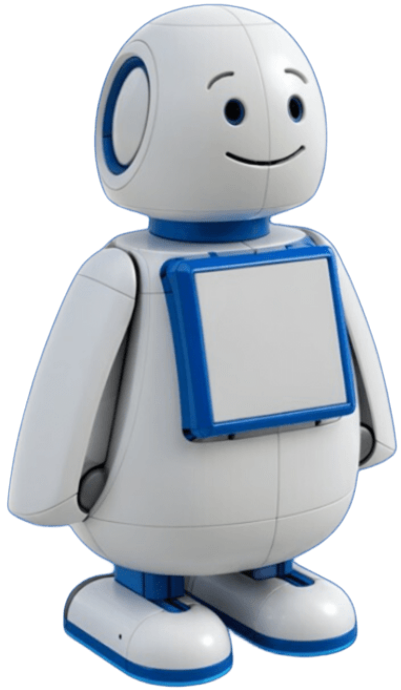
El papel de la IA en la IHR

Dr. Carlos Ricardo Cruz Mendoza

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM

23 de febrero de 2026

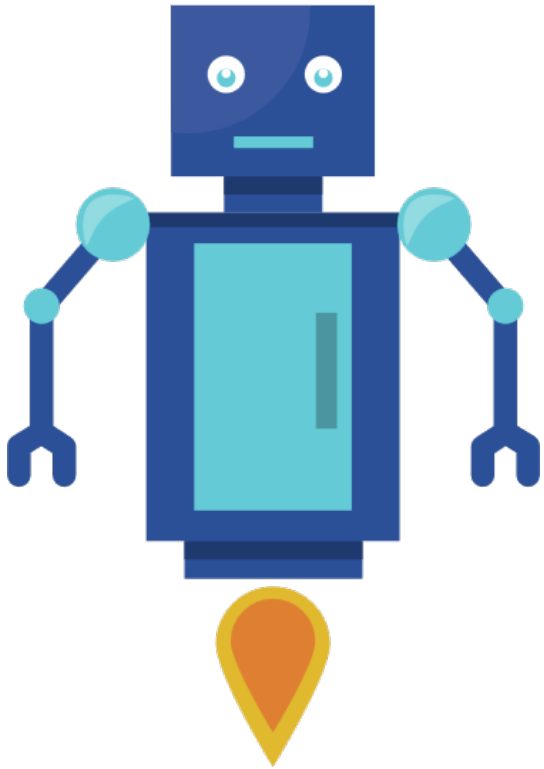


¿Qué es la IHR?

Campo interdisciplinario que estudia cómo los humanos y robots se comunican e interactúan.

Componentes Esenciales

- **Entrada.** Voz, gestos, expresiones faciales, tacto.
- **Procesamiento.** Interpretación de intenciones y contexto.
- **Salida.** Respuestas verbales y acciones físicas.



La IA como núcleo facilitador

Capacidad Humana	**Equivalente en Robot**	**Tecnología**
Entender lenguaje	Procesamiento de lenguaje	Procesamiento de Lenguaje Natural y Audición Robótica
Reconocer emociones	Análisis afectivo	Visión por computadora, Aprendizaje Profundo
Aprender de experiencias	Adaptación de comportamiento	Aprendizaje de Máquina
Memoria	Recordar contexto	Redes Neuronales, LLMs con Memoria Externa, Memoria Asociativa Entrópica.

Procesar Lenguaje Natural

Capacidad Humana
Procesamos lenguaje con comprensión contextual, inferencia pragmática y adaptación a distintos registros.

- **Natural Language Processing (NLP)**: Procesamiento computacional del lenguaje.
- **Large Language Models**: Sistemas diseñados para entender el lenguaje natural

Herramientas

- **Hugging Face**: Modelos pre-entrenados (BERT, BART, GPT).
- **spaCy/NLTK**: Pipelines para NLP (recopilación, preprocesamiento, análisis semántico y sintáctico, extracción de características, modelos y visualización).
- **Tensor Flow/ PyTorch**: Frameworks de machine learning.

Procesar Lenguaje Natural - Ejemplo

Ejemplo HRI

Robot interpretando comandos casuales.
Usuario: *'Hace mucho calor'*

Audio a Texto (Speech-To-Text)

Se procesa el audio para convertirlo a texto (Whisper o Vosk).
Opcional: VAD mediante webrtcvad.

