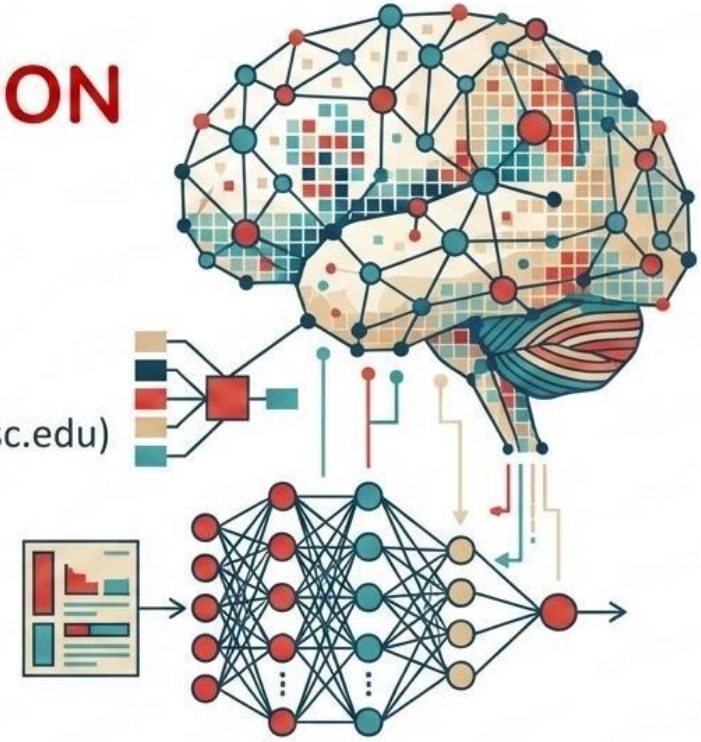
DL POSTER PRESENTATION

Team Pixels

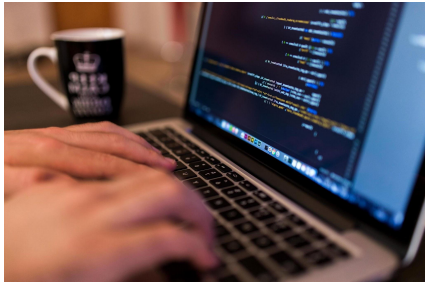
Team Member 1: Zahra Shergadwala (shergadw@usc.edu)
Team Member 2: Prakash Brahmananda Kadiyala (pkadiyal@usc.edu)
Team Member 3: Dhyey Desai (dhyeydes@usc.edu)
Team Member 4: Shubham Chaudhari (shubhamc@usc.edu)



