

# Cursos Extraordinarios

## Verano 2024

**“Inteligencia Artificial y Grandes Modelos de Lenguaje: Funcionamiento, Componentes Clave y Aplicaciones”**

**Zaragoza, del 3 al 5 de julio**

# MODELOS DE LENGUAJE MULTIMODALES

## RECUPERACION DE INFORMACIÓN MULTIMODAL BASADA EN ESPACIOS SEMÁNTICOS

Los sistemas tradicionales de IR (Recuperación de Información) se basan en el indexación de texto completo de los documentos en sí mismos o en metadatos que describen los datos en el caso de audio, imágenes o video.

**Nuestro problema:** solo contenido de video, **sin metadatos**

**¿Cómo representamos el contenido solo de video?**

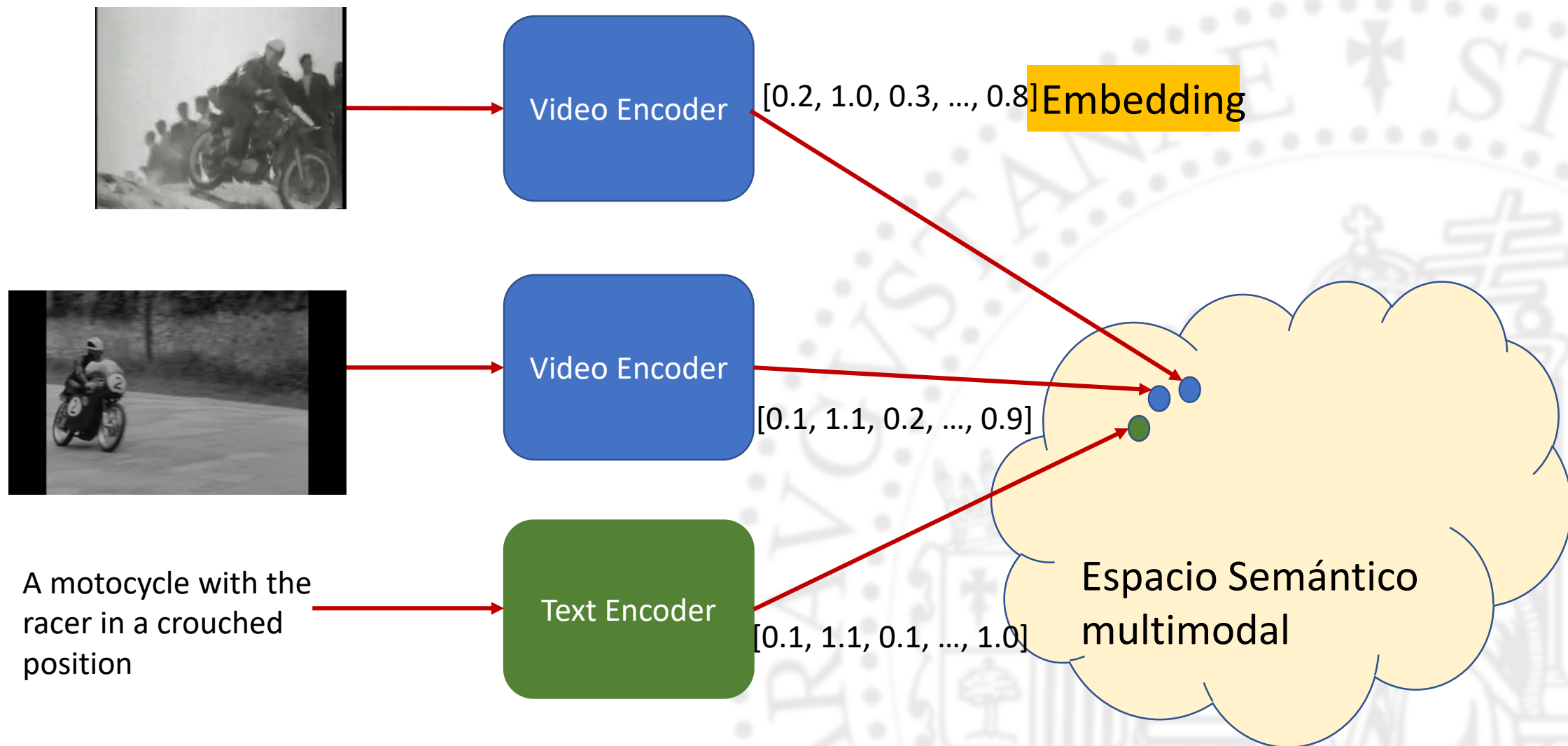
- Consumir miles de horas de documentalistas creando metadatos.
- Utilizar nuevos sistemas de descripción automática de imágenes para crear metadatos. Estos han tenido grandes avances en los últimos años, pero todavía enfrentan varios desafíos (alucinaciones, flexibilidad en la descripción, sesgos, ...) pero podrían ser una opción en el futuro próximo utilizando VLM (Modelos de Lenguaje Visuales) como Llava, GPT-4v, PaLlGemma, ...
- Usar una representación conjunta en el espacio semántico.