Comprensión del Lenguaje

El usuario no dio una orden explícita pero pragmáticamente es una petición indirecta por lo que el robot debe inferir que probablemente se quiere reducir la temperatura

1. Clasificación de intención mediante un LLM o modelos ligeros.

Procesar Lenguaje Natural - Ejemplo cont...

Actuación

Una vez que se tiene el comando se envía la señal al GPIO o al bus de datos para encender un ventilador (ROS).

Texto a Audio (Text-To-Speech)

Se procesa el texto para convertilo a voz (Piper, Bark, MeloTTS).

En IHR, es importante la prosodia ya que la voz debe sonar adecuada para la situación, el usuario y la intención comunicativa. Es decir, se debe tener en cuenta la entonación, el acento, el ritmo, el tono y las pausas.

Reconocer emociones

Capacidad Humana
Detectamos emociones a través de expresiones faciales, lenguaje corporal, tono de voz y contexto.

- **Visión por computadora**: Análisis de datos obtenidos por cámaras.
- **Redes Neuronales Convolucionales (CNNs)**: Extracción de características visuales.

Herramientas

- **OpenCV**: Framework para el procesamiento de imagenes.
- **Media Pipe**: Marcas faciales y detección de pose.
- **Yolo o RT-DETR**: Detección de objetos en tiempo real.
- **Vision-Language-Models**: CLIP, SigLIP, LLama 3.2-Vision, DeepSeek-VL, SmolVLM2.

Reconocer emociones - Ejemplo

Ejemplo HRI

Robot interpretando la frustración de un estudiante al resolver un exámen.

Análisis de expresiones Faciales

- Face detector (MediaPipe, OpenCV)
- CNN o ResNet o MobilNet entrenada en reconocer emociones (FER-2013, RAF-DB, AffectNet)

Análisis de postura

Detección de puntos clave mediante MediaPipe, OpenPose, MoveNet. Extracción de características buscando hombros caidos, manos en la cabeza o sien.

Reconocer emociones - Ejemplo cont...

Clasificación multimodal

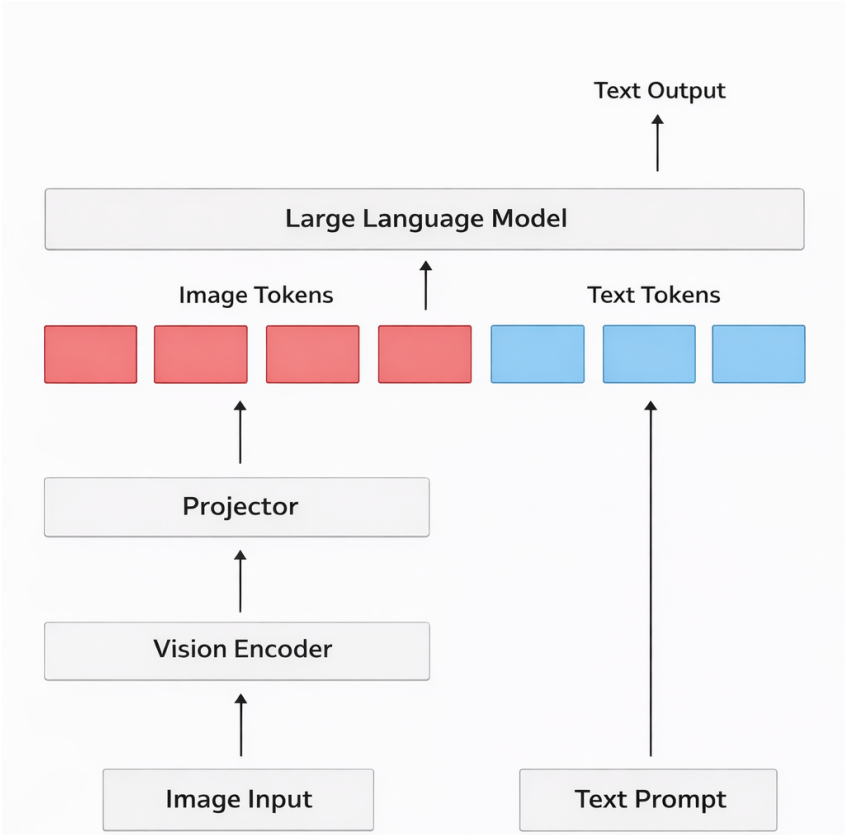
- Se combinan las probabilidades obtenidas al reconocer expresiones faciales con los datos obtenidos en el análisis de postura y se utilizan como entrada de otra red para dar un resultado final.
- Se utiliza un VLM con un fine tunning utilizando un dataset de emociones.

