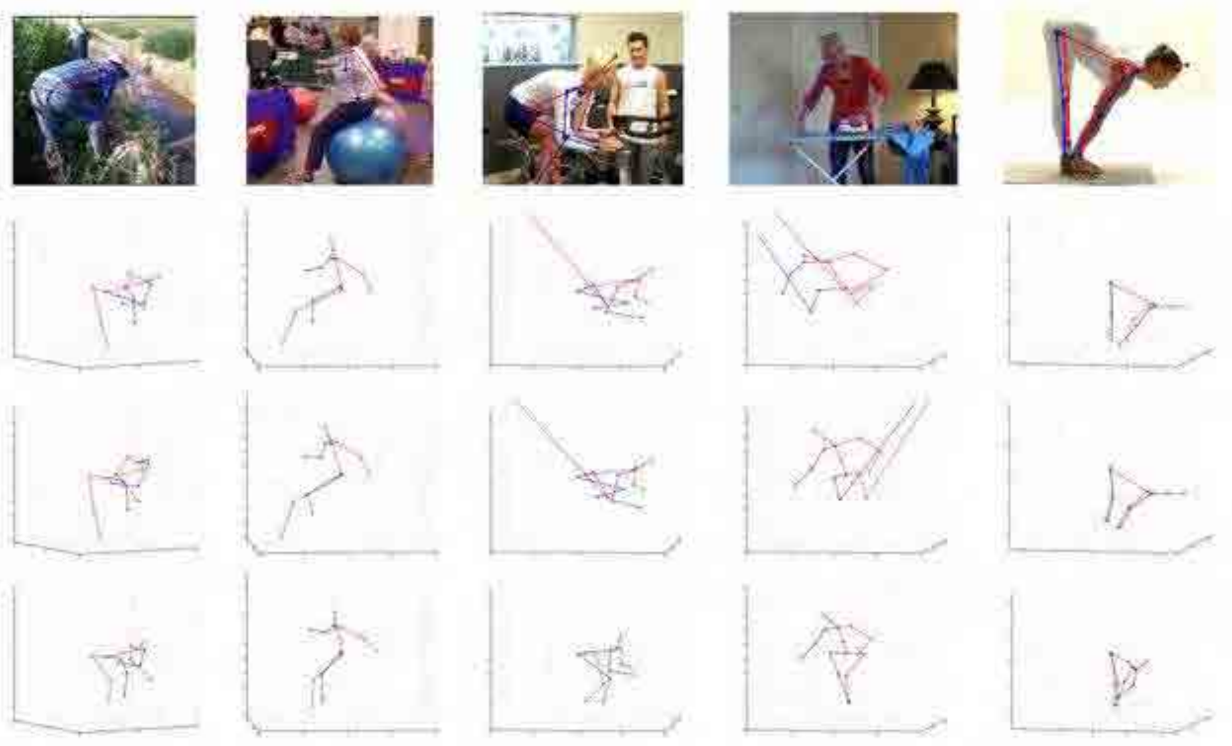
Pose and spatial movement as input for dynamic content search & generation

iptc

POLITÉCNICA

Pose and spatial movement as input for dynamic content search & generation

Main purpose? To explore the potential of posture correctness analysis and multimodal feedback delivery for different applications (ergonomics, yoga, others).



Tasks

1) Task scope definition ✓
2) Dataset search and evaluation. ✓
3) State of the art on building postural models and postural analysis. ✓
4) Setting up an environment for posture classification from images. ✓
5) Model concept proposal for posture analysis, based on angles. Built from reference datasets and literature. Limited scope. ✓
6) First beta prototype set up. ✓

Pose and spatial movement as input for dynamic content search & generation

CLIP as classifier

DEEPER ANALYSIS OF CLIP AS CLASSIFIER:

Yoga-82 dataset

- **Low zero-shot performance**
- **Significant improvement after fine tuning**
  - Metrics on **82 classes**:

	Precision	Recall	F1-score	MCC	Support
Weighted avg	0,861	0,859	0,857	0.855	3826

BUT STILL MARGIN TO IMPROVE !

iptc

Pose and spatial movement as input for dynamic content search & generation

CLIP as classifier

- **How can we improve the performance ?**
  - Balancing the classes in the dataset (μ= 186, σ=105)
  - Boosting classes that CLIP encounters issues with. (Already identified thanks to the representation based on hierarchical order groups. Confusion matrices)
    - E.g.:

Makara Adho Mukha Svanasana

F1 score: 0,5 | **36 train imgs**.

Chaturanga Dandasana

F1 score: 0,796 | **165 train imgs**.

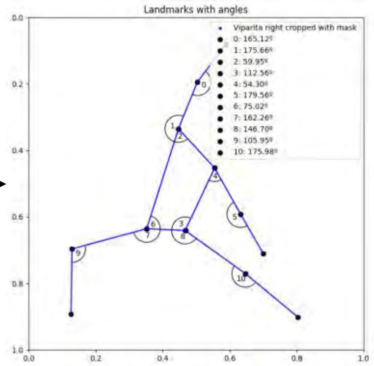
  - Combining with other types of images (infrared, joints, etc.)
  - Trying to fine-tune various models of visual encoders

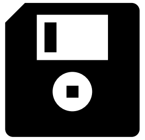
Pose and spatial movement as input for dynamic content search & generation

Mediapipe for pose evaluation

Angles extraction pipeline **applied to all images of each class.**

Dataset → Cropped dataset (image segmentation) → Joint extraction + angle computation (12 angles per posture) → Data manipulation JSON + pandas





Building the posture evaluator model.

Still necessary to evaluate it on real data.

- **Various attempts:**
  - Rules engines
  - KNN
  - XGBoost

Trained and tested using synthetic data generated from the angle's extraction pipeline.

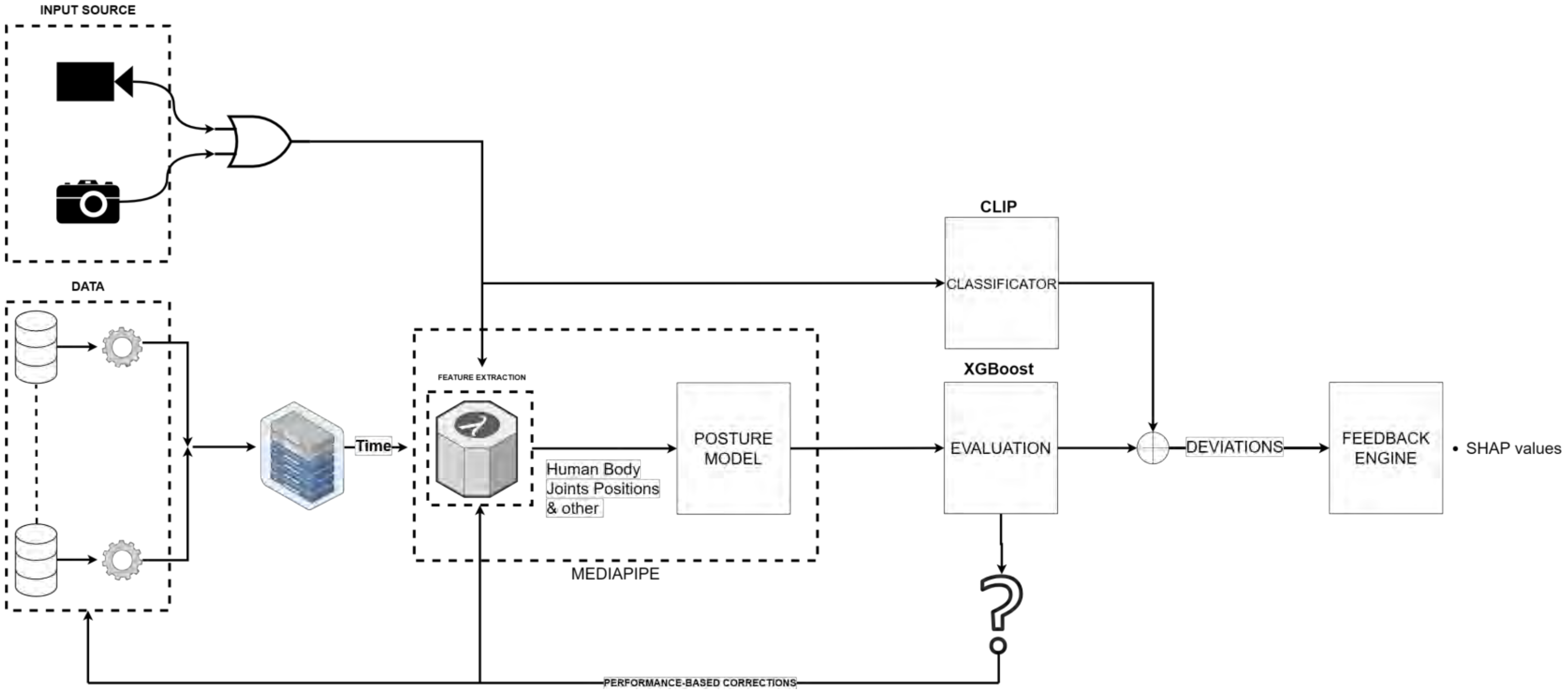
- **Best results: XGBoost. 1 model by posture.**

