ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Máster

Sistema autónomo comprensivo de clasificación y consulta de contenidos multimedia

Universidad Internacional de Valencia

Por: Martín González Domínguez
Directora TFM: Yudith Cardinale
Co-Director: Ismael de Fez
Colaboradora: Vanessa Moscardó

De:
Planeta Formación y Universidades

14/01/2025

Contenidos

- Introducción y Objetivos
- Marco Tecnológico – Tecnologías empleadas, justificación técnica, requisitos tecnológicos y plataforma de despliegue.
- Estado del Arte – Compendio de estudios previos que han servido de base para este trabajo.
- Metodología utilizada.
- Desarrollo del proyecto – Funcionamiento, componentes principales y descripción de procesos.
- Evaluación de los resultados obtenidos.
- Conclusiones obtenidas y desarrollos futuros.

Universidad Internacional de Valencia

ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

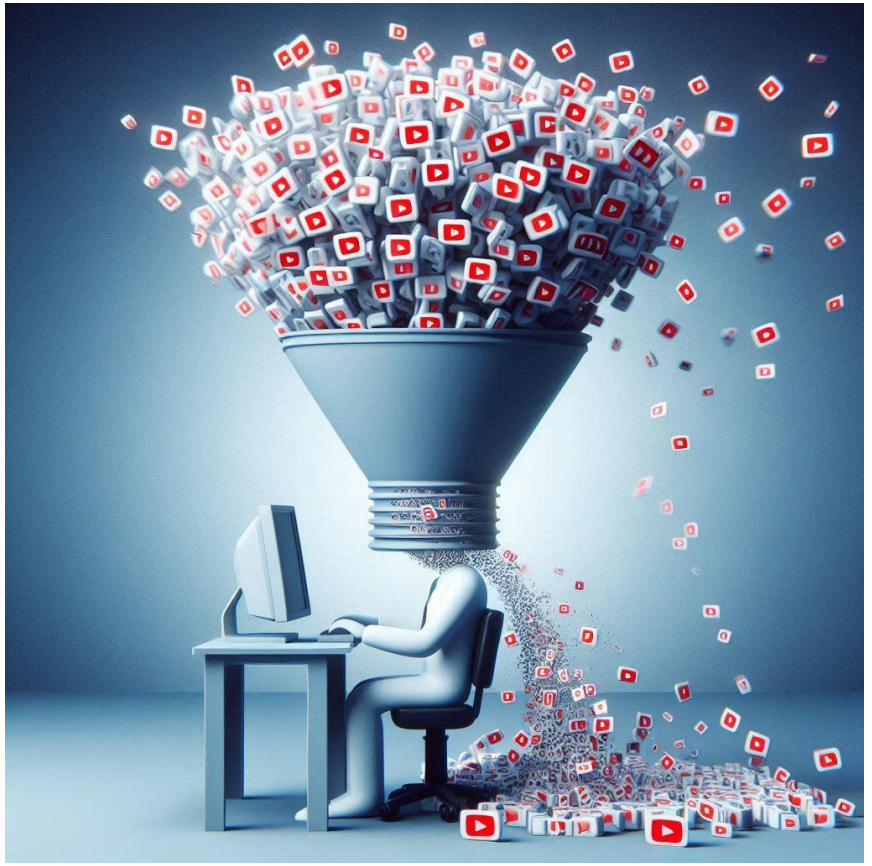
01
Introducción y Objetivos

Introducción

Existencia de múltiple cantidad de horas de contenido audiovisual en Internet (plataformas como Vimeo, Dailymotion...).

En el ámbito educativo existen otras plataformas de vídeo, lo que hace inviable el visionado de todo el material audiovisual.

Necesidad: Una herramienta que permita procesar vídeos educativos con el fin de facilitar la formación y consulta de contenidos.



Objetivos

- Objetivo general
  - Desarrollar un software que realice la visualización, comprensión y extracción de información de vídeos, que pueda ofrecer facilidades de consultas a los usuarios, sin que tengan que visualizar los vídeos completos.

- Objetivos específicos
  - Desarrollar una funcionalidad que **extraiga el audio y los fotogramas** de un vídeo.
  - Crear un proceso automático de **identificación de objetos** en cada uno de los fotogramas mediante CNN.
  - Diseñar e implementar un sistema que haga una **descripción de cada fotograma** mediante LLM descriptivo.
  - Desarrollar un **transcriptor de audio a texto** mediante un LLM, multiidioma.
  - Implementar un automatismo que permita **resumir** la transcripción obtenida.
  - Desarrollar una funcionalidad que **almacene** toda la información generada en una base de datos de vectores.
  - Desarrollar un proceso automático que permita **contestar** un conjunto de preguntas utilizando un LLM.

02
Marco Tecnológico

Tecnologías Empleadas

Lenguaje + IDE

Base de Datos

Chroma

Control de Versiones

GitHub

Contenedores

Gestor de LLM

Modelos LLM

- nomic-embed-text-v1
- Mistral
- LLAMA3
- LLaVA
- openai/whisper-large
- clu-ling/whisper-large-v2-spanish
- t5-small

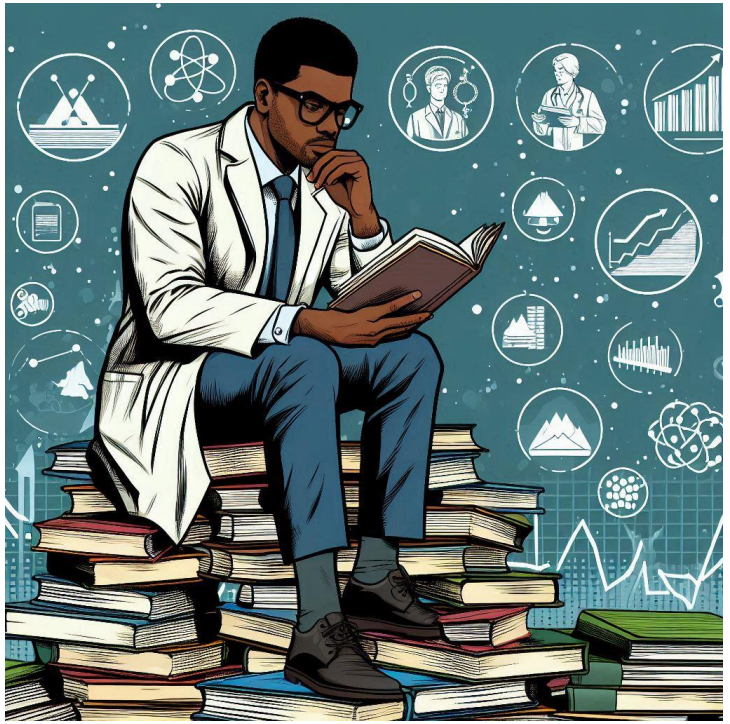
03
Estado del Arte

Trabajos Previos

Se han analizado 11 trabajos relacionados con este campo.