PixelSense

Multimodal Transformer Framework for Interactive Image Understanding

Our Work

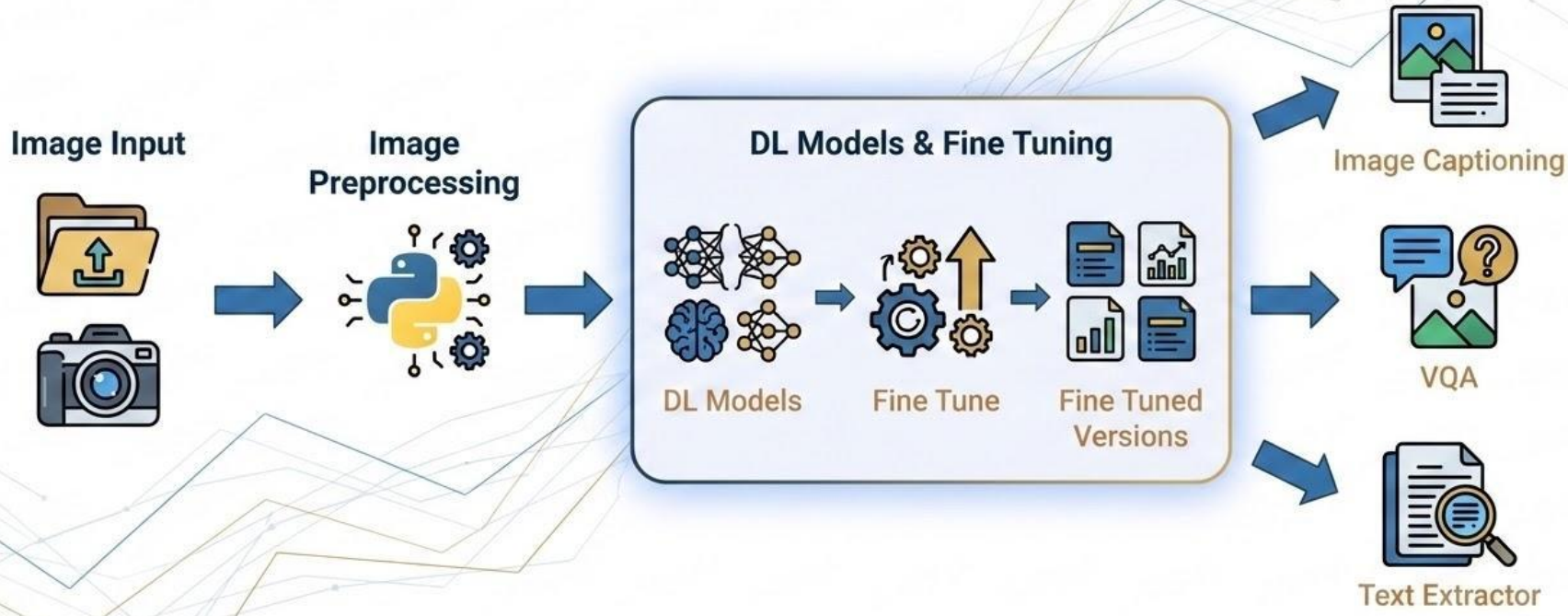


PixelSense uses a hybrid vision language pipeline that combines BLIP, BLIP-2, BLIP-Large, and InstructBLIP with a CLIP based reranking module. This setup gives us captions that are more accurate, more detailed, and more visually grounded than using any single model alone.

To extend the system beyond captioning, we added BLIP-VQA for question answering, YOLOv8 for object detection, and Pytesseract for OCR. These modules allow PixelSense to recognize objects, understand scenes, and extract text from images.

We fine tuned our models on a curated subset of the Flickr8k dataset using a mixed precision FP16 setup designed for limited compute. We also built a consistent evaluation framework using CLIP similarity, readability, and latency to measure gains in accuracy, fluency, and real time performance.