	Precision	Recall	F1-score	Support
Average	0,981	0,982	0,982	1000

Pose and spatial movement as input for dynamic content search & generation

Results

- **Results:**
  - Journal article:
    - **Exploring the Use of Contrastive Language-Image Pre-Training for Human Posture Classification: Insights from Yoga Pose Analysis**
  - Yoga-82 app



Pose and spatial movement as input for dynamic content search & generation

Results

Real time

Virabhadrasana II posture

Pose and spatial movement as input for dynamic content search & generation

Results

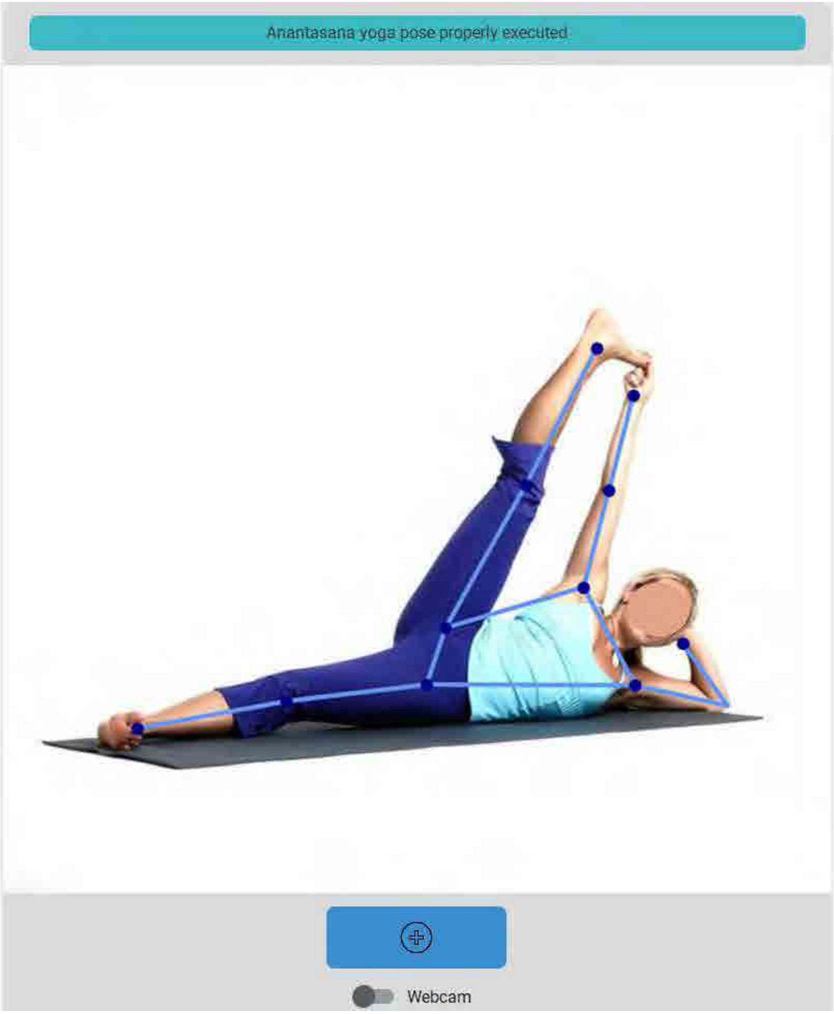
Real time

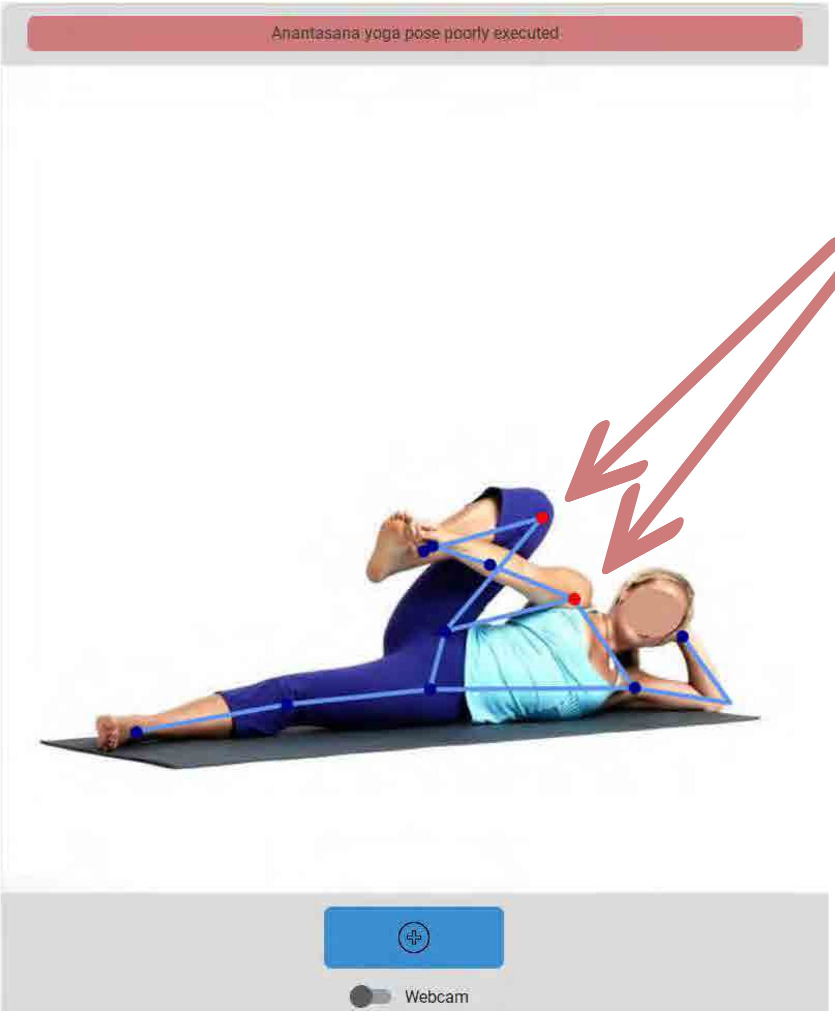
Marjaryasana posture

Pose and spatial movement as input for dynamic content search & generation

Results

Image based feedback





SHAP VALUES

Zero-shot sonorizing of video sequences

iptc

POLITÉCNICA

Zero-shot sonorizing of video sequences

Entangling AI-audio synthesis models and multimodal representations

Signal Processing Applications Group
GAPS – IPTC - UPM

How should this scene sound?