# MODELOS DE LENGUAJE MULTIMODALES

## RECUPERACION DE INFORMACIÓN MULTIMODAL BASADA EN ESPACIOS SEMÁNTICOS

- El **objetivo** es aprender un espacio semántico conjunto que pueda capturar las relaciones inherentes entre ambas modalidades (texto y video), también conocido como modelo de espacio vectorial multimodal.
- Los modelos de espacio vectorial multimodal representan el significado (texto o video) como puntos en espacios vectoriales de alta dimensionalidad, también conocidos como **espacios semánticos multimodales**.
- Cada "**punto**" en el espacio semántico multimodal se conoce como "**embedding**".
- El texto y el video se representan mediante **embeddings**.
- Las piezas de información de diferentes modalidades que representan el mismo concepto semántico estarán "**cerca**" en el espacio semántico multimodal.
- La noción de "**similaridad**" o "**proximidad**" entre conceptos se reduce a la "distancia" entre los vectores de representación en el espacio vectorial.

# MODELOS DE LENGUAJE MULTIMODALES



# MODELOS DE LENGUAJE MULTIMODALES

## Three important definitions

**Semantic Shot:** A consecutive sequence of nearby video frames embeddings in semantic space.

**Semantic Scene:** A consecutive sequence of nearby semantic shots in semantic space.

**Semantic Class:** A set of nearby video frames embeddings in semantic space. Classes are like scenes but without considering the time position.

The multimodal semantic space enables the following search functionalities:

1. Text-to-Video search, which retrieves **semantic shots** based on text queries
2. Image-to-Video search, which retrieves **semantic shots** that are “similar” to the image query
3. Text+Image-to-Video search, which retrieves **semantic shots** based on a combination of text and image query (augmented retrieval)

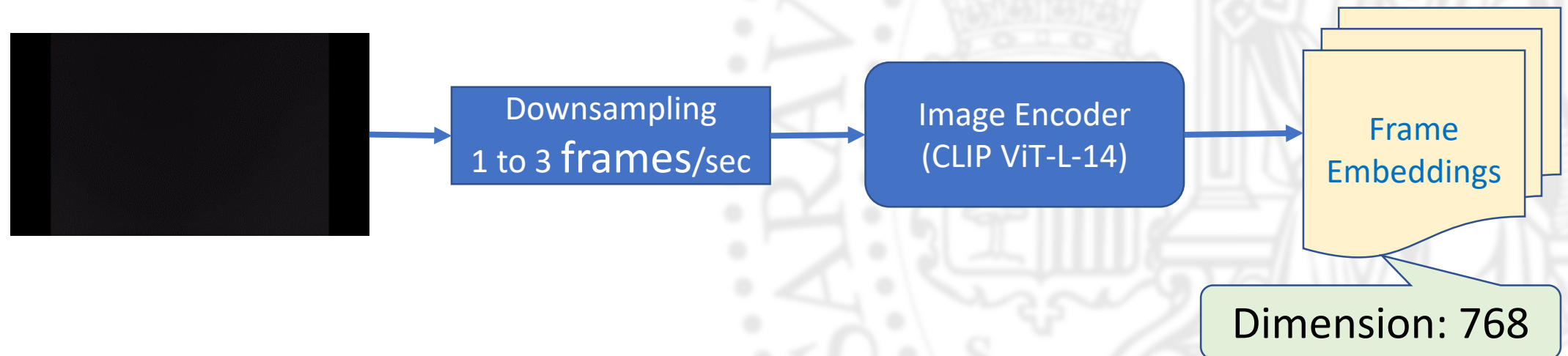
# MODELOS DE LENGUAJE MULTIMODALES

## VIDEO ANALISYS

INPUT: VIDEO (mp4 format)

1. STEP: DONWSAMPLING VIDEO FRAME RATE (1 TO 3 FRAMES/SEC)
2. STEP: COMPUTE FRAME EMBEDDINGS USING A IMAGE ENCODER (CLIP ViT-L-14)

OUTPUT: FRAME EMBEDDINGS (768 DIMENSIONAL VECTORS)



# MODELOS DE LENGUAJE MULTIMODALES

## SEMANTIC ANALYSIS

INPUT: FRAME EMBEDDINGS (768D x (time video length in seconds x 3))

1. COARSE GROUPING: Find Clusters in the Semantic Space → SEMANTIC CLASSES
2. CLASS DIARIZATION: Distribute Semantic Classes in the Time-Line → SEMANTIC SCENES
3. FINE GROUPING: Find Clusters in the Semantic Scenes → SEMANTIC SHOTS
4. SHOT EMBEDDINGS: Use Embedding Mean Average over Semantic Shots

OUTPUT: SHOT EMBEDDINGS (768D x (#semantic shots))



# MODELOS DE LENGUAJE MULTIMODALES

## Semantic hierarchy

Clustering over Semantic Space  
(coarse grouping)

Semantic Class

Diarization over Semantic Classes

Semantic Scene

Semantic Scene

Clustering over Semantic Scenes  
(fine grouping)