Conclusión: Cada uno de los artículos estudia una parte del proceso, pero no hay un estudio que englobe una solución global, ya que se centran únicamente en el audio o las imágenes del vídeo.

Propuesta: este trabajo abarca diferentes aspectos del soporte de la información, tanto audio como vídeo, consiguiendo una mayor cantidad de información y contexto en cada análisis.

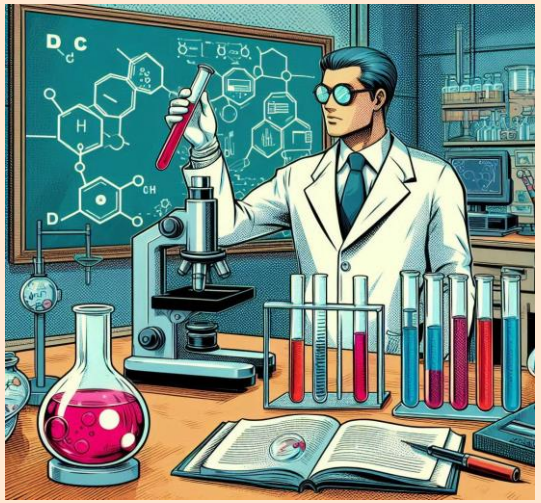


Universidad Internacional de Valencia

ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

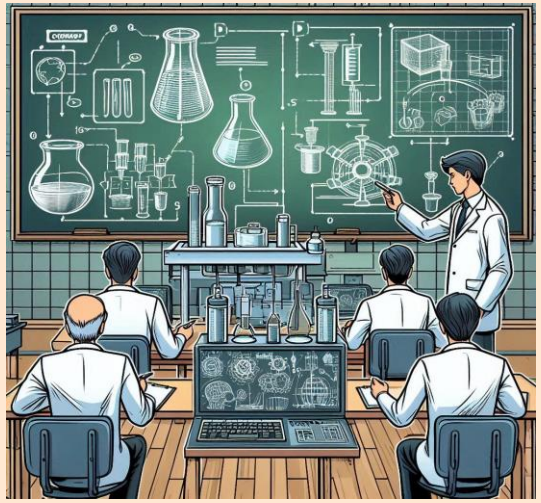
04
Metodología

Fases



Investigación preliminar

Evaluación de modelos



Diseño

Implementación y pruebas

05
Desarrollo del Proyecto

Requisitos Tecnológicos

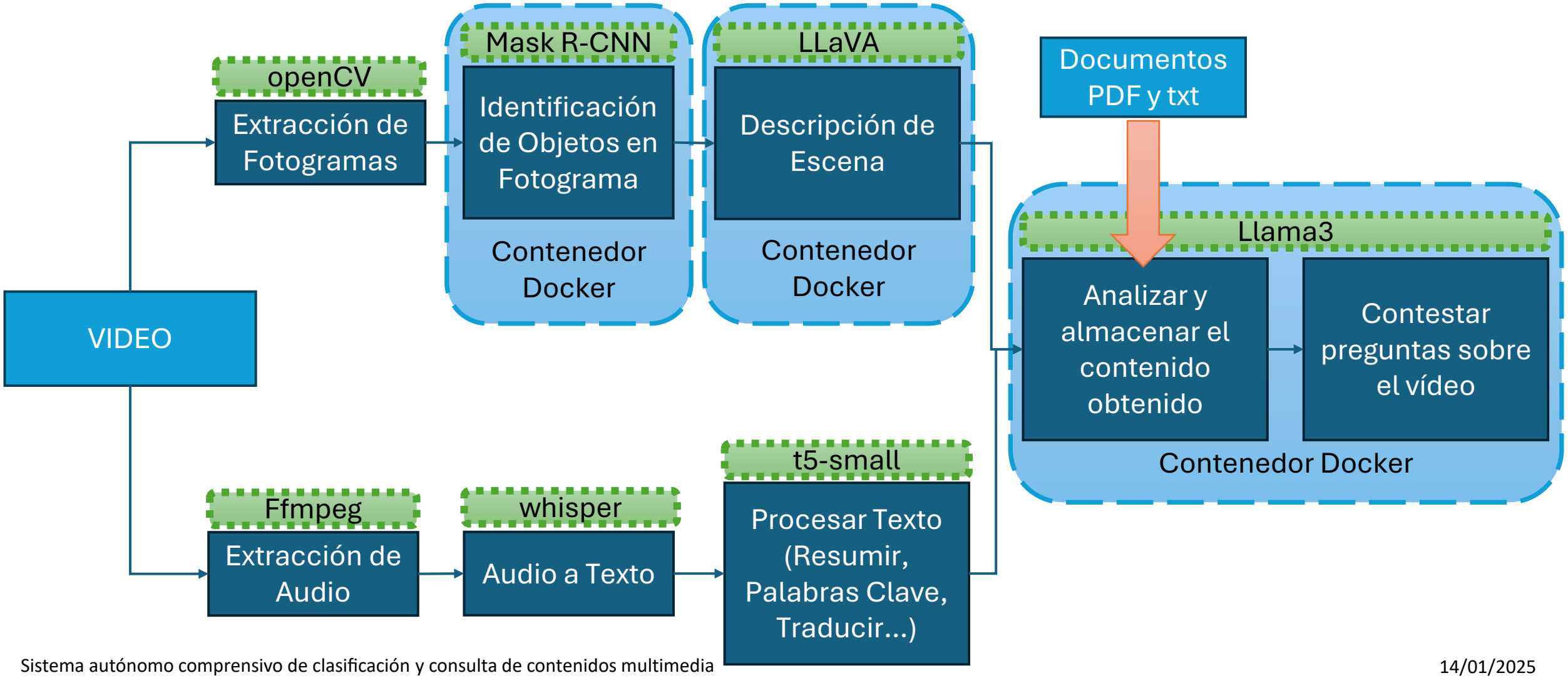
- Crear proceso que facilite el responder a preguntas formuladas por parte del usuario, partiendo de un vídeo y complementándolo con otros textos didácticos en texto o PDF.

- Facilitar la usabilidad, incluyendo una interfaz que sea agradable y sencilla de utilizar al usuario.