Should this sound similar?

The Question

And a shared goal

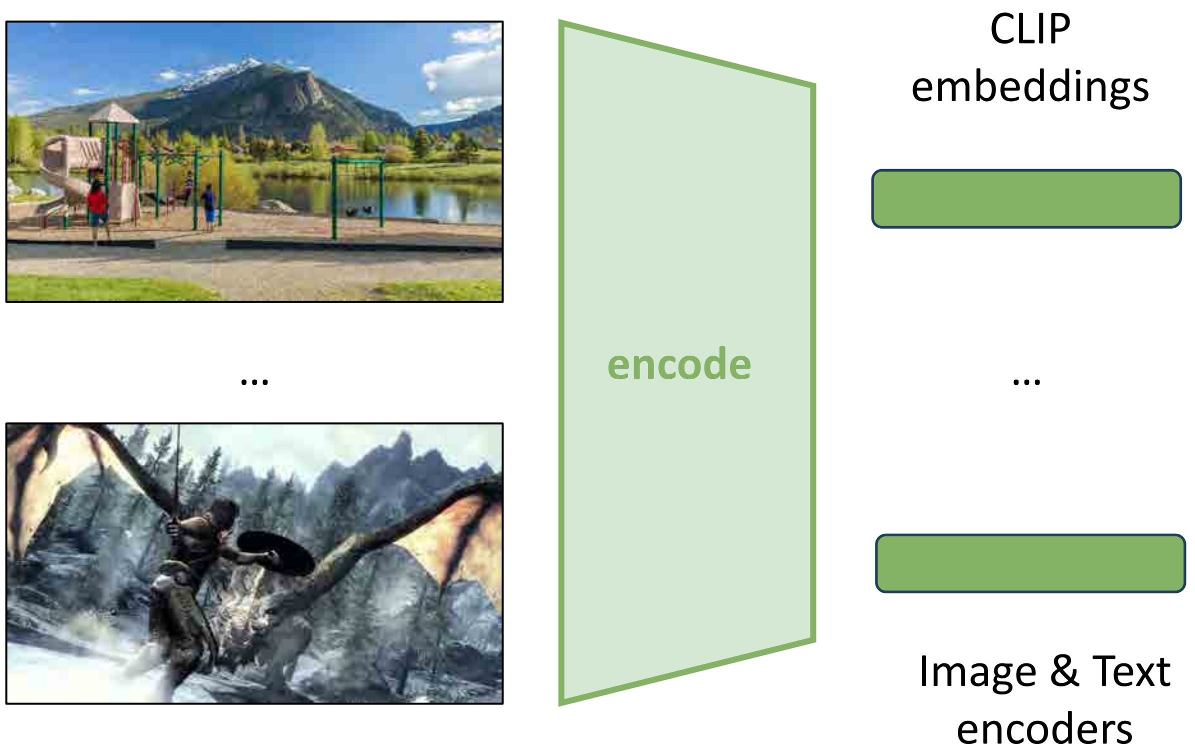
What is **a suitable audio** for a given image or video sequence?

How do we **search** or **create** a matching audio?

How should we **evaluate** if this match is *coherent*?