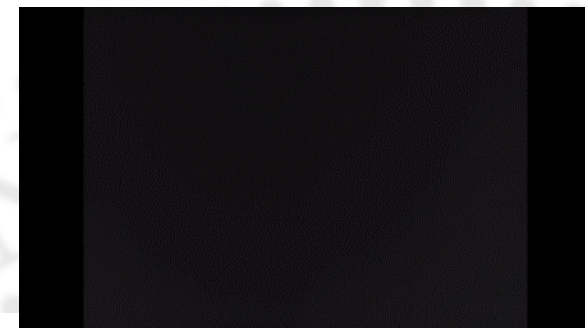
Semantic Shot

Semantic Shot

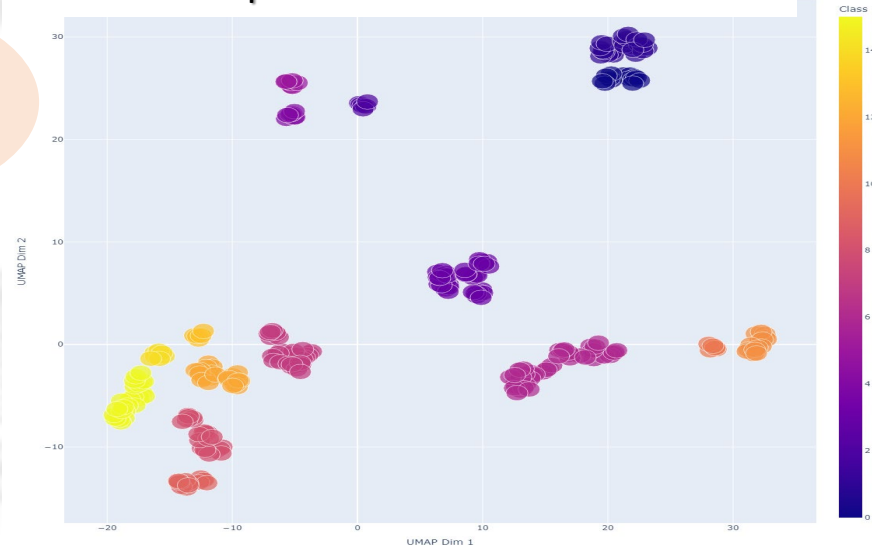
Mean Pooling

Shot Embedding

Shot Embedding

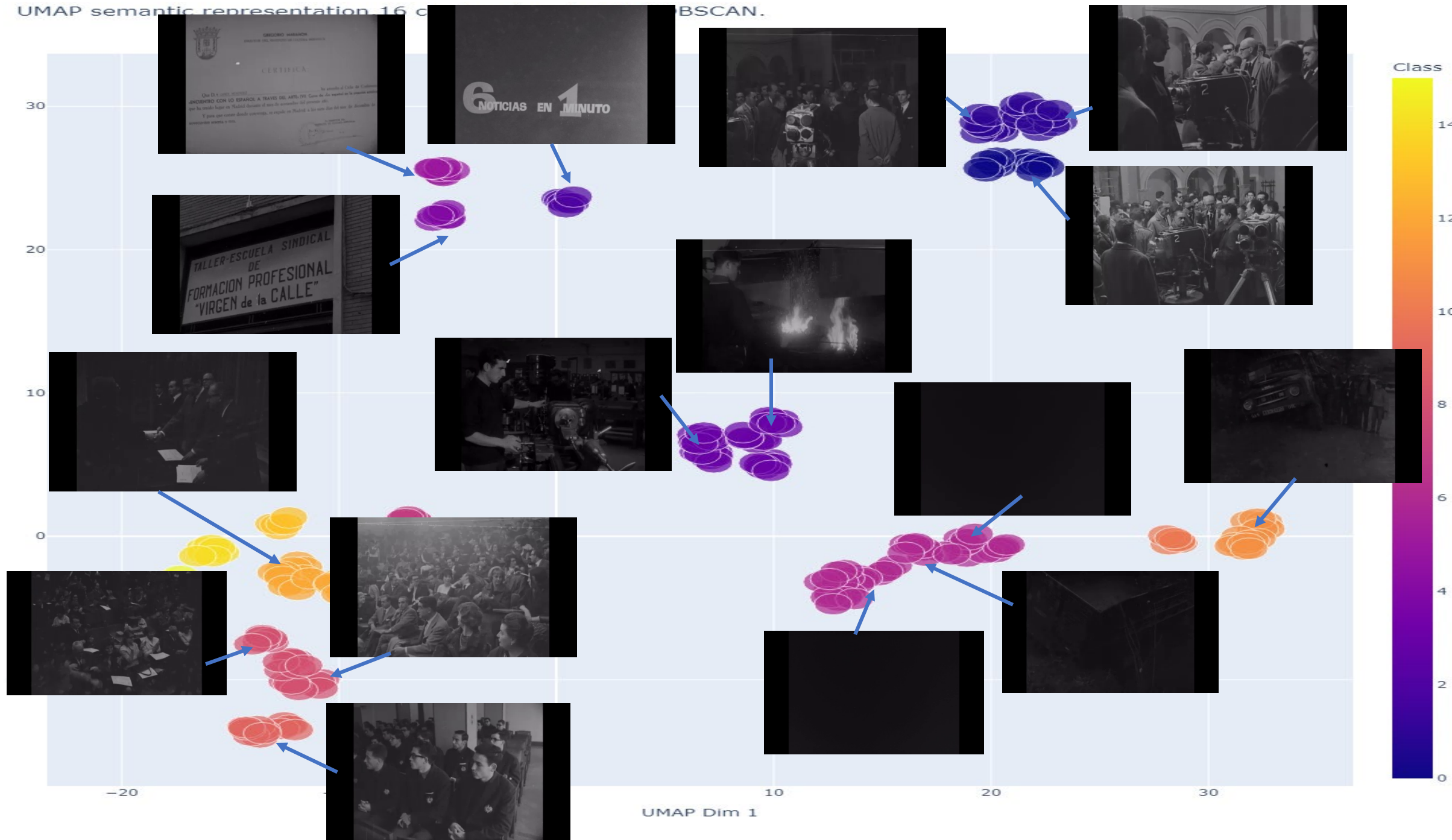