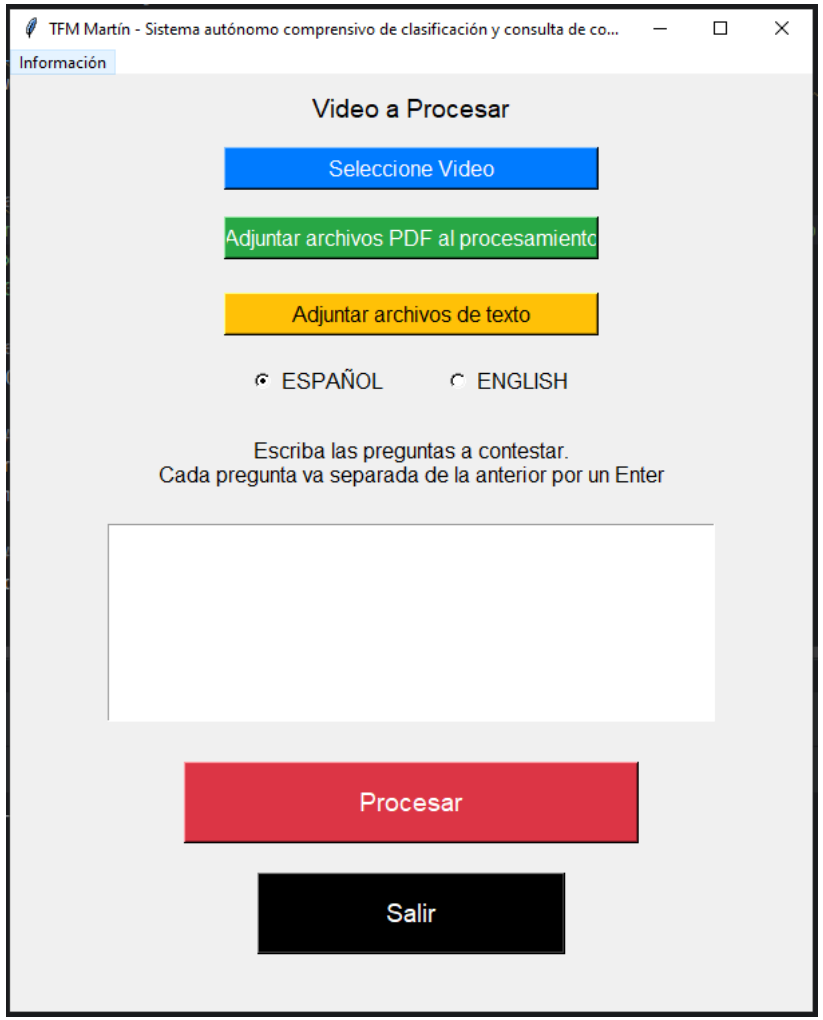
- Economizar costes, para permitir una implantación viable al ser un proyecto económico.

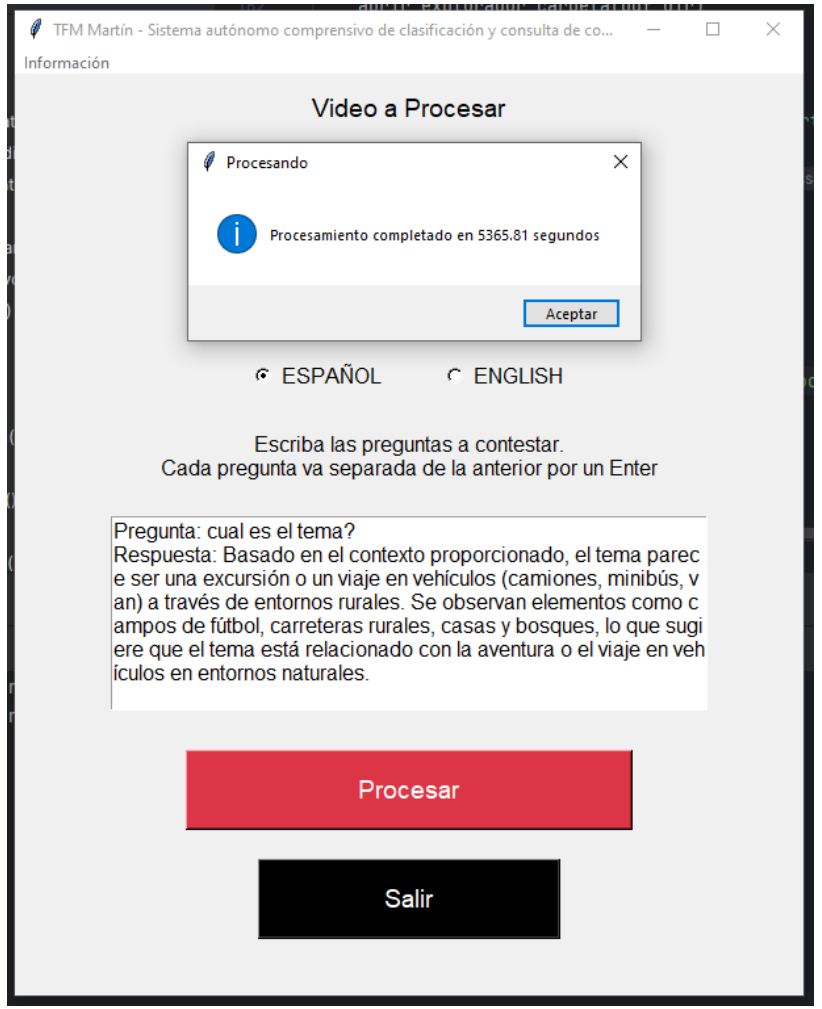
- Reducir la dependencia de Internet en el mayor grado posible.