The approaches

Two ways to address the task + a novel way to evaluate



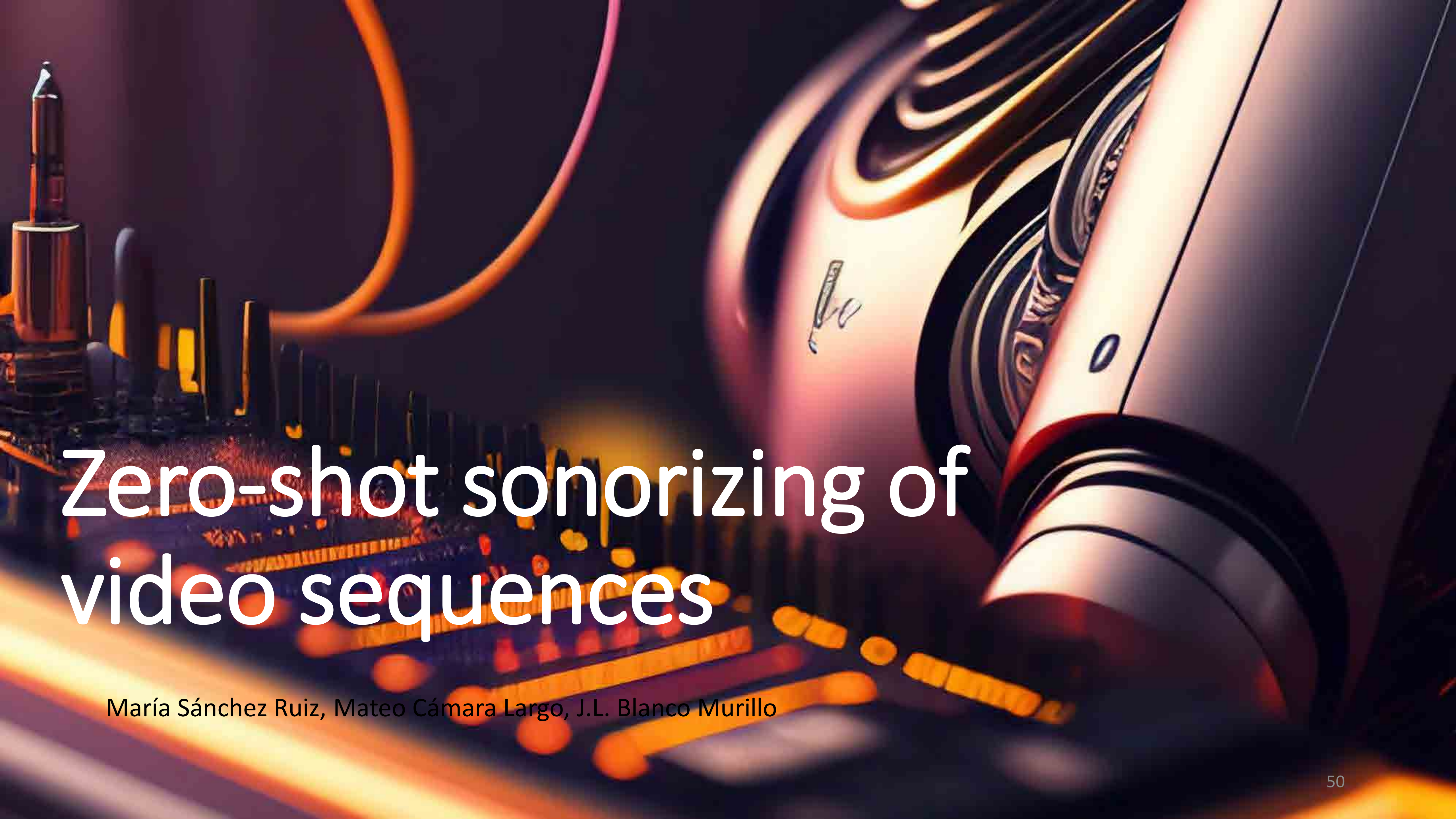
Multimodal representations

on Pretrained Models + Gen AI

Image & Text

Text & Audio

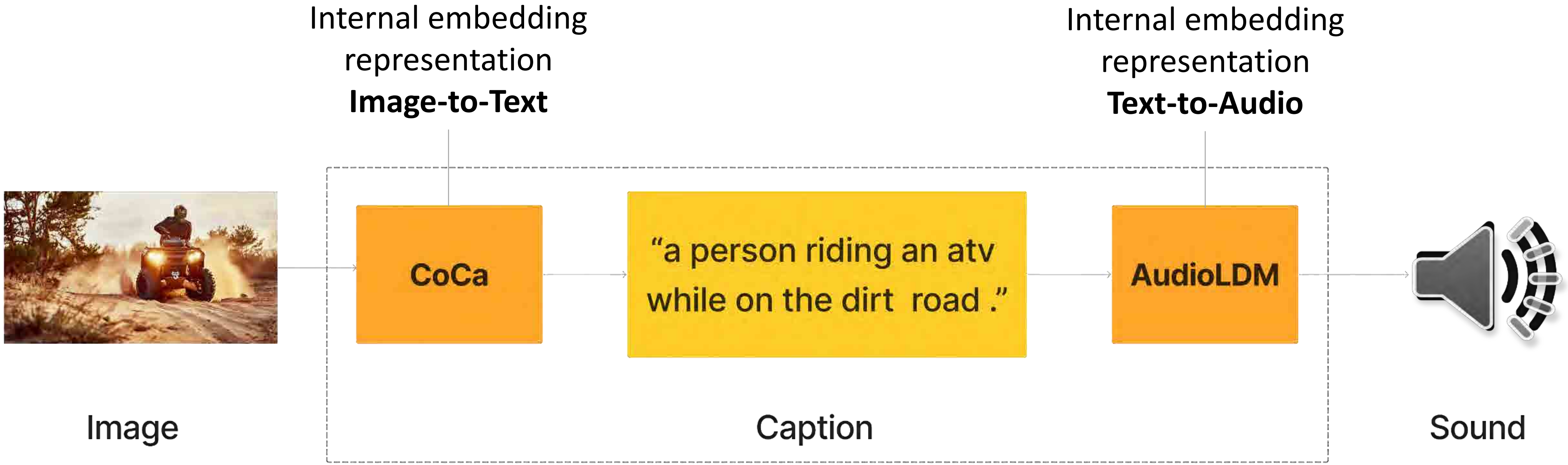
Image & Audio



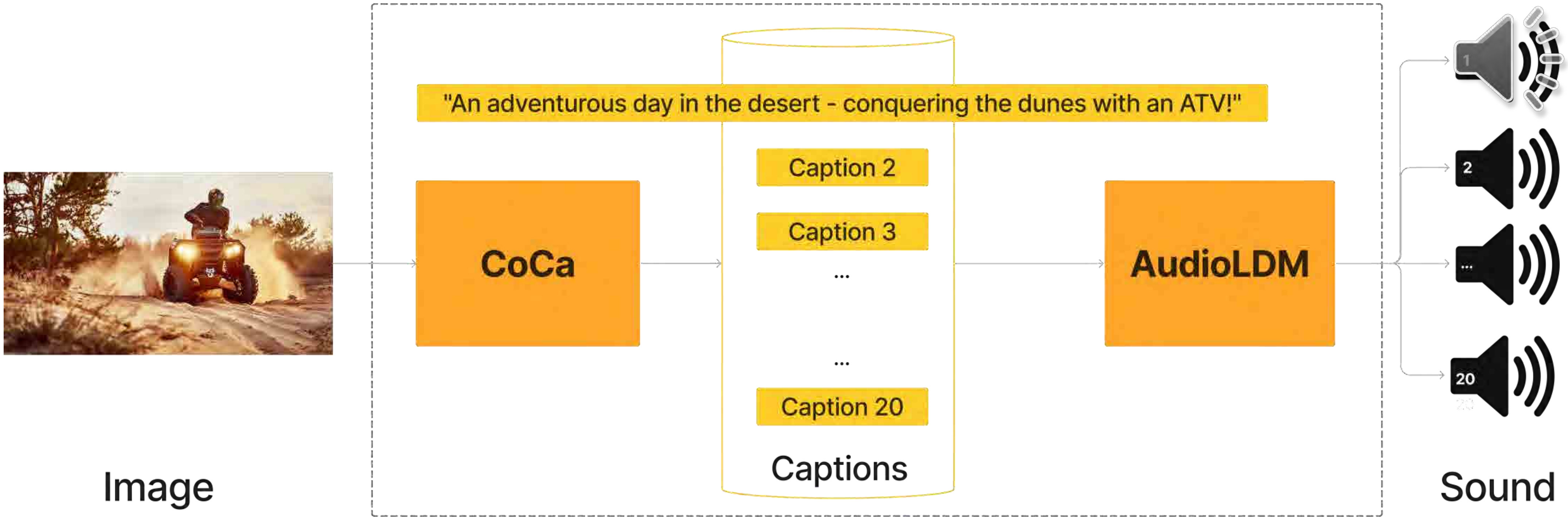
Zero-shot sonorizing of video sequences