¿Qué es Vision-Language Model (VLM)?

- Modelo multimodal que procesa **imagen + texto**.
- Aprende una representación conjunta en un espacio compartido.
- Permite tareas como:
  - Análisis de imagenes
  - Resumen de un video
  - Parseo de documentos (OCR)
- Se pueden aplicar a la creación de chatbots multimodales.

Arquitectura general de un VLM

- Tres bloques principales
  - **Codificador de visión:** transforma imágenes o video en embeddings (extractor de características visuales).
  - **Fusor o proyector:** alinea los espacios de visión y lenguaje en un mismo dominio semántico (Linear, MLP y Q-Former).
  - **LLM:** interpreta o genera texto a partir de la información visual.
- Generalmente se basa en arquitecturas tipo Transformers como CLIP o ViT.



Aprender de Experiencias

Capacidad Humana

Ajustamos nuestro comportamiento basándonos en resultados previos y generalizamos ese comportamiento.

- **Aprendizaje supervisado**: Aprender a partir de ejemplos etiquetados.
- **Aprendizaje no supervisado**: Encontrar patrones sin supervisión.
- **Aprendizaje por refuerzo**: Aprender por prueba y error.

Herramientas

- **Scikit-learn**: Framework con algoritmos de ML clásico (Regresión Lineal, SVM, K-Means, Random Forest).
- **Stable-Baselines**: Implementaciones de RL (PPO, DQN, etc).
- **TensorFLow / Pythorch**: Frameworks de Deep Learning (CNN, Transformers, LLMs, VLMs).
- **Vision-Languages-Action VLA**

Ejemplo

Robot moviendo el brazo hacia una dirección para ejecutar una acción.

- Mediante Aprendizaje por Refuerzo. Ejemplo donde se entrena un brazo robótico virtual de 3 grados de libertad en Gymnasium usando aprendizaje por refuerzo (PPO), para que aprenda a mover su brazo hasta alcanzar un punto objetivo en el espacio.

  `https://colab.research.google.com/drive/1-LdZvh120yFUtcqnt976VxWINVKdvycv?usp=sharing`

- Mediante un VLA

¿Qué es Vision-Language-Action VLA?

Un modelo VLA integra tres capacidades fundamentales:

- **Visión:** interpreta imágenes o video del entorno.
- **Lenguaje:** comprende instrucciones en lenguaje natural.
- **Acción:** ejecuta movimientos en el mundo físico.

Idea central:

$$\text{Percibir} \rightarrow \text{Entender} \rightarrow \text{Actuar}$$

A diferencia de un VLM, un VLA no solo describe la escena, sino que toma decisiones y ejecuta acciones.

Arquitectura de un VLA

Componentes principales:

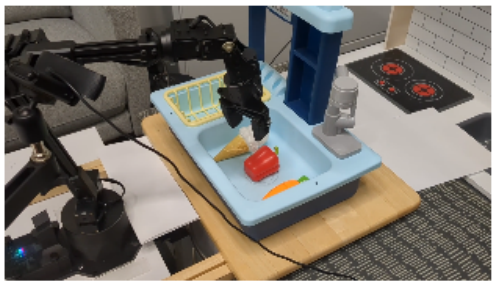
1. **Vision Encoder** Extrae características visuales del entorno.
2. **Language Encoder / LLM** Procesa la instrucción del usuario.
3. **Fusión Multimodal** Integra visión y lenguaje en un espacio compartido.
4. **Policy / Action Head** Traduce la representación en acciones (movimientos, trayectorias o comandos motores).