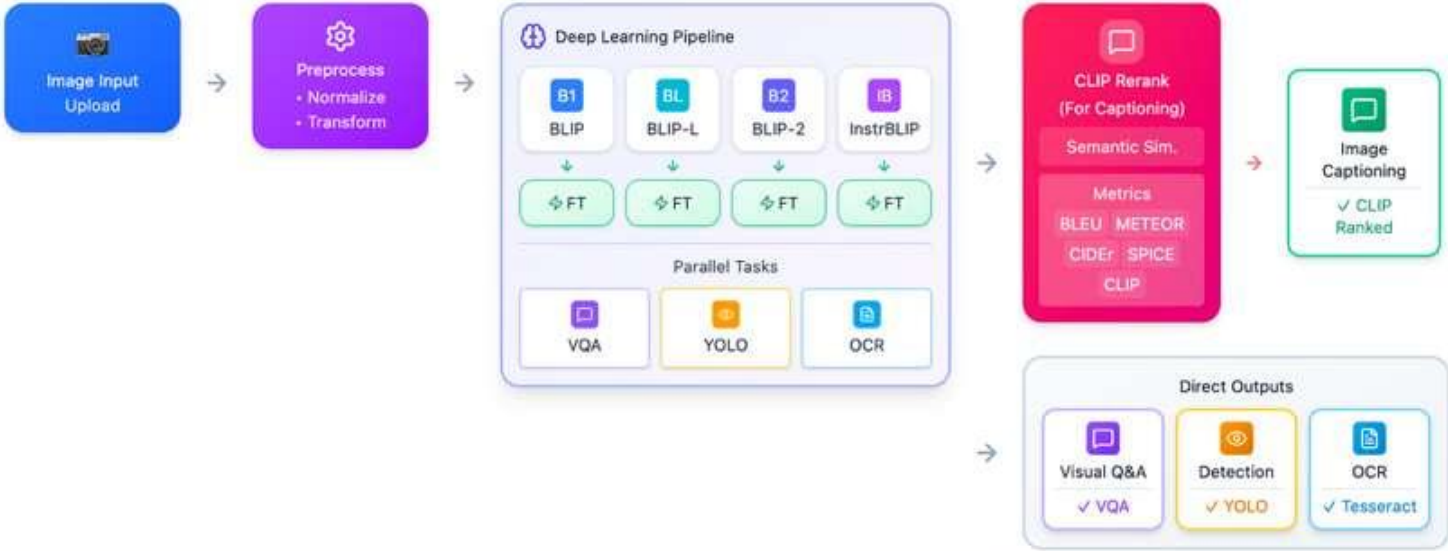
Pre Proposal Project Flow Chart



PixelSense Architecture

Interactive Multimodal Transformer Framework for Comprehensive Image Understanding

Python | Google Colab | Gradio



PixelSense: Unified Multimodal Understanding

Fine-tuned BLIP models + CLIP semantic ranking for captioning + parallel vision tasks

Mixed Precision Training

40% memory reduction with torch.cuda.amp

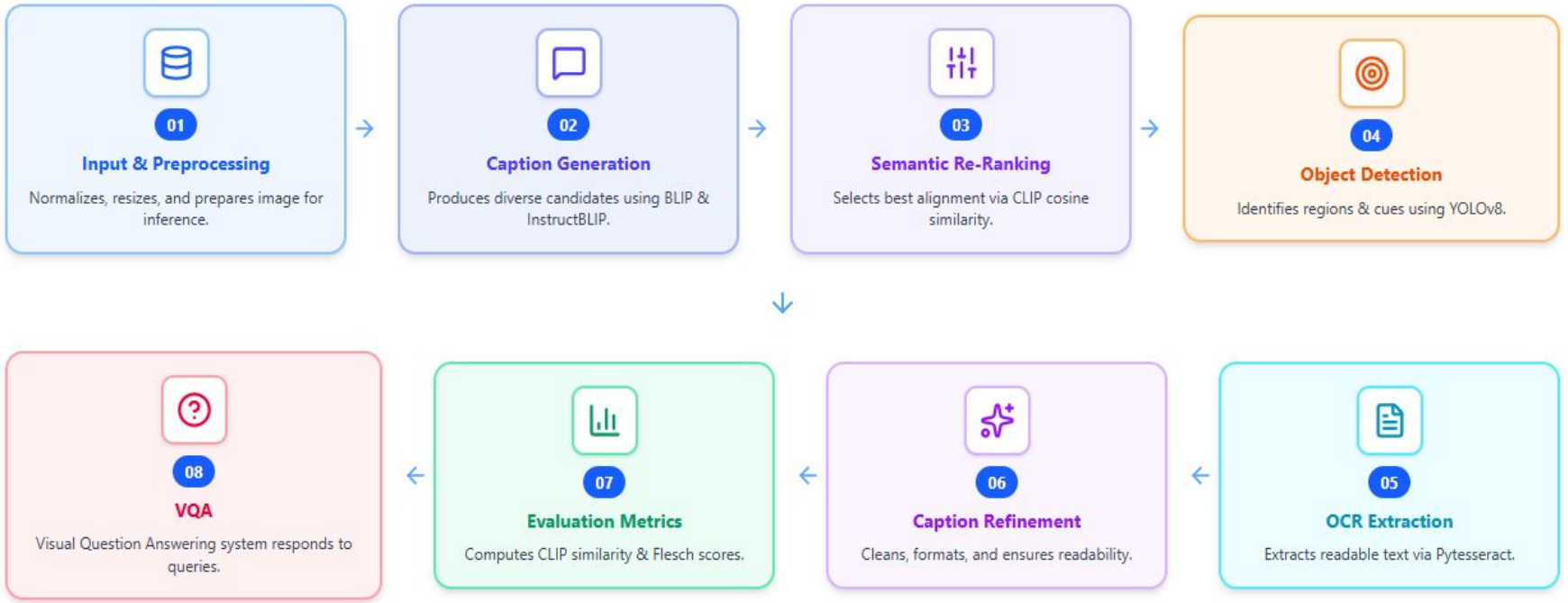
CLIP Semantic Scoring

Visual-text alignment with cosine similarity

Best Model Performance

InstructBLIP-FT: 0.872 CLIP score, 83.94 CIDEr

Modules of PixelSense



FrontEnd(Image Input to Our Project)