María Sánchez Ruiz, Mateo Cámara Largo, J.L. Blanco Murillo

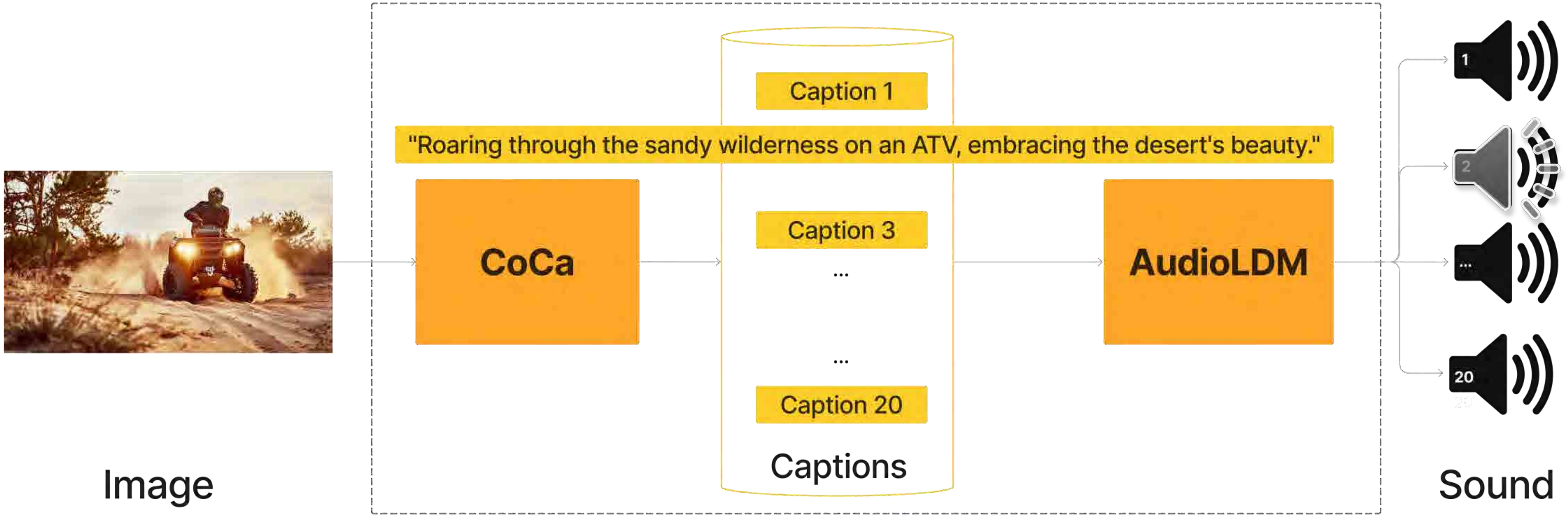
Text-guided sonirization



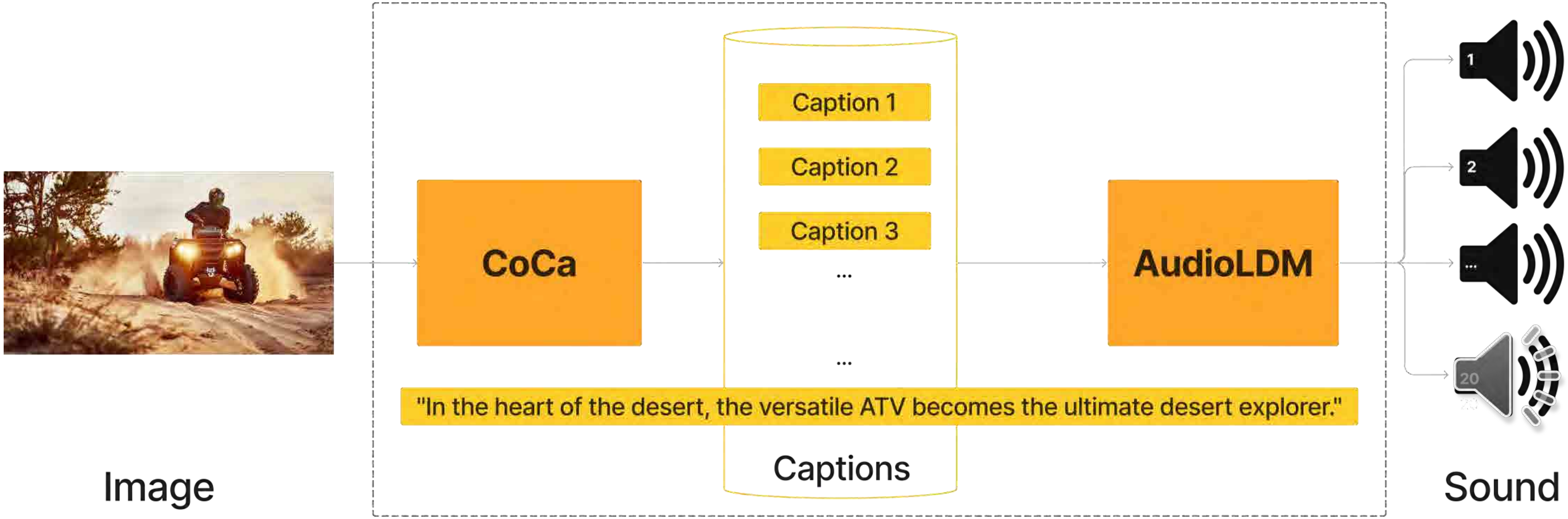
Multi-captioning



Multi-captioning



Multi-captioning

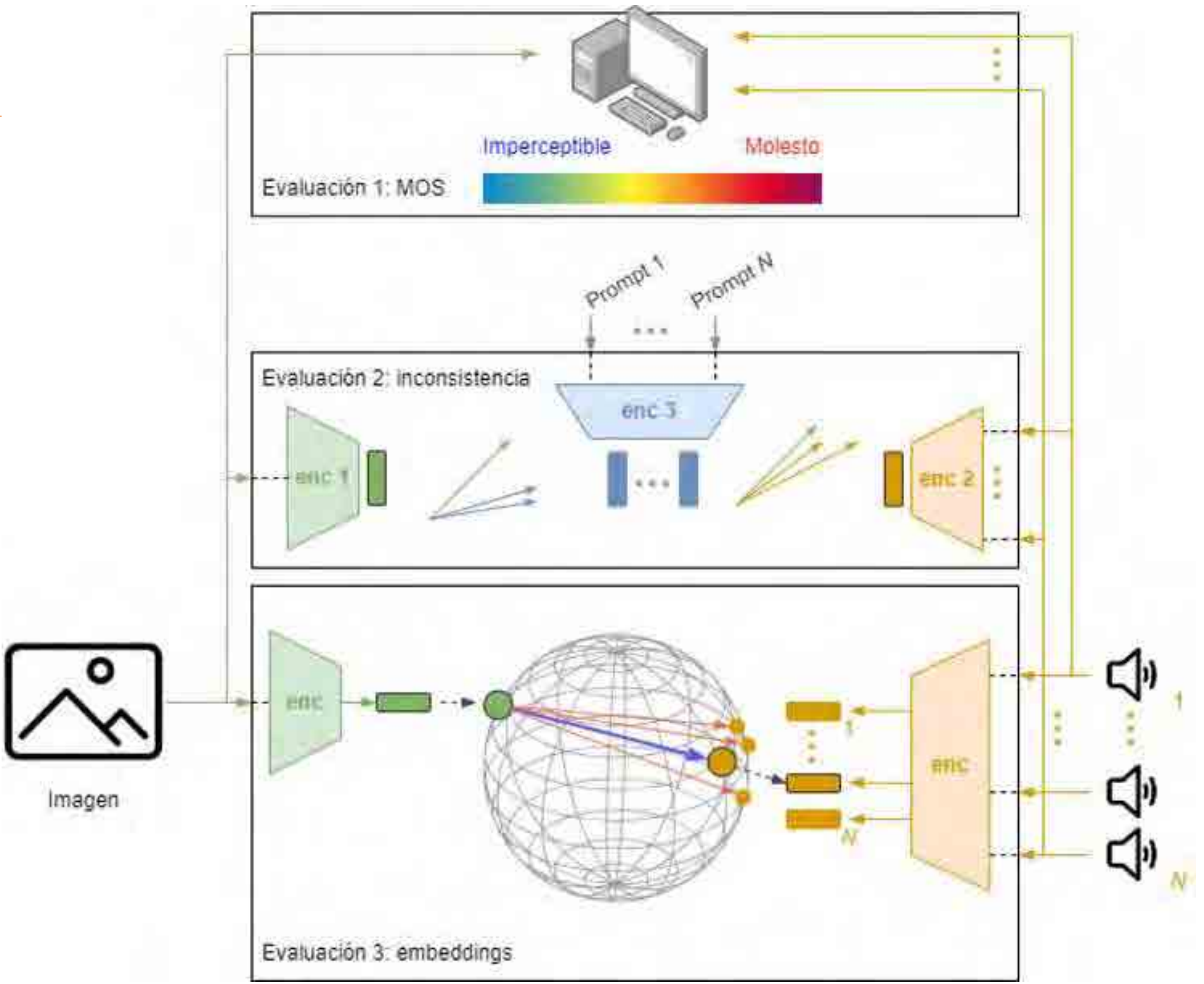


Evaluation

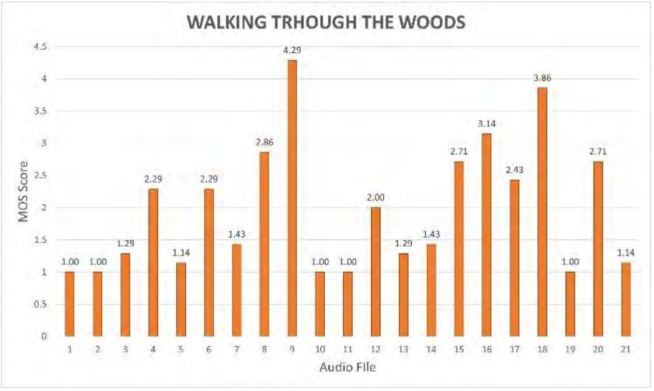
1. **Subjective** user experience.

