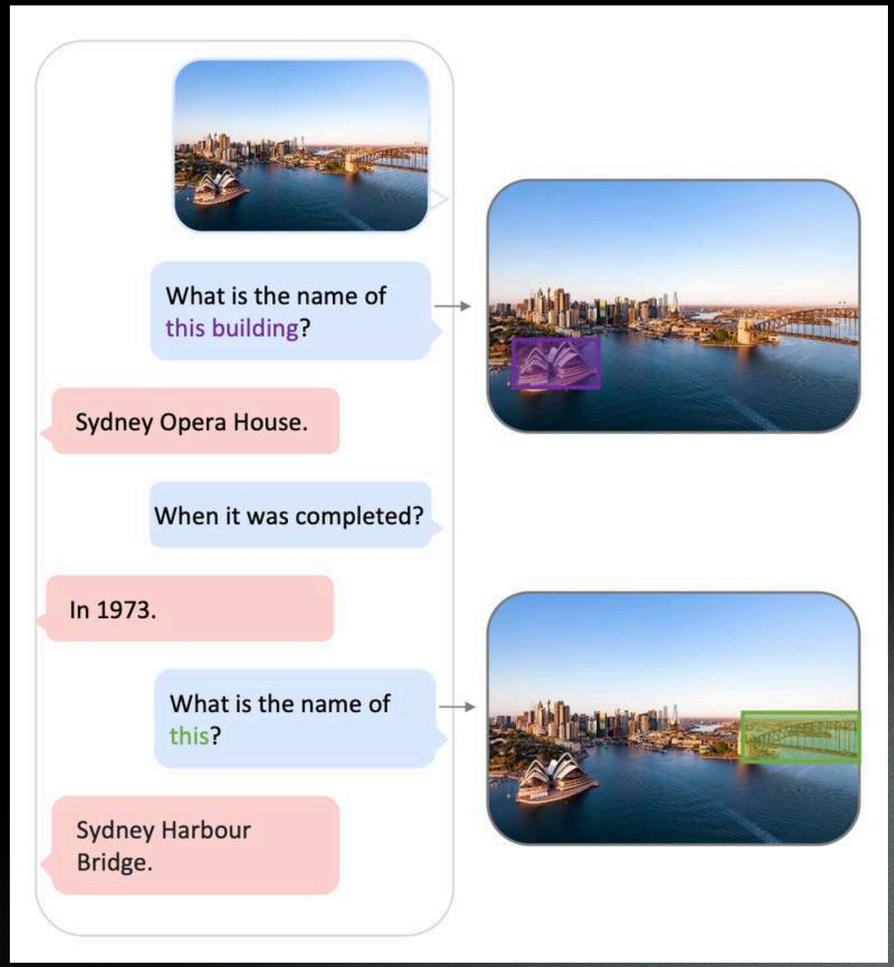
Table 7: Training hyperparameters of KOSMOS-2

Hyperparameters	
Training steps	10,000
Warmup steps	375
Learning rate	1e-5
Batch size of language instruction data	117
Batch size of vision-language instruction data	351
Batch size of grounded image-text pairs & grounded instruction data	1404
Batch size of text corpora	30
Batch size of interleaved data	15

Table 8: Instruction tuning hyperparameters of KOSMOS-2

Kosmos-2 Results





PEC
EXPLORE INNOVATE EXCEL

Real World Applications

Pix2Struct and Kosmos convert visual data into meaningful text, enabling automation, improving accuracy, and reducing manual effort in real-world systems.

PEC
EXPLORE INNOVATE EXCEL

Applications Enabled by Pix2Struct & Kosmos

1. UI & Document Automation
Automatically converts screenshots, invoices, forms, and webpages into structured text.

2. Visual Question Answering Systems
Answers user queries directly from images (useful for education, support, accessibility).

3. Accessibility for Visually Impaired Users
Generates clear descriptions of websites, documents, and diagrams.

SUMMARY COMPLETED ✓

PEC
EXPLORE INNOVATE EXCEL

Thank You