Semantic Space with 16 Semantic Classes

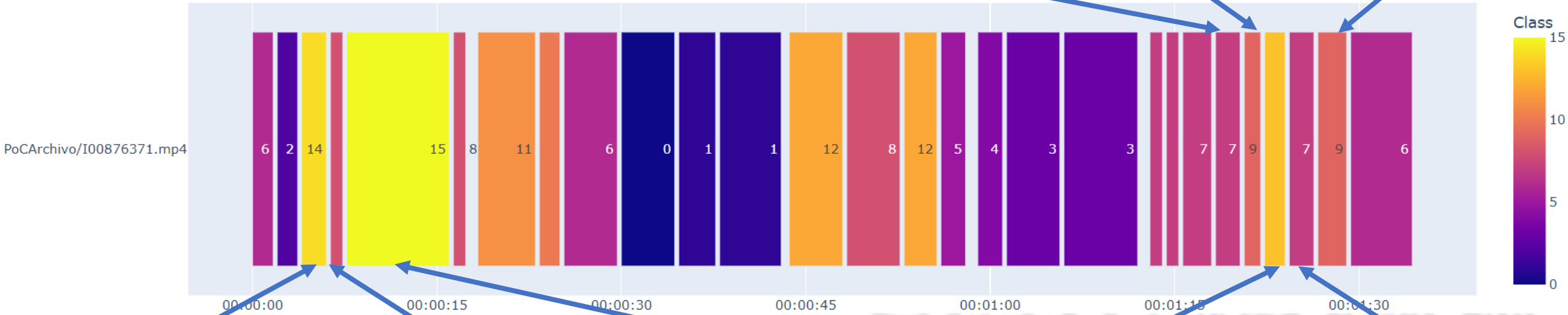




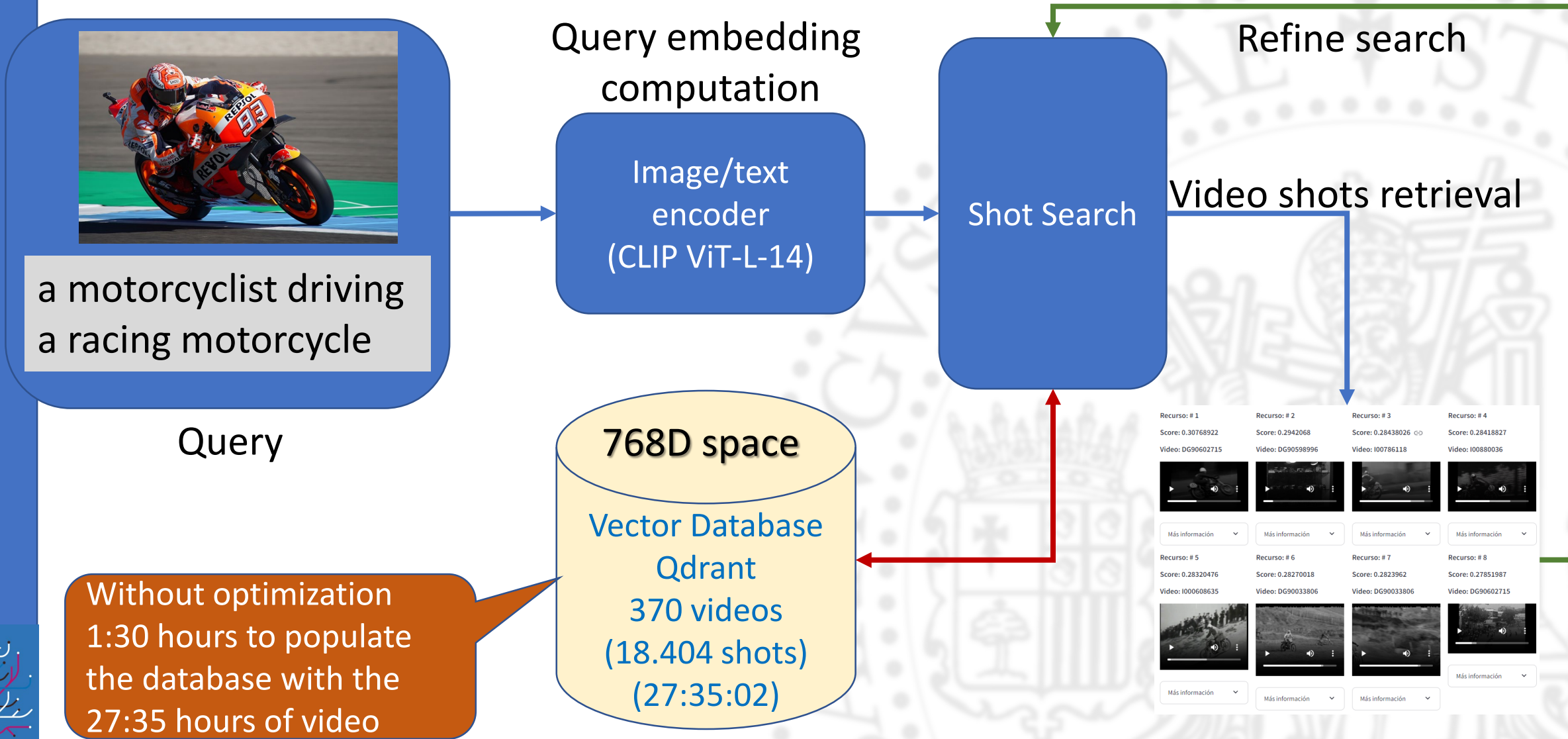
UMAP Dim 2



Video shots Timeline (23 scenes y 28 shots)