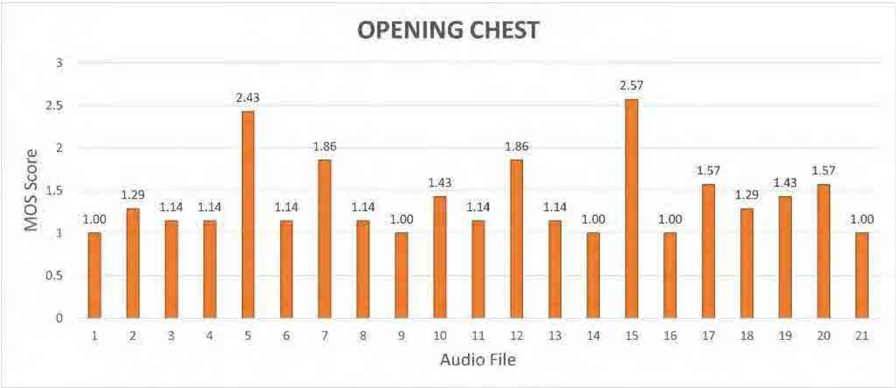
2. Embeddings **consistency**.

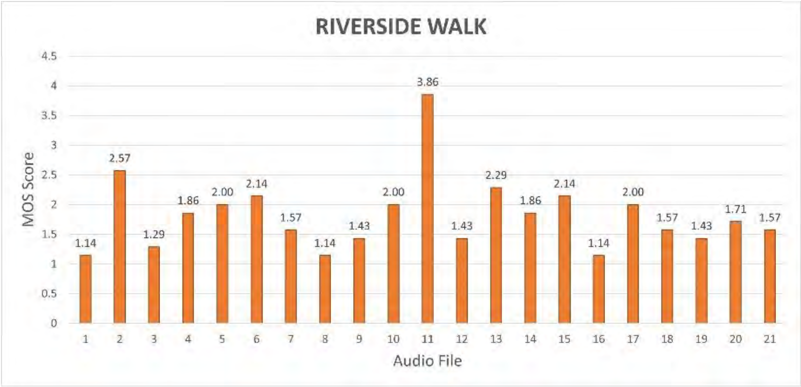
3. Embeddings **projection**.



Evaluation 1: MOS







Evaluation 2: Inconsistency

Image-Caption Cosine Distance

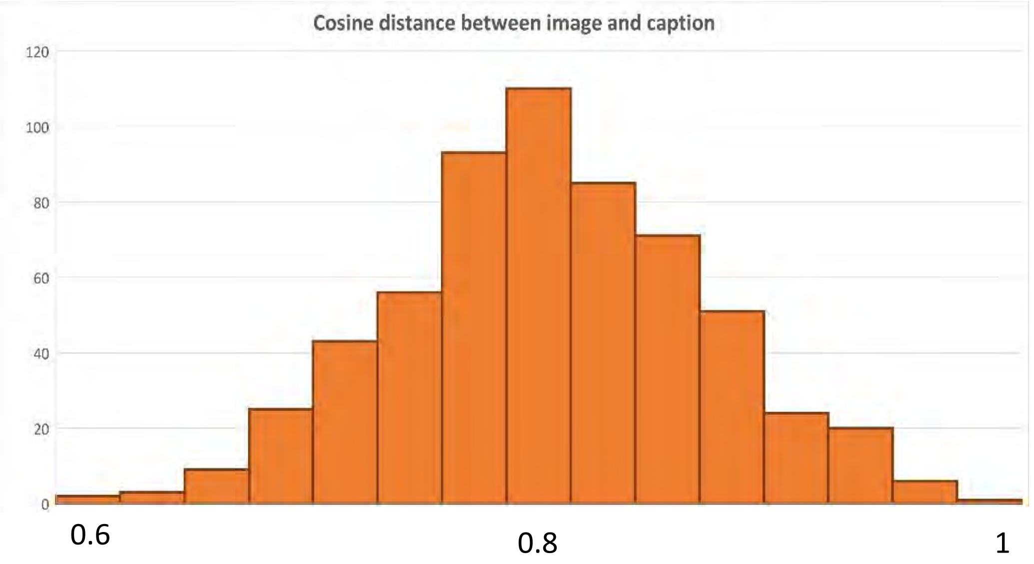
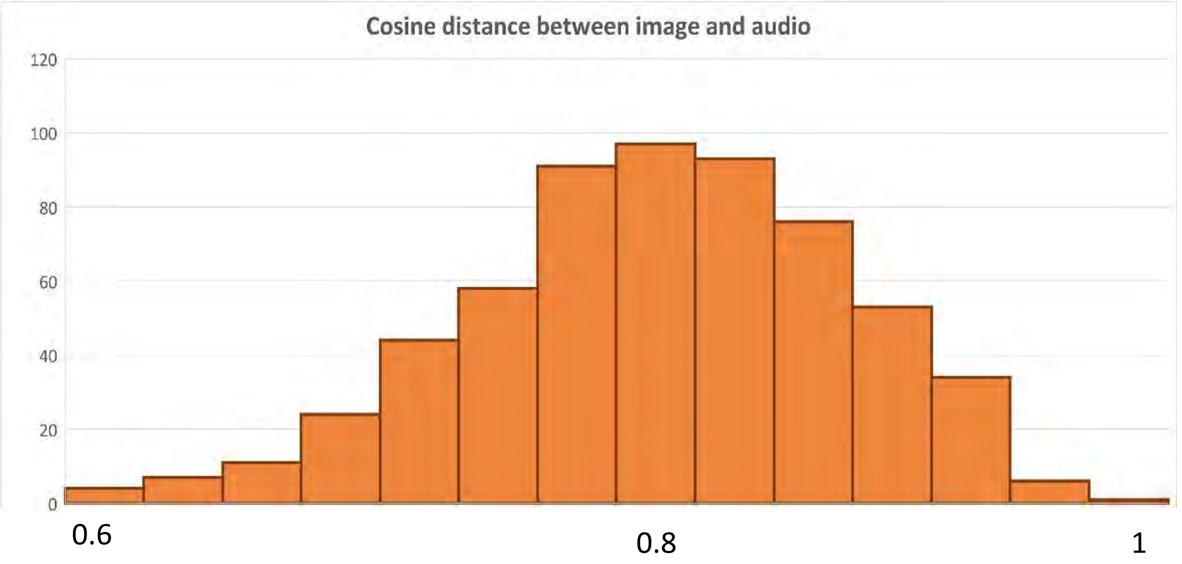
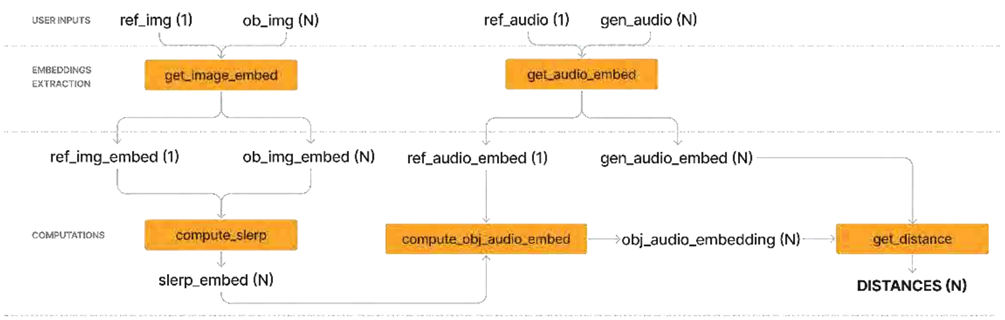