Users can interact with our vision pipeline in two convenient ways:

1. **Upload an Image** from their device
2. **Capture a Live Photo** using the webcam

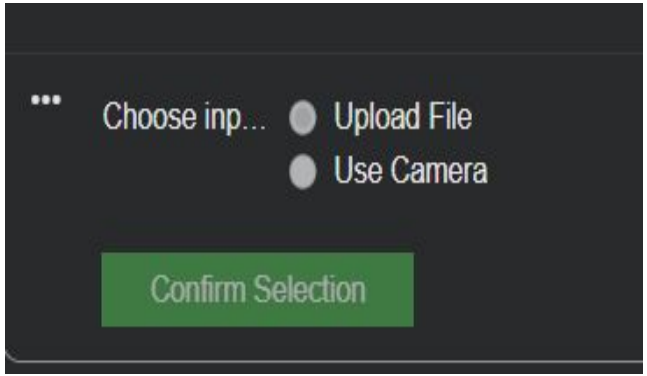
Once the image is received, it flows through a unified preprocessing layer and is passed to all downstream modules **Image Captioning, VQA, Object Detection, and OCR**.

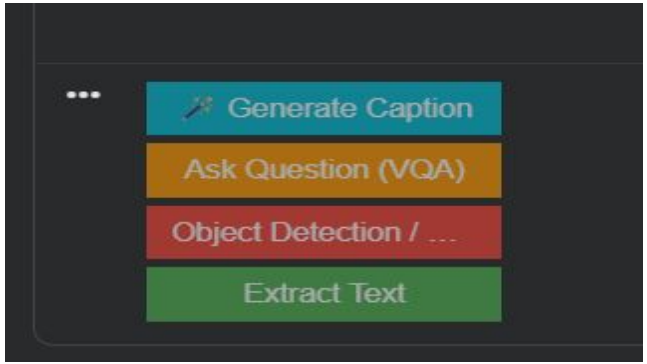
To support this workflow, we developed **two frontends**:

- **Colab IPyWidgets UI** — lightweight interface ideal for development and debugging
- **Gradio Web Interface** — modern, responsive frontend designed for a real user experience

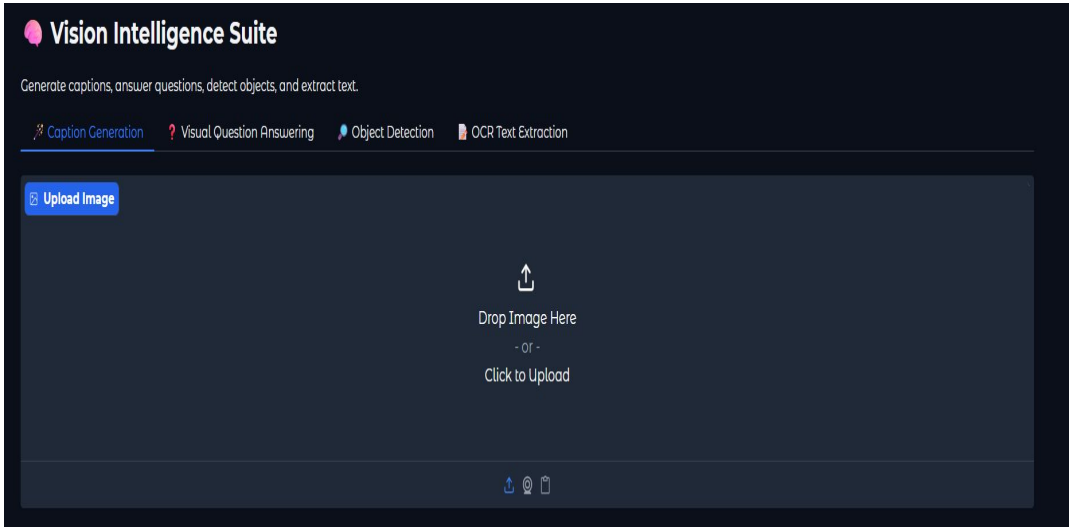
Together, these interfaces make the system accessible not only for experimentation but also during interactive use.