Esquema de Funcionamiento

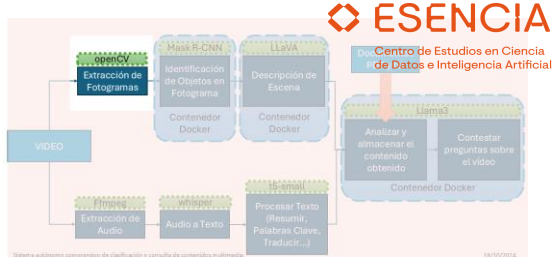


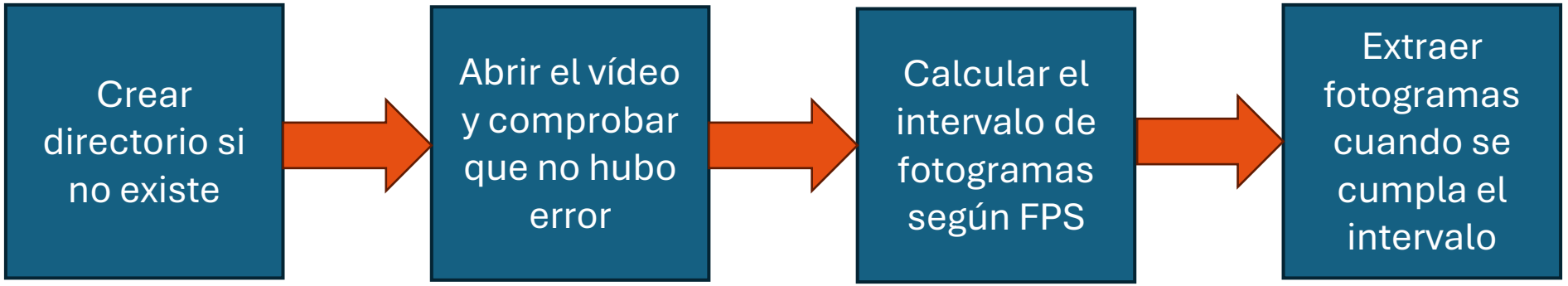
Interfaz Gráfica





Extracción de fotogramas

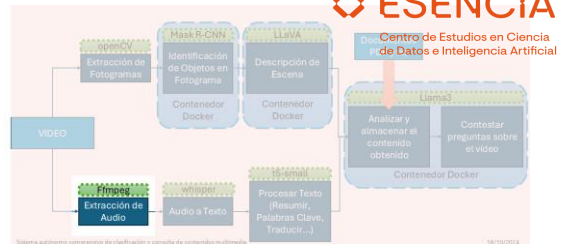




Universidad Internacional de Valencia

ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

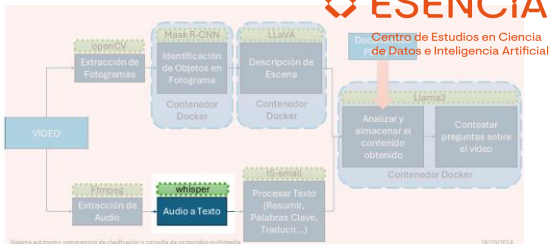
Extracción de audio



Universidad Internacional de Valencia

ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

Transcripción de audio



AUDIO extraído

Configurar CUDA si disponible

Seleccionar LLM según idioma configurado

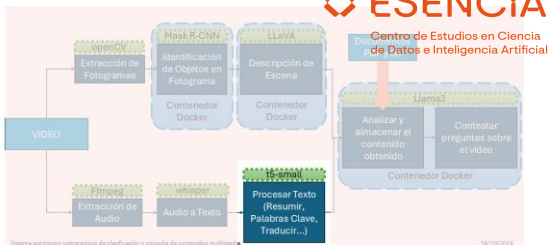
Dividir el audio en bloques de audio de 30 segundos