Ejemplos:

- RT-2 y RT-2-X (Google DeepMind)
- OpenVLA
- Gemini Robotics-ER 1.5
- Π_0 (Universidad de California)

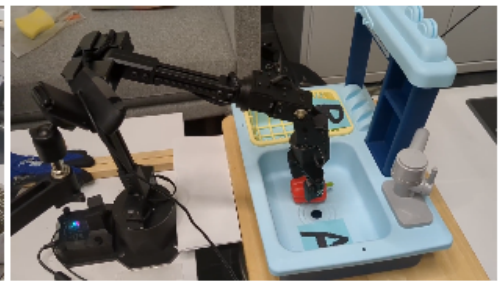
OpenVLA

- Modelo VLA de código abierto entrenado con miles de datos reales para realizar tareas de manipulación robótica.
- El modelo es entrenado utilizando el dataset Open X-Embodiment:Robotic Learning Datasets (ASU, DeepMind, Standford, UC San Diego, Toyota, etc...). Contiene más de 1 millón de trayectorias de robots reales que abarcan 22 tipos de configuraciones robóticas (embodiments), desde brazos robóticos individuales hasta robots bimanuales y cuadrúpedos.



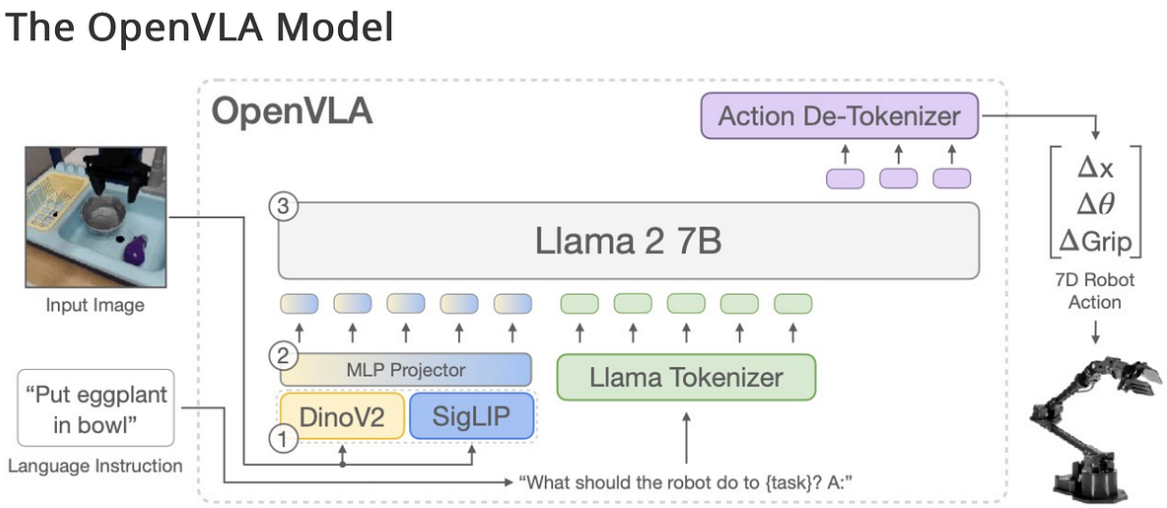
move red pepper to tray

pick ice cream



move red pepper to A

Arquitectura de OpenVLA



Dada una imagen y una instrucción en lenguaje natural, el modelo predice acciones para controlar un robot con 7 grados de libertad.

1. Un Encoder de Visión que concatena las caracteristicas obtenidas por DINO V2 y SigLIP.
2. Un Proyector que integra visión y lenguaje en un espacio compartido (language embedding space).
3. Llama 2 7B como backbone.
4. Un Destokenizador para obtener las acciones del robot.

https://www.youtube.com/watch?v=-0s0v3q7mBk

Memoria

Capacidad Humana Recordamos eventos (memoria episódica) y conocimiento (memoria semántica) para interacciones coherentes.

- **Redes Neuronales**: Almacena información en sus pesos sinápticos.
- **LLMs con Memoria Externa**: A través de un RAG dotamos a los LLMs de memoria externa.
- **Memoria Asociativa Entrópica**: Modelo computacional de memoria inspirado en la memoria natural humana, diseñado para almacenar, reconocer y recuperar información (como imágenes o datos).

Herramientas

- **Bases de datos vectoriales**: ChromaDB, Faiss, pgvector.
- **LangChain Memory**: Framework para mantener información en LLMs.
- **Grupo Mare**:

`https://turing.iimas.unam.mx/~luis/mare/index.html`

Ejemplo

Ejemplo HRI

Robot recordando preferencias de usuario.

- **Memoria episódica**: "Ayer a las 8:10 el robot le sirvió una cerveza a Ricardo"
- **Memoria semántica**: "Ricardo prefiere su cerveza bien fría"

RAG (Retrieval-Augmented Generation)

Se usan embeddings almacenados en bases de datos semánticas para recuperar información utilizada por LLMs.