Colab IPYWidgets UI





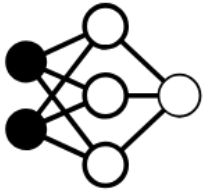
Gradio Enhanced UI



Backend(Models,Datasets,Metrics etc.)

Core Vision–Language Models

PixelSense is built on **BLIP ,BLIP- Large, BLIP-2, and InstructBLIP** as its primary captioning engines. These models use frozen ViT-G/14 vision encoders paired with Flan-T5-XL language backbones. BLIP-2's Q-Former enables efficient cross-modal alignment, allowing the system to generate detailed and context-aware captions.



Semantic Re-Ranking (CLIP Module)

A **CLIP ViT-Base/Patch32** encoder evaluates multiple caption candidates generated through beam search and selects the one with the highest image-text cosine similarity. This Hybrid Ranking Framework combines generative captions with contrastive embeddings, improving semantic precision and relevance.

Grounding and OCR Integration

Lightweight object detection with **YOLOv8s** and text extraction using **Pytesseract** provide grounding cues that enrich scene understanding. These modules supply object-level and textual context, enabling captions that are more descriptive, interpretable, and aligned with real-world visual details.

VQA Support

The system also incorporates **BLIP-based VQA** modules for interactive querying, allowing users to ask questions about objects, attributes, and scene relationships. This extends PixelSense beyond captioning into broader vision-language interaction.

BLIP Model Family

Bootstrapping Language-Image Pre-training

BLIP

Base Captioning Model

Generates fluent, image-grounded captions using a ViT encoder and transformer decoder.

BLIP-Large

Enhanced Capacity

Larger version with stronger visual-text alignment and better language fluency.

BLIP-2

Efficient Multimodal

Frozen image encoder + LLM via query transformer. Efficient reasoning with lower cost.

InstructBLIP

Instruction-Tuned

Instruction-tuned BLIP-2 that follows prompts for detailed, context-aware captions.

Related Vision Models

Complementary models for comprehensive visual understanding

VQA-BLIP Base

Visual Q&A

Specialized for Visual Question Answering with strong visual grounding.

YOLOv8

Object Detection

Real-time object detection with bounding boxes and scene-level grounding.

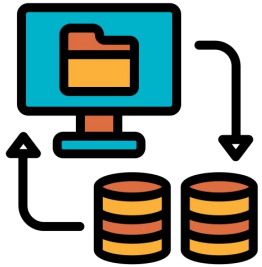
Pytesseract

OCR Engine

Extracts embedded text from images for signs, documents, and text-rich scenes.

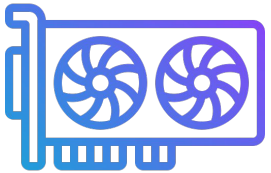
Dataset

A curated subset of 200 images from the **Flickr8k dataset,** paired with 1000 human-written reference captions, is used for fine-tuning. This controlled subset supports efficient experimentation under limited compute resources while preserving enough visual and linguistic diversity to evaluate semantic alignment and caption robustness.



Training Setup

Fine-tuning is conducted for **3 epochs** using **AdamW** with a learning rate of **5e-5, 8-step** gradient accumulation, and **FP16 mixed precision** to reduce memory usage and stabilize training. Both base and fine-tuned variants of the **BLIP family** models are evaluated to quantify improvements in caption quality, semantic consistency, and inference efficiency.



Evaluation Metrics

Performance is assessed using traditional captioning metrics such as **BLEU, METEOR, SPICE and CIDEr**, supplemented by **CLIP Similarity** to measure visual-semantic alignment. Linguistic clarity is evaluated with the Flesch Reading Ease score, and inference latency is recorded to examine whether the pipeline can support real-time interaction within the PixelSense framework.