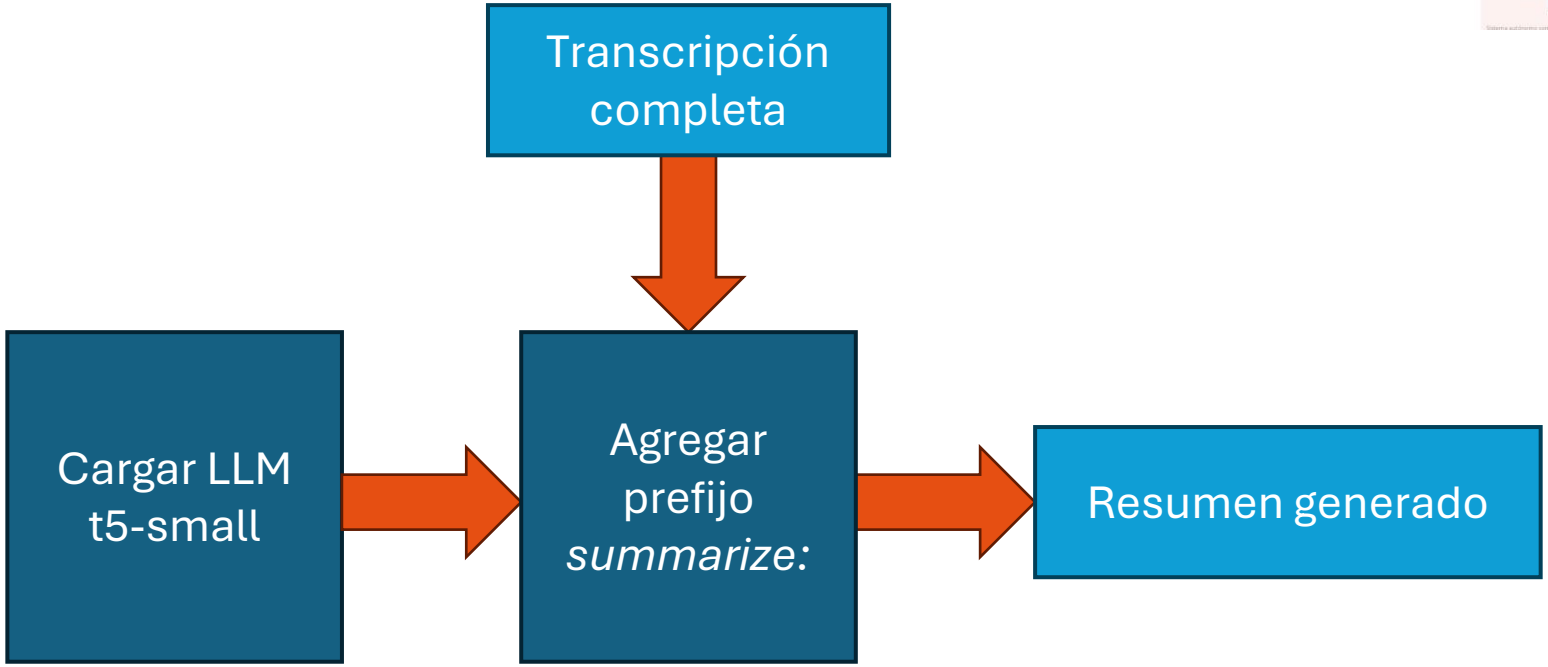
Transcribir cada segmento

Combinar transcripciones

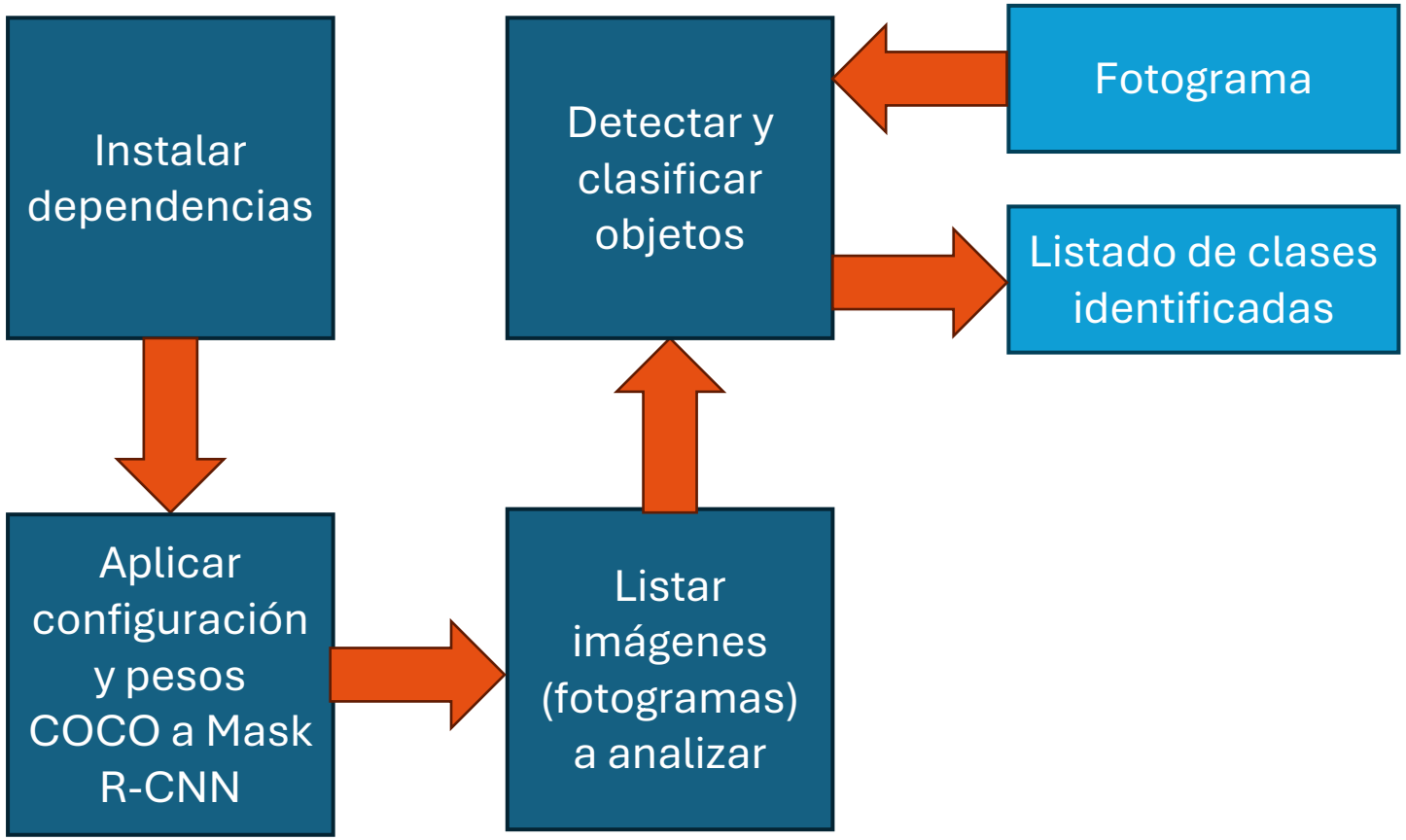
Transcripción completa

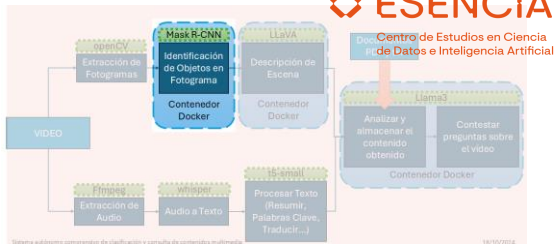
Procesamiento de texto (Generar resumen)