Recordar contexto - RAG

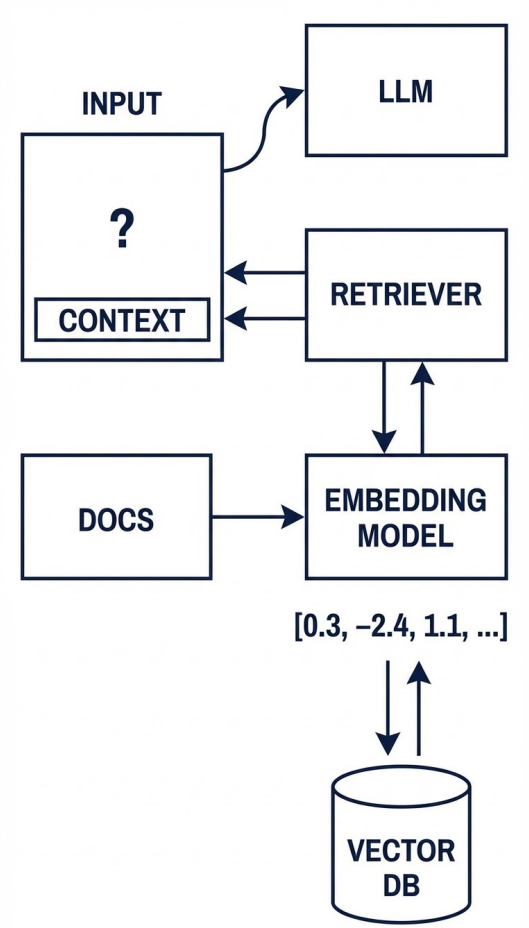
RAG (Retrieval-Augmentation Generation)

Técnica que combina una base de datos vectorial, un modelo de embedding y un LLM para generar respuestas con información guardada en memoria.

Dos Fases:

1. **Recuperación de vectores densos (embeddings)**
2. **Aprendizaje en contexto**

Embeddings: Representaciones vectoriales de datos que capturan características semánticas y sintácticas.



https://colab.research.google.com/drive/1bRUASj2K9zcM0rjaLFcvy3XzjYv0odF9?usp=sharing

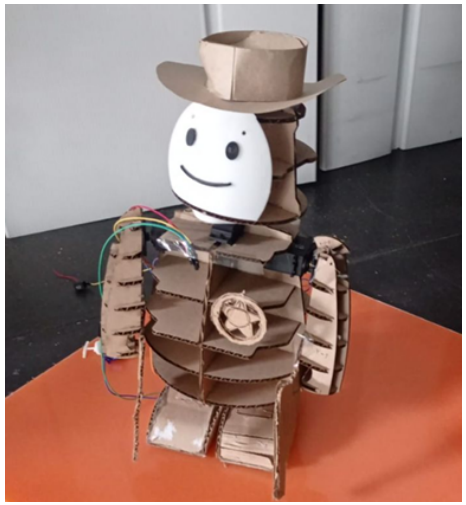
Golem IV

Aspectos de Interacción Humano-Robot e Inteligencia Artificial a investigar:

- **Percepción social y comprensión del usuario**
  - Reconocimiento de rostros, estimación de edad
  - Detección de emociones, postura y gestos
  - Identificación de prosodia (tono de voz)
  - Localización del hablante y detección de turnos
  - Multimodalidad para detección de estados afectivos
- **Expresión social y comportamiento del Robot**
  - Expresividad: Gestos, miradas, cambios de voz, imágenes en pantalla
- **Comunicación natural**
- **Aprendizaje personalizado y adaptable**
  - Adecuación de personalidad respecto al usuario

Golem IV (continuación)

- **Compañía y apoyo emocional**
  - Reducción de ansiedad en niños, adultos mayores o personas con deterioros cognitivos (autismo, depresión, etc.)
- **Educación y aprendizaje**
  - El robot puede fungir como un tutor educativo
- **Ética, confianza y aceptación social**
  - Qué diseño aumenta la confianza, qué comportamientos prefiere una persona, etc.
- **Aspectos técnicos**
  - Uso de modelos y algoritmos de IA en general aplicados a IHR
  - Uso en contextos reales



MUCHAS GRACIAS

cricardo.cruz@iimas.unam.mx