Flickr 8k DataSet Sample

start two large tan dogs play along a sandy beach end
start two dogs playing together on a beach end
start two dogs playing in the sand at the beach end
start two dogs are making a turn on a soft sand beach end
start two different breeds of brown and white dogs play on the beach end

start climber climbing an ice wall end
start a person in blue and red ice climbing with two picks end
start an ice climber scaling a frozen waterfall end
start an ice climber in a blue jacket and black pants is scaling a frozen ice wall end
start a man uses ice picks and crampons to scale ice end

start a wet black dog is carrying a green toy through the grass end
start a dog in grass with a blue item in his mouth end
start a black dog has a blue toy in its mouth end
start a black dog carrying something through the grass end
start a black dog carries a green toy in his mouth as he walks through the grass end

start man and child in yellow kayak end
start a man and young boy ride in a yellow kayak end
start a man and child kayak through gentle waters end
start a man and a little boy in blue life jackets are rowing a yellow canoe end
start a man and a baby are in a yellow kayak on water end

start the chocolate lab jumps too late to get the toy as the black lab captures it in the driveway end
start black dog snaps at red and black object as brown dog lunges end
start a brown and black lab are outside and the black lab is catching a toy in its mouth end
start a black dog and a brown dog play with a red toy on a courtyard end
start a black dog and a brown dog are jumping up to catch a red toy end

CLIP Similarity

Contrastive Language-Image Pre-training

Semantic Image-Caption Matching

- Measures how well caption matches the image semantically
- Uses vision-language embeddings