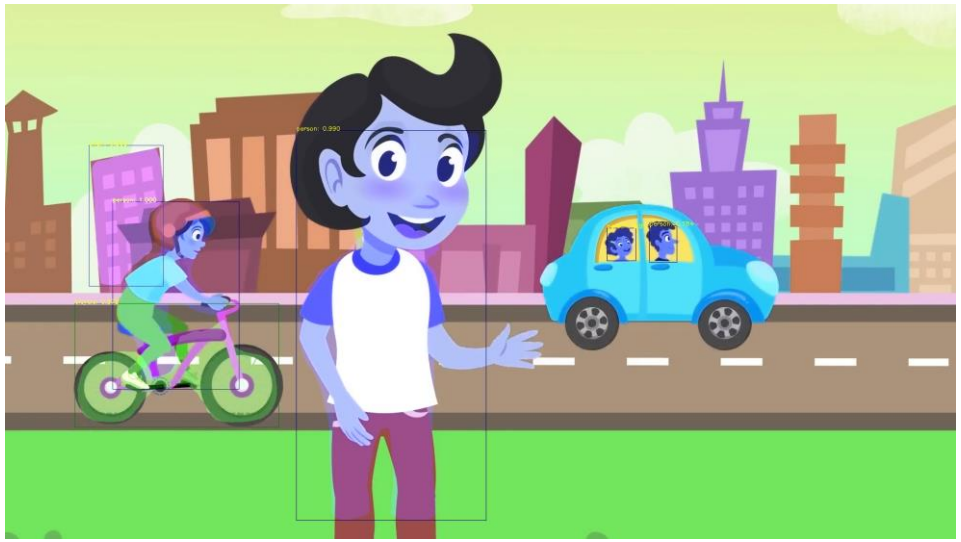




Identificación de objetos



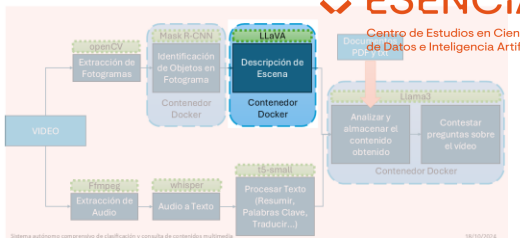


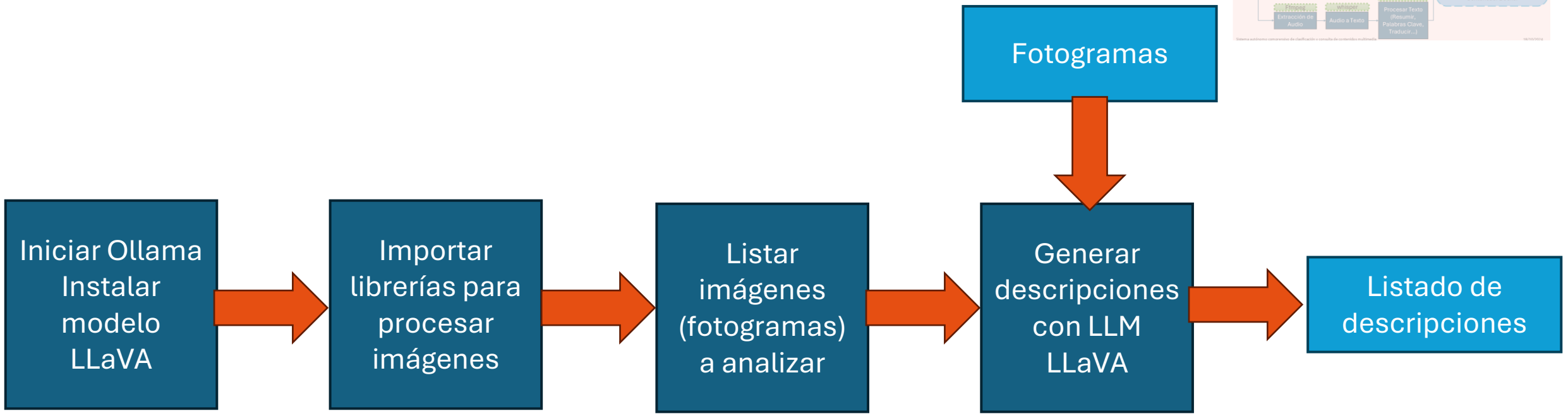


Universidad Internacional de Valencia

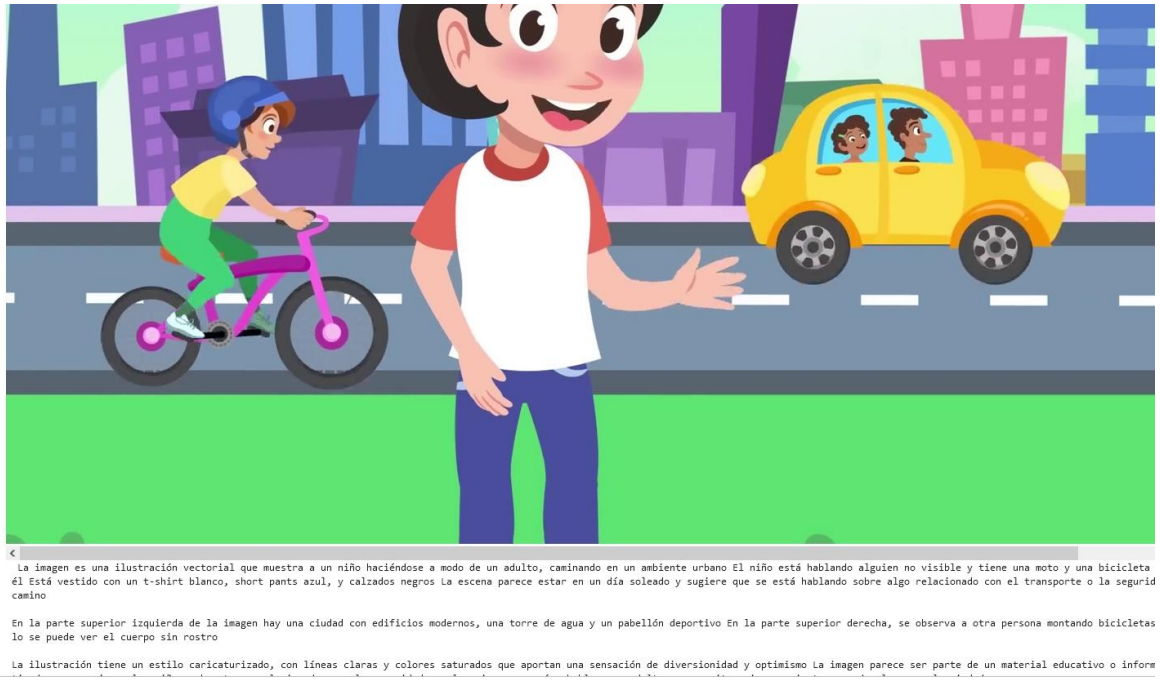
ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

Descripción de la escena





Descripción de la escena - Muestra

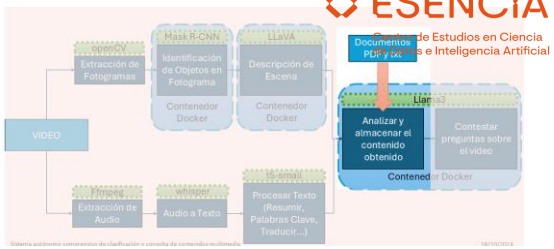


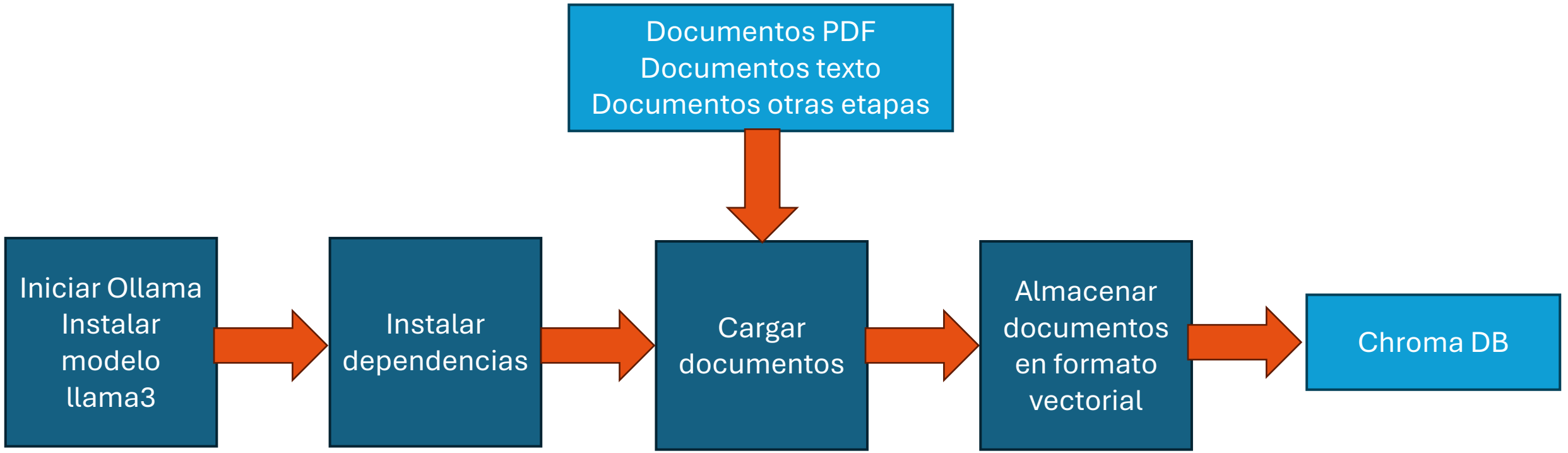
Descripción obtenida:

“La imagen es una ilustración vectorial que muestra a un niño haciéndose a modo de un adulto, caminando en un ambiente urbano El niño está hablando alguien no visible y tiene una moto y una bicicleta detrás de él Está vestido con un t-shirt blanco, short pants azul, y calzados negros La escena parece estar en un día soleado y sugiere que se está hablando sobre algo relacionado con el transporte o la seguridad en el camino.