Image-Audio Cosine Distance



Evaluation 3: Embedding distance

Embedding Distance Scheme



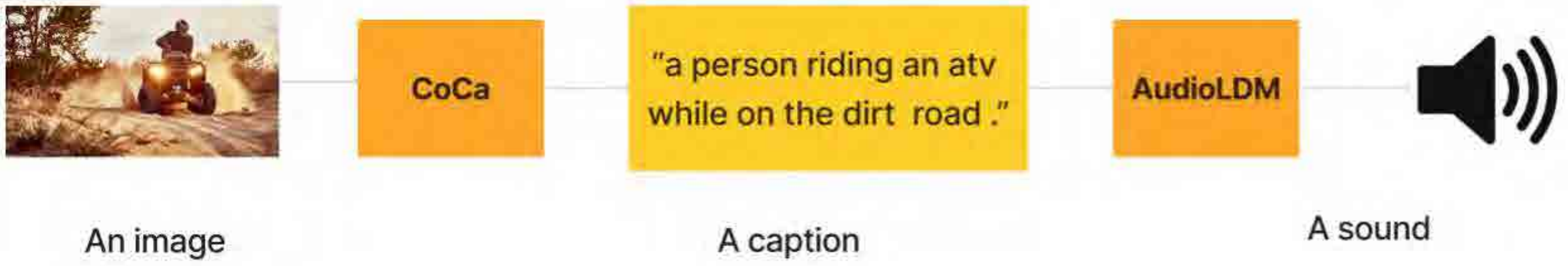
Summary of Validation Test Results

Test	Audios	Frames	Mean	STD
1	ref_audio, gen_audio	ref_img, ob_img	0	0
2	ref_audio, gen_audio	ref_img, ob_img	0	0
3	ref_audio, gen_audio	ref_img, ob_img	0	0
4	ref_audio, gen_audio	ref_img, ob_img	24.9	2.5
5	ref_audio, gen_audio	ref_img, ob_img	24.7	2.5
6	ref_audio, gen_audio	ref_img, ob_img	12.5	3.2
7	ref_audio, gen_audio	ref_img, ob_img	12.5	3.2

1. **Valid sonorization approach & evaluation procedure**
2. **Consistency** of the metrics with user subjective assessment.
3. Embeddings **consistency** metric is robust.

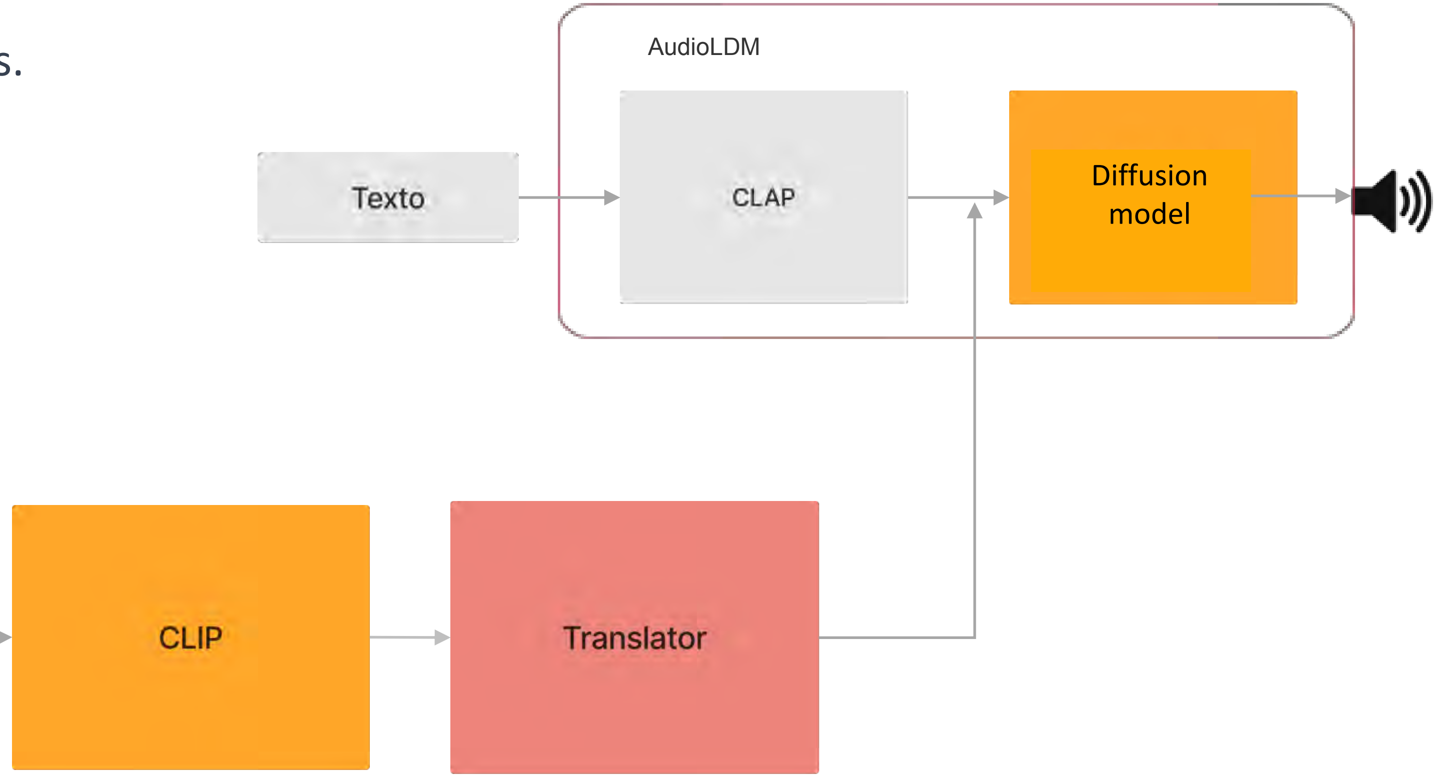
Entangling AI-audio synthesis models and multimodal representations

CoCa-AudioLDM integrated model.



CLIP-T-AudioLDM

- Eliminate the need for CoCa.
- Utilizes CLAP for audio representations.
- Translate CLIP embeddings into CLAP.