BLEU

Bilingual Evaluation Understudy

N-gram Precision Scoring

- Measures textual closeness to references
- Based on n-gram precision

METEOR

Metric for Evaluation of Translation with Explicit ORdering

Advanced Alignment Scoring

- Considers synonyms and stemming
- Alignment-based scoring

CIDEr

Consensus-based Image Description Evaluation

Consensus-Based Evaluation

- Weighs rare words higher
- Correlates with human-like descriptions

SPICE

Semantic Propositional Image Caption Evaluation

Scene Graph Analysis

- Measures objects, attributes, and relations
- Semantic accuracy focused

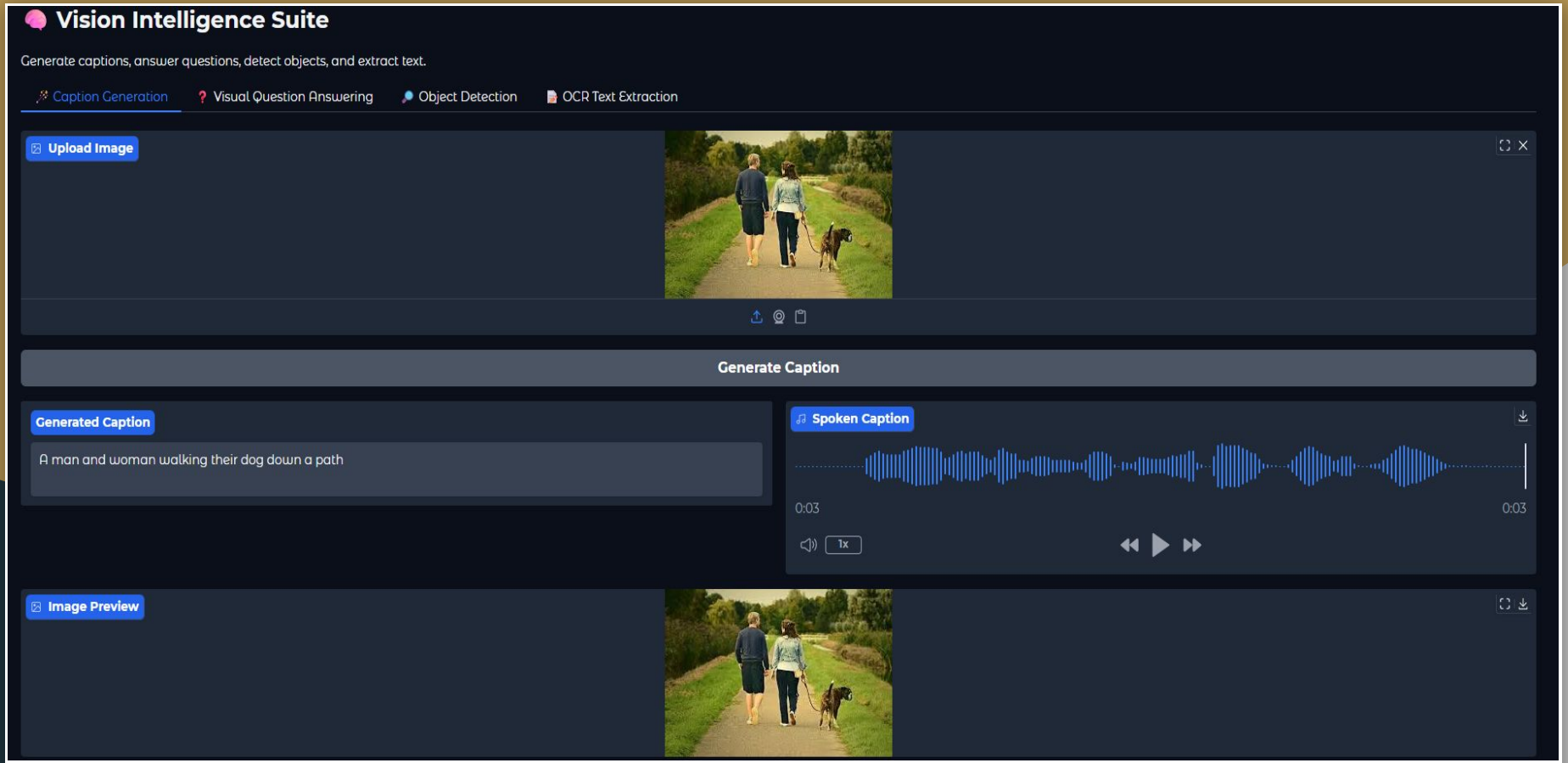
Readability Score

Human Interpretability Metrics

Human Interpretability

- Fluency and clarity measurement
- Grammatical correctness

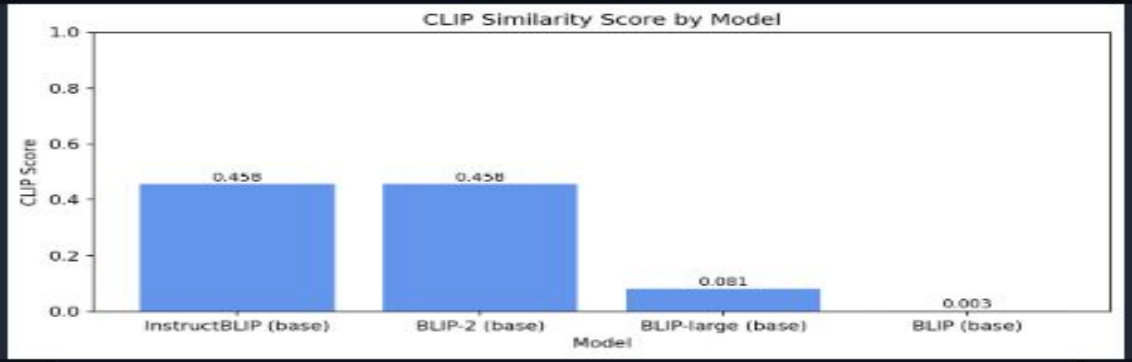
Evaluation Metrics For Image Captioning



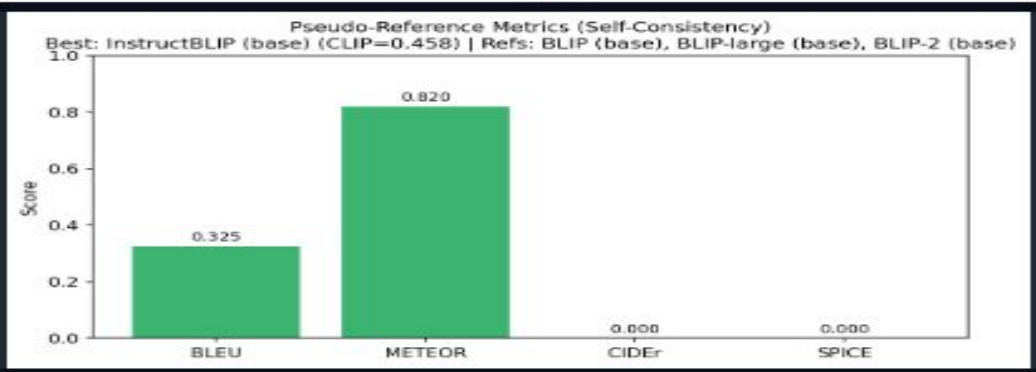
Gradio UI Based Image Captioning Output

Image Preview

CLIP Ranking Plot



Metric Scores Plot



Model	Caption	CLIP Score	Length	Time (s)	Readability
InstructBLIP (base)	A man and woman walking their dog down a path	0.457628	10	1.02	95.17
BLIP-2 (base)	A man and woman walking their dog down a path	0.457628	10	1.34	95.17
BLIP-large (base)	There are two people walking down a path with a dog on a leash	0.081492	14	0.52	95.84
BLIP (base)	A woman walking a dog down a path	0.003251	8	0.95	92.97

Generating visualizations...