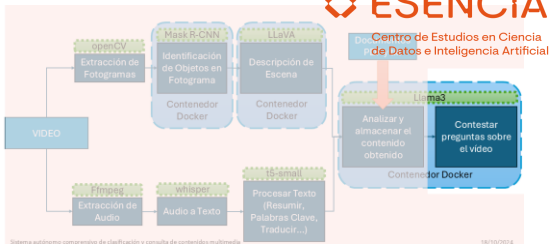
En la parte superior izquierda de la imagen hay una ciudad con edificios modernos, una torre de agua y un pabellón deportivo En la parte superior derecha, se observa a otra persona montando bicicletas, pero solo se puede ver el cuerpo sin rostro.”

Analizar y almacenar el contenido generado





Contestar preguntas del usuario sobre el vídeo



Iniciar Ollama
Instalar modelo llama3

Instalar dependencias

Configurar RAG. Establecer cuantas variaciones por pregunta (*MultiQueryRetriever*)

Preguntas del usuario Chroma DB

Recuperar respuestas (contexto)

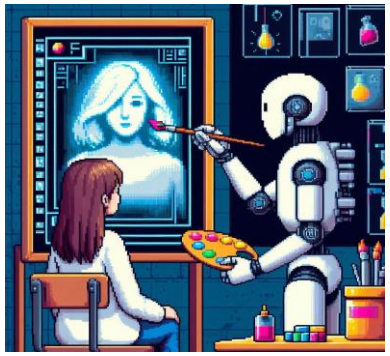
Preguntas de usuario

Generar respuestas coherentes mediante ChatOllama

Listado de preguntas y respuestas

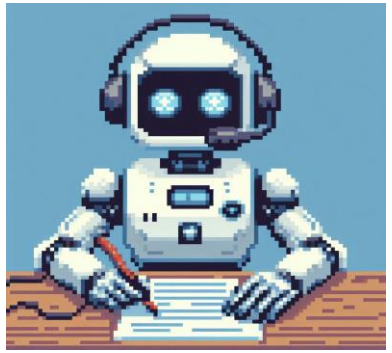
Etapas y LLM

Extracción de Fotogramas

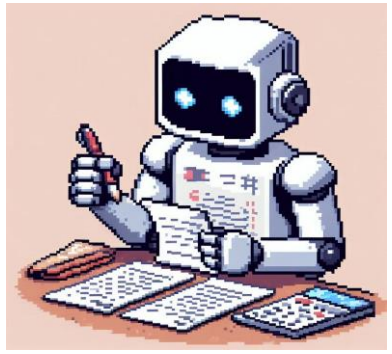


Identificación de Objetos en Fotograma

Descripción de Escena



Extracción y Transcripción de Audio



Procesar Texto (Resumir)

Analizar y almacenar

Contestar preguntas del usuario

Devolver informe con respuestas

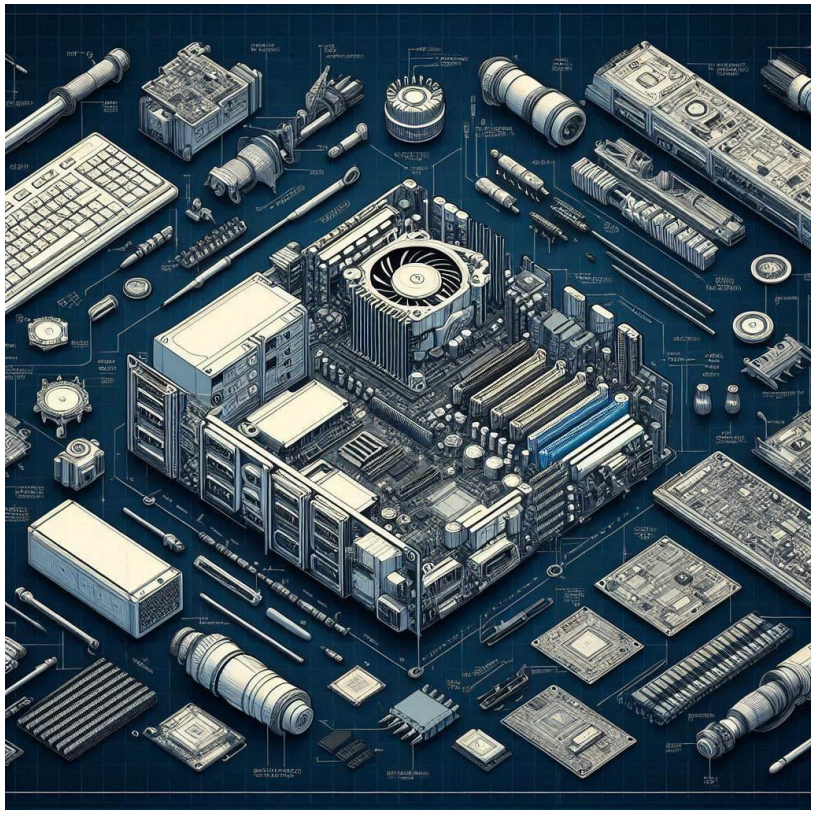
Justificación de la fiabilidad del proceso

Etapa	Modelo	Grado de Fiabilidad
VIDEO	-	-
Extracción de fotogramas	-	-
Identificación de Objetos en Fotograma	MaskRCNN	Variable
Descripción de Escena	LLaVA	75.3
Extracción de Audio	-	-
Transcripción (Español)	clu-ling/whisper-large-v2-spanish	Loss: 0.1466 Wer: 0.0855
Transcripción (Inglés)	openai/whisper-large	Wer: 0.03 - 0.05
Resumen de texto	t5-small	83.28
Analizar textos y almacenar	-	-
Contestar preguntas	Llama3	68.4

06
Evaluación y Resultados

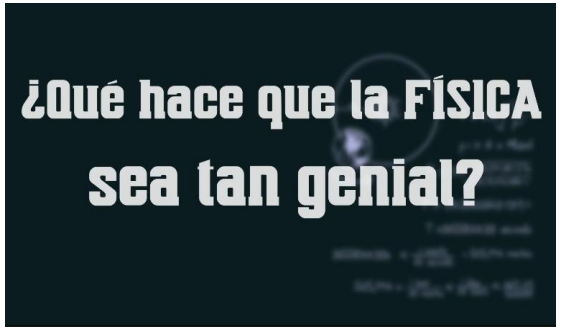
Características del *benchmark* utilizado

Componente	Nombre
Procesador	Intel Core i7-14700K 3.4/5.6GHz
Memoria RAM	32GB DDR5
Tarjeta Gráfica	Asus ROG Strix RTX 2060
Tarjeta Gráfica (VRAM)	6GB GDDR6
Tecnología Disco Duro	M.2 PCI Express



Tiempos obtenidos

Vídeo 1 – “Los medios de transporte para niños | Transportes terrestres, acuáticos y aéreos” (6:03 minutos)



Vídeo 2 – “¿Qué es la FÍSICA? | En 2 minutos” (2:20 minutos)