# MODELOS DE LENGUAJE MULTIMODALES




# MULTIMODAL INFORMATION RETRIEVAL BASED ON SEMANTIC SPACES

Search examples using the time-coded logging of the footage performed by a documentalist:

- DG90602601.mp4 00:02:22.205 - 00:02:41.861 BOTELLAS CON PRODUCTOS QUIMICOS EN EL LABORATORIO, EL TRABAJADOR MEZCLA EL TINTE CON AGUA EN UN TARRO Y REMUEVE. INTRODUCE EL LIQUIDO EN LA MAQUINA PARA TEÑIR A TRAVES DE UN EMBUDO (BOTTLES WITH CHEMICALS IN THE LABORATORY, THE WORKER MIXES THE DYE WITH WATER IN A JAR AND STIRS. INTRODUCE THE LIQUID INTO THE DYEING MACHINE THROUGH A FUNNEL)
- DG90602601.mp4 00:07:32.738 - 00:07:40.615 CUATRO BIDONES PEQUEÑOS DE CLENSEOL CON LA PALABRA "CASABLANCA" (FOUR SMALL DRUMS OF CLENSEOL WITH THE WORD "CASABLANCA")
- DG90318484.mp4 00:00:00.000 - 00:02:37 CHICOS Y CHICAS JOVENES EN UNA DISCOTECA O SALA DE BAILE EN LOS AÑOS 60. LOS CHICOS VESTIDOS CON TRAJE Y CORBATA Y LAS CHICAS CON MINIFALDA Y MEDIAS DE RED. UNOS ESTAN SENTADOS TOMANDO UNA BEBIDA Y CHARLADO, OTROS DE PIE CHARLANDO EN GRUPOS Y OTROS BAILANDO EN ESTILO LIBRE. (YOUNG BOYS AND GIRLS IN A NIGHTCLUB OR DANCE ROOM IN THE 60S. THE BOYS DRESSED IN SUITS AND TIES AND THE GIRLS IN MINISKIRT AND FISHNET STOCKINGS. SOME ARE SITTING HAVING A DRINK AND CHATTING, OTHERS STANDING CHATTING IN GROUPS AND OTHERS DANCING FREESTYLE.)
- DG90602864.mp4 00:01:24.164 - 00:01:27.820 UN HOMBRE CON UNA CAMARA DE CINE PROFESIONAL EN EL TELESILLA. (A MAN WITH A PROFESSIONAL FILM CAMERA ON THE CHAIRLIFT)
- I000608689.mp4 00:00:34.840 - 00:00:49.200 TRABAJADORES CAVANDO ZANJAS PARA LA POSTERIOR INTRODUCCION DEL TRONCO. (WORKERS DIGGING TRENCHES FOR THE LATER INTRODUCTION OF THE TRUNK)

# BuscaEscenas: rtvePoCArchivo3

Selección de la base de datos de vídeos:  rtvePoCArchivo3

Parámetros de búsqueda 

Limpia la consulta

Consulta sobre el campo de imagen  

Introduzca la descripción textual de la escena:

Escribe tu consulta a la base de datos de escenas

y/o suba una imagen



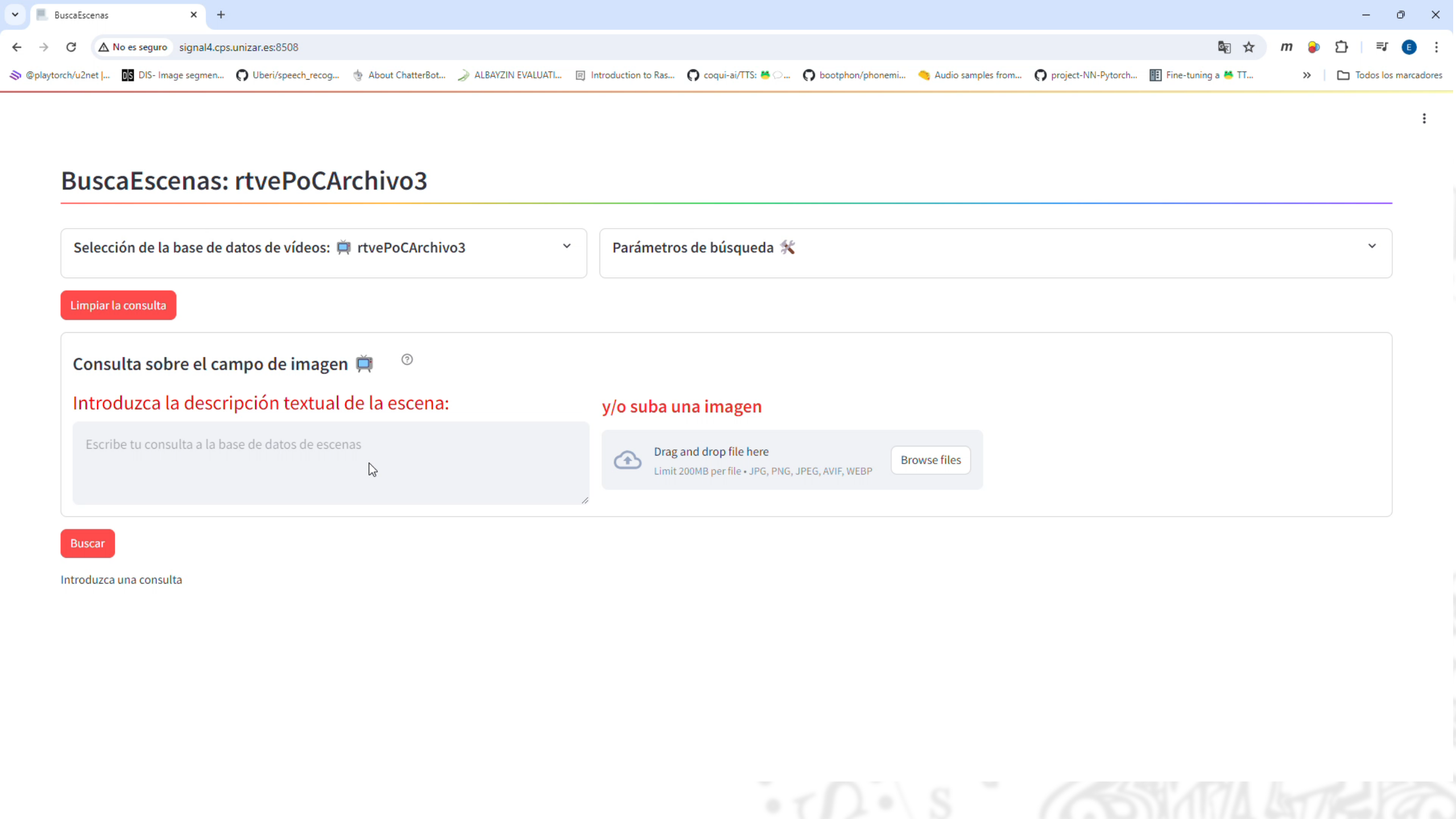
Drag and drop file here

Limit 200MB per file • JPG, PNG, JPEG, AVIF, WEBP

Browse files

Buscar

Introduzca una consulta



# BuscaEscenas: rtvePoCArchivo3

Selección de la base de datos de vídeos: rtvePoCArchivo3

Parámetros de búsqueda

Limpiar la consulta

Consulta sobre el campo de imagen

Introduzca la descripción textual de la escena:

Escribe tu consulta a la base de datos de escenas

y/o suba una imagen



Drag and drop file here


Limit 200MB per file • JPG, PNG, JPEG, AVIF, WEBP

Browse files

Buscar

[Introduzca una consulta](#)

# BuscaEscenas: rtvePoCArchivo3

Selección de la base de datos de vídeos:  rtvePoCArchivo3

Parámetros de búsqueda 

Limpia la consulta

Consulta sobre el campo de imagen  

Introduzca la descripción textual de la escena:

Escribe tu consulta a la base de datos de escenas

y/o suba una imagen



Drag and drop file here

Limit 200MB per file • JPG, PNG, JPEG, AVIF, WEBP

Browse files

Buscar

Introduzca una consulta

# MODELOS DE LENGUAJE MULTIMODALES

## PERFORMANCE IN VIDEO RETRIEVAL USING A BROAD DESCRIPTION OF THE VIDEO CONTENT AS :

FLAMENCO SHOW AND FEMALE GYMNASTICS EXHIBITION. People dancing flamenco. Women doing Artistic Gymnastics. (I00546653.mp4) (6)

DESCENT OF THE SEGURA RIVER. people rowing. people in boat (I00786279.mp4) (3)

FACTORY. People working in a Factory. (I00786311.mp4) (-)

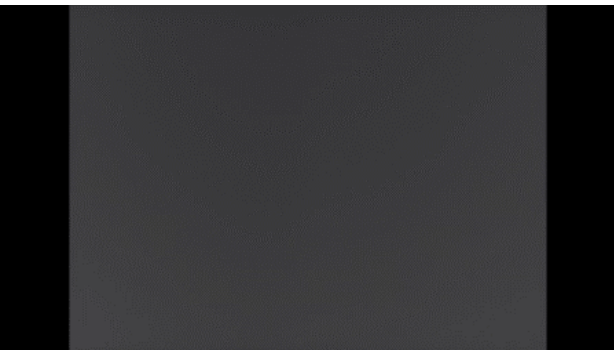
CHRISTMAS LIGHTING IN ALICANTE. Christmas lights. Christmas decoration (I00876540.mp4) (15)

CIVIL AUTHORITY VISIT. FOOTBALL TEAMS. WAITER CAREER. INAUGURATION OF THE HEADQUARTERS OF "TEACHING SCHOOLS". Waiters. Women dancing (I00786332.mp4) (42)