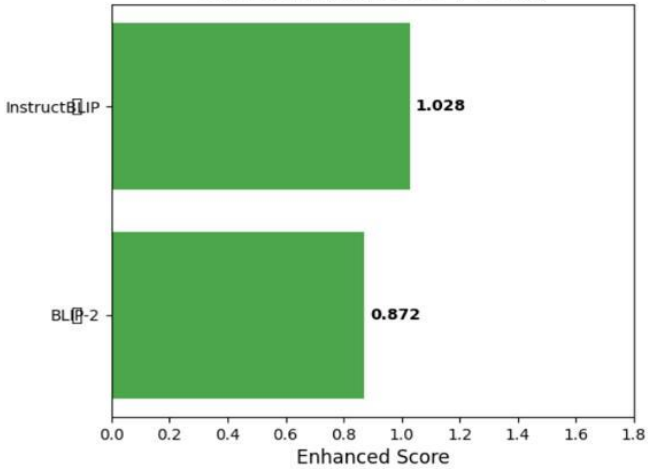
Comprehensive Image Captioning Results

Best Caption (InstructBLIP)

a glass of orange juice, lemons and limes on a table

Original Image

Caption Rankings (CLIP + Detail)



Model Comparison

Model	Status	Enhanced	Caption
BLIP-2	Fine-tuned	0.872	a glass of orange juice
InstructBLIP	Fine-tuned	1.028	a glass of orange juice, lemons and limes on a tab...

System Status

BLIP-2: Fine-tuned
InstructBLIP: Fine-tuned

Enhanced fine-tuning active
3 epochs per model
CLIP ranking active

Summary Statistics

Best Enhanced Score: 1.028
Best Model: InstructBLIP
Caption Length: 11 words

Enhanced Scores:
BLIP-2: 0.872
InstructBLIP: 1.028

Future Scope

Sentiment-Aware Image Understanding

3D-Based Object Detection

Video-Level Captioning & Analysis

Multilingual Captioning & VQA

Scene Graph–Based Reasoning

Domain-Specific Captioning (Medical, Retail, Accessibility)

Enhanced OCR-Grounded Understanding

Real-Time Multimodal Interaction (Voice + Vision)

Conclusion

- PixelSense demonstrates a unified multimodal intelligence system that brings together image captioning, visual question answering, object detection, and OCR within a seamless and interactive workflow.
- By leveraging the BLIP family of models BLIP, BLIP-Large, BLIP-2, and InstructBLIP and fine-tuning them on the Flickr8k dataset, the system learns to generate richer, more descriptive, and context-aware captions tailored to diverse scene types.
- A core strength of PixelSense is its hybrid captioning pipeline: multiple BLIP-based models generate candidate captions, and CLIP semantic reranking selects the most visually aligned result.
- Combined with readability scoring and pseudo-reference metrics (BLEU, METEOR, CIDEr, SPICE), this provides a robust and holistic evaluation of caption quality. These improvements remain effective even under compute limitations, thanks to mixed-precision fine-tuning, model optimization, and streamlined data handling.

Continuation...

- Beyond captioning, the system integrates YOLOv8 for object detection and Tesseract for OCR, enabling grounded, multi-signal scene understanding. The addition of BLIP-VQA enables users to ask natural questions about images, expanding the system's reasoning capabilities.
- PixelSense also incorporates voice output via gTTS and offers two complete front-end experiences: an interactive ipywidgets interface and a polished Gradio application supporting captioning, VQA, detection, and OCR within a single platform.
- Together, these components show how multimodal transformers, semantic ranking, and lightweight fine-tuning can be combined into a practical, user-friendly, and extendable framework for real-time visual understanding.
- PixelSense sets the groundwork for future expansions in sentiment-aware perception, 3D reasoning, multilingual output, and domain-specific applications, reinforcing its potential as a versatile platform for advanced multimodal AI.

Thank You

Team Pixels