Vídeo 3 – “¿como funciona un motor de inducción?” (5:18 minutos)

Vídeo 4 – “¡VAMOS AL AEROPUERTO! AVIONES, PILOTOS Y MÁS | Vídeos educativos para niños y niñas | Papumbas” (3:15 minutos)

Etapa	Vídeo 1 Tiempo (segundos)	Vídeo 2 Tiempo (segundos)	Vídeo 3 Tiempo (segundos)	Vídeo 4 Tiempo (segundos)
Extracción fotogramas	26.80	3.36	12.96	24.58
Descripción de fotogramas	3623.26	442.62	509.57	480.05
Clasificación de fotogramas	551.47	1744.04	3108.24	1966.54
Transcripción de fotogramas	817.61	410.01	732.39	350.83
Resumen transcripción	3.86	4.04	3.86	4.25
Contestar Preguntas (RAG)	263.21	173.03	207.04	195.38
TOTAL	5365.81	2906.28	4660.74	3248.37

Comparación respuestas – Sin/Con documentación aportada

Sin aportar PDF

- Pregunta: cual es el tema?
- Respuesta: "*Según el contexto, el tema parece ser relacionado con la seguridad policial en un entorno familiar y divertido, como el del viaje. Esto se infiere de los elementos presentes en las imágenes, como la presencia de personajes animados vestidos de policía, uniformes de viaje y accesorios similares a juegos de roles.".*

- Pregunta: cuantos medios de transporte aparecen?
- Respuesta: "*Según la descripción del contexto, se mencionan los siguientes medios de transporte:*
  - *Aviones (mencionados varias veces)*
  - *Automóviles*
  - *Camionetas (probablemente destinados a la logística del aeropuerto)*

  En total, hay 4 medios de transporte mencionados.".

Aportando PDF sobre transportes

- Pregunta: cual es el tema?
- Respuesta: "*El tema es "Transporte y Desarrollo".*".

- Pregunta: cuantos medios de transporte aparecen?
- Respuesta: "*Según el contexto, aparecen dos medios de transporte:*
  1. *Aviones (aeronaves)*
  2. *Nave espacial (no especifica qué tipo de nave espacial, pero se menciona la órbita lunar)*".

Comparación respuestas – Sin/Con documentación aportada

Sin aportar PDF

- Pregunta: cuantas personas?
- Respuesta: *“Según el contexto, se pueden contar las siguientes personas:*
  - *En la imagen "fotogramas/frame_97.jpg", hay varias personas sentadas en asientos a lo largo del pasillo, y algunas de ellas están de pie. No puedo dar un número exacto porque no hay una cuenta precisa.*
  - *La imagen "fotogramas/frame_98.jpg" muestra solo al piloto y posiblemente a la co-piloto, pero no estoy seguro.En resumen, hay varias personas en la primera imagen, pero no puedo dar un número exacto. En la segunda imagen, hay solo dos personas: el piloto y posiblemente la co-piloto.*

  En resumen, hay varias personas en la primera imagen, pero no puedo dar un número exacto. En la segunda imagen, hay solo dos personas: el piloto y posiblemente la co-piloto.”.

Aportando PDF sobre transportes

- Pregunta: cuantas personas?
- Respuesta: “*La respuesta a esta pregunta es cero. No hay ninguna mención a personas en el contexto proporcionado. El texto se refiere a transporte, infraestructuras, políticas públicas y regulaciones relacionadas con la actividad aérea, pero no contiene información sobre personas.*”.

07
Conclusiones y Desarrollos Futuros

Conclusiones

- Se han **cumplido los objetivos marcados**, e incluso se han podido ampliar con la adición de poder incorporar documentos de texto y PDF para ser procesados y ser tenidos en cuenta a la hora de generar respuestas.

- Se ha creado un diseño que puede **funcionar de forma autónoma**, a excepción del servidor de almacenamiento de vectores, que no es de funcionamiento local, pero están trabajando para que lo pueda ser.

- Se ha podido comprobar de forma empírica que **las respuestas entre modelos varían**, siendo los modelos más pequeños y de menor número de parámetros los que dan unas respuestas más generalistas, como el caso de t5-small, que da una visión tan general del texto procesado que se centra en resumir, frente a modelos como llama3 que dan respuestas más concretas.

- A mayor **cantidad de parámetros** mayores son las necesidades de hardware, por contrapartida las respuestas son de mayor precisión, mejor calidad y están mejor redactadas. No se han podido probar modelos superiores a 7 billones de parámetros por limitaciones del hardware local.

- Se ha podido observar que el **tiempo** depende de la cantidad de información a analizar y del grado de complejidad del modelo, siendo los modelos más sencillos los más rápidos, como se puede apreciar en el tiempo empleado para realizar un resumen (t5-small) frente al tiempo de describir un fotograma (LLaVA).

- El **diseño modular** favorece poder intercambiar y ampliar el diseño del proceso, pudiendo variarlo o ampliarlo según se fuese necesitando.

- Cuanto mayor es el modelo y más parámetros tiene, necesita un **hardware cada vez más caro** y una mejor refrigeración, lo que repercute en el consumo del centro de datos o en el añadir refrigeración a la sala donde se encuentre el PC.

Desarrollos futuros

- Actualización de modelos: A medida que se vayan desarrollando LLM nuevos se pueden ir actualizando los modelos existentes con modelos que hagan las mismas funciones.

- Adición de nuevos tipos de archivo: Se puede añadir otros tipos de archivo como presentaciones PowerPoint, archivos de Excel, etc.

- Proceso por lotes: Podría ser interesante evitar que el procesamiento esté limitado a un único vídeo y permitir que se procesen una serie de vídeos, y así tener un volumen mayor de información recogida.

- Acumulación de datos: Se podría hacer que el vaciado de la base de datos fuese opcional, para ir acumulando colecciones de datos sobre ciertos temas, con el fin de hacer una base de datos vectorial más versátil y evitar el procesar algún vídeo varias veces en caso de hacerse preguntas de la misma temática en ejecuciones distintas.

La modularidad del proyecto favorece la actualización de la solución desarrollada a través de la actualización de componentes, LLM y librerías, permitiendo la adaptación de la solución a las nuevas tecnologías que surjan en el futuro.

Disponibilidad de Código

El código de este proyecto se encuentra disponible en el repositorio GitHub

https://github.com/martinVIU/TFM

Este Trabajo Fin de Máster ha sido financiado por el Proyecto
“VIU24004 - UbiqDataEnergyAware: Ciencia de Datos para ambientes de Computación Ubicua conscientes del consumo energético" de la Universidad Internacional de Valencia

¡Gracias por su atención!

ESENCIA
Centro de Estudios en Ciencia de Datos e Inteligencia Artificial

viu

Universidad Internacional de Valencia

De:
Planeta Formación y Universidades