R@1	R@2	R@5	R@10	R@20	R@50
22%	31%	45%	61%	73%	85%



# FACTORY. People working in a Factory. (I00786311.mp4)



Recurso: # 1

Score: 0.26436812

Video: DG90347831



Más información

Recurso: # 2

Score: 0.2605429

Video: DG90347831



Más información

Recurso: # 3

Score: 0.25045383

Video: DG90347831



Más información

Recurso: # 4

Score: 0.2488241

Video: DG90033229



Más información

Recurso: # 5

Score: 0.24868599

Video: I00786096



Más información

Recurso: # 6

Score: 0.24504258

Video: DG90021789

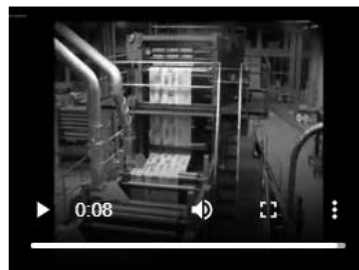


Más información

Recurso: # 7

Score: 0.24170998

Video: I900528772



Más información

Recurso: # 8

Score: 0.24167752

Video: DG90033601



Más información

Recurso: # 9

Score: 0.24079113

Video: DG90033692



Más información

Recurso: # 10

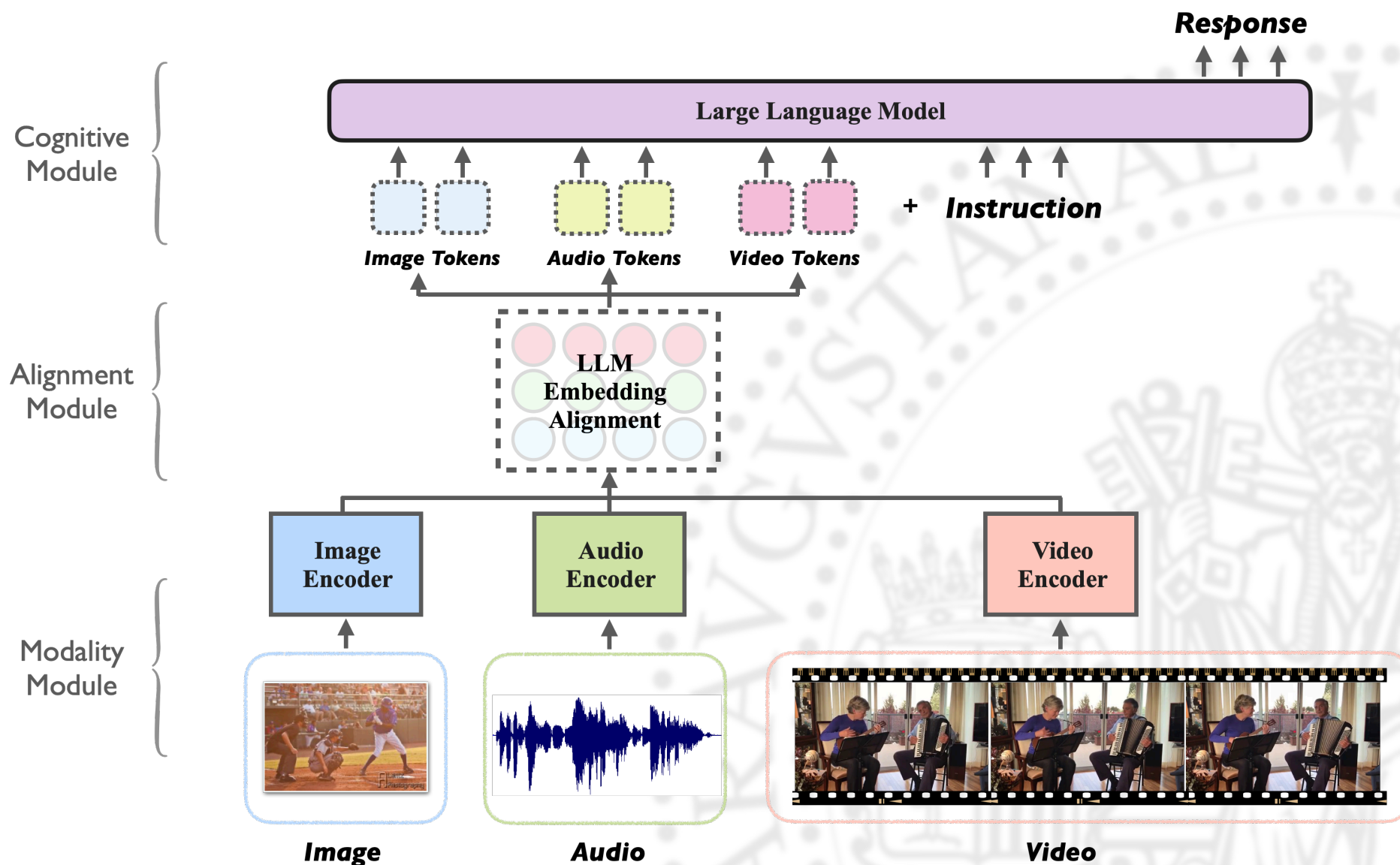
Score: 0.23922843

Video: DG90347831



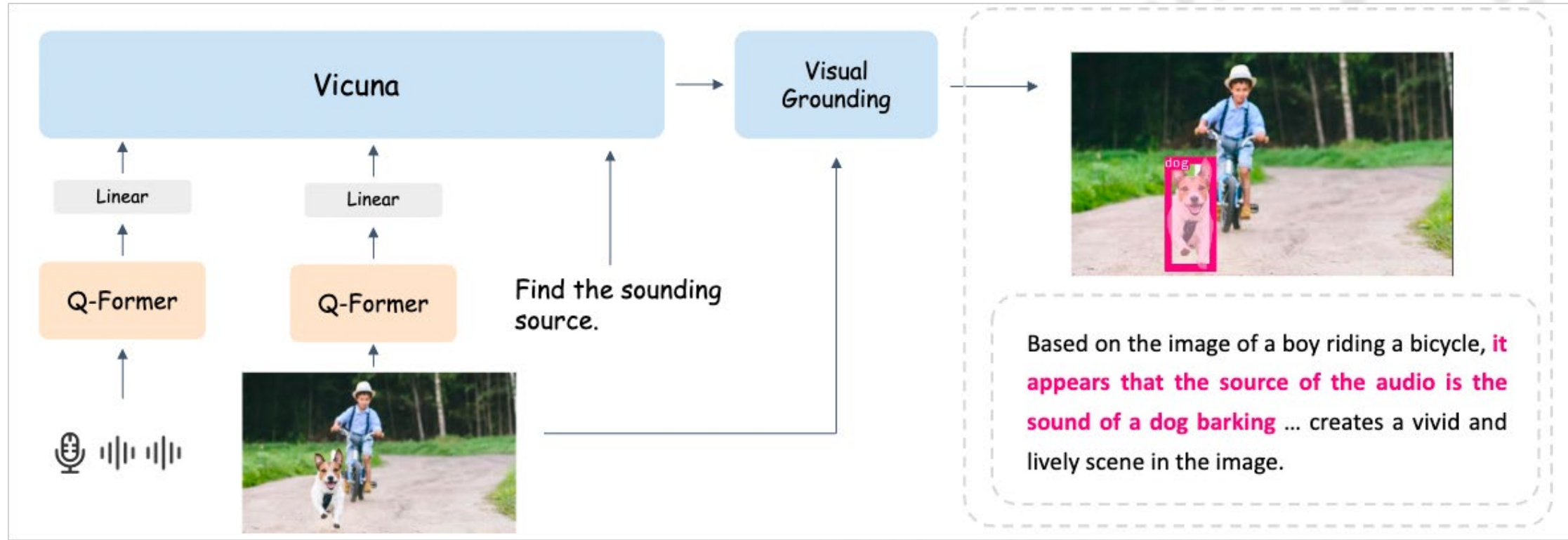
Más información

# MODELOS DE LENGUAJE MULTIMODALES



<https://github.com/lyuchenyang/Macaw-LLM>

# MODELOS DE LENGUAJE MULTIMODALES



<https://bubo-gpt.github.io/>

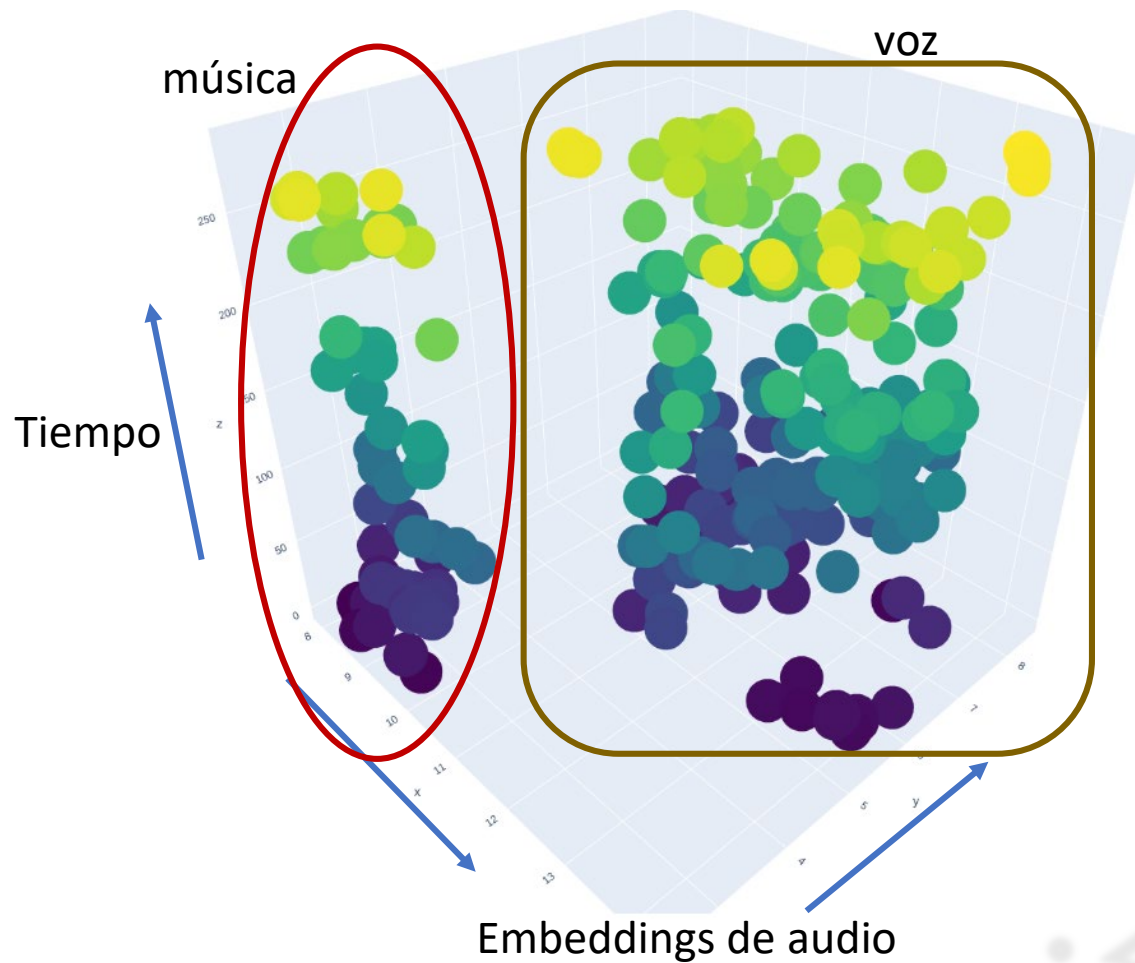
# MODELOS DE LENGUAJE MULTIMODALES



**Video-ChatGPT**

# MODELOS DE LENGUAJE MULTIMODALES

AudioCLIP: espacio semántico Audio-Texto-Vídeo



- ✓ Programas de televisión: Subtítulos / Audio / Vídeo
- ✓ AudioCLIP permite consultas-respuestas multimodales