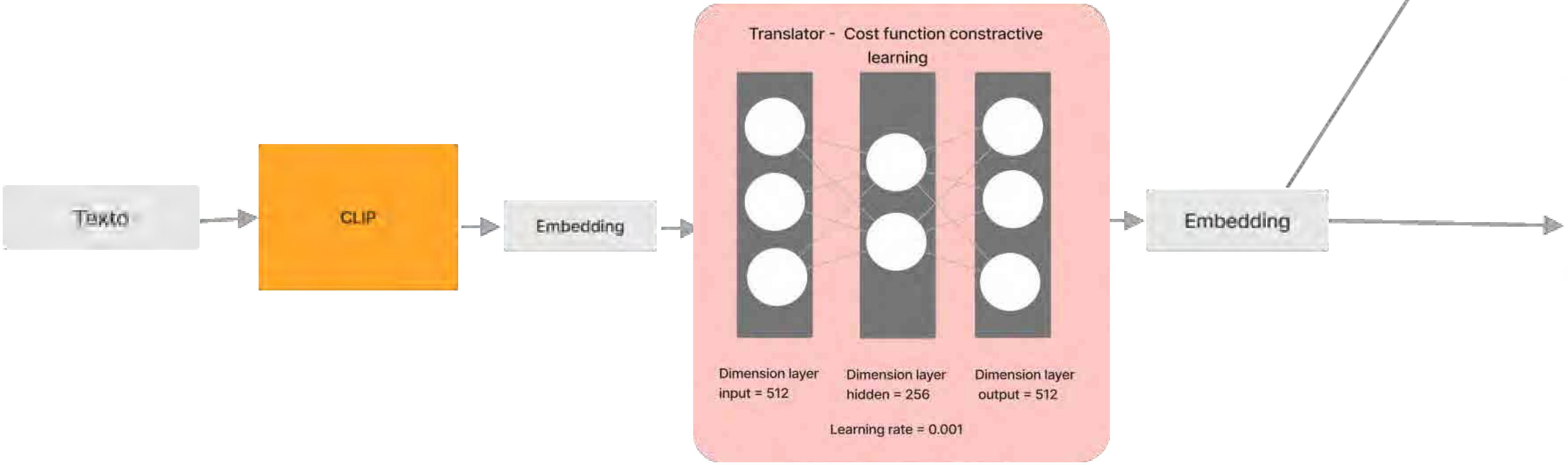


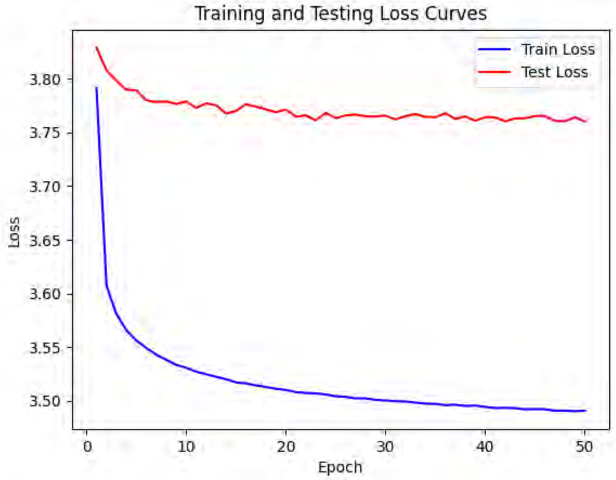
CLIP-T-AudioLDM

18 models implemented:

- 2 different hidden layer dimensions: 256 and 512.
- 3 different cost functions: MSE, CD and CL.
- 3 training databases: Audiocaps (57K), WIT(10K), Conceptual Captions(10K).

Model trained with all databases.





Evaluation of the two models

CoCa-AudioLDM

Metric	Gigs	Horses	Kids	Piano	Train	Average	Std desviation
FAD	17	12,43	10,88	4,78	15,56	12,03	4,65
DistanceMetric-Pipeline	23,24	24,85	24,18	24,82	28,5	25,12	2,00

FAD and Distance Metric-Pipeline metrics for the CoCa-AudioLDM integrated model.

CLIP-T-AudioLDM

Model	DROP	Learning rate	Gigs	Horses	Kids	Piano	Train	Average	Std desvia tion
CL512	0	0,001	14,12	23,85	**16,25**	**10,5**	25,8	**18,11**	**6,51**
	0,5		12,59	22,69	17,67	28,48	**24,9**	21,27	6,23
		0,002	**11,43**	**22,34**	17,51	30,18	27,13	21,72	7,5

FAD metrics for the CLIP-T-AudioLDM integrated model.

Model	DROP	Learning Rate	Gigs	Horses	Kids	Piano	Train	Average	Std desv iation
CL512	0	0,001	14,12	23,85	**16,25**	**10,5**	25,8	**18,11**	**6,51**
	0,5		12,59	22,69	17,67	28,48	**24,9**	21,27	6,23
		0,002	**11,43**	**22,34**	17,51	30,18	27,13	21,72	7,5

DistanceMetric-Pipeline metrics for the CLIP-T-AudioLDM integrated model.

Demo

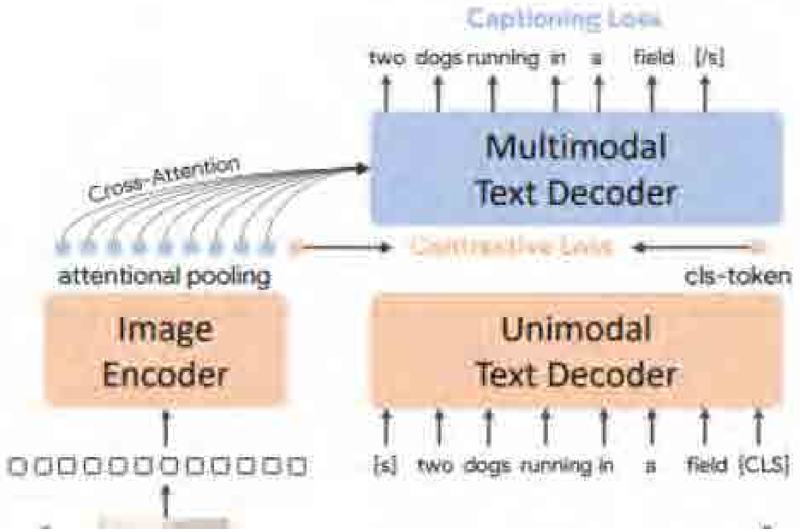
Home CoCa-AudioLDM CLIP-T-AudioLDM

Generator of Audio from Images

Welcome to our platform where two powerful models collaborate to generate audio from images. Let's explore the capabilities of these models:

Contractive Captioner (CoCa)

CoCa is designed to describe the content of an image. It employs an image encoder and a text decoder to obtain unimodal text representations. These representations are then used to create multimodal image and text representations. CoCa captures both global and regional characteristics of images and texts, making it versatile in various tasks such as visual recognition, image caption generation, and more.



React App
Imágenes de Google
localhost:3000

Generator of audio

Submit

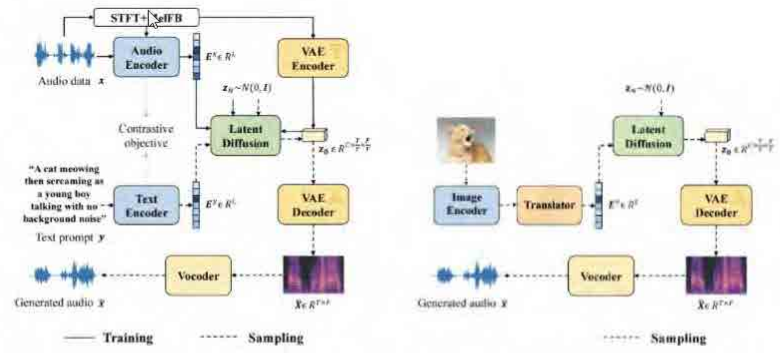
Results: submitted & under preparation

International Conference paper:

Del Visual Al Auditivo: Sonorización De Escenas Guiada Por Imagen

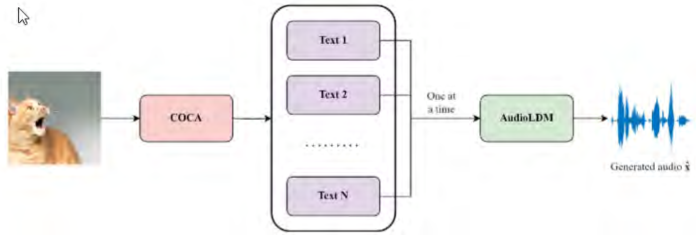
arxiv working draft:

Image-conditioned audio generation and evaluation using deep learning models



(a) Training and sampling process of AudioLDM. Image taken from [1].

(b) Modified AudioLDM sampling process. Image adapted from [1].



Results: and more than just that…

María Sánchez Ruiz
Masters Degree: **DTU+MUIT ETSIT**

12/12

Laura Fernández Galindo
Masters Degree: **MUIRST-ETSIT**

10/10

Thanks

Amazon Team. Particularly to our coworkers:

Guilia Comini **Adam Gabrys**