Informe final - AI Clip Generator

Carles Fornés Mas

Resum— El present projecte desenvolupa una eina basada en intel·ligència artificial per automatitzar la generació de clips curts a partir de vídeos llargs, com ara podcasts o conferències. El sistema és capaç de descarregar el vídeo original, transcriure'n l'àudio, identificar-ne les parts més rellevants mitjançant LLMs, aplicar reenquadrament adaptatiu amb detecció de cares i afegir subtítols personalitzables. El resultat són clips optimitzats per a xarxes socials en format vertical, amb alta qualitat visual i amb criteris definits per l'usuari. L'arquitectura modular i l'ús de paral·lelisme i GPU al núvol permeten una execució eficient i escalable. Els experiments realitzats demostren una millora significativa en el temps de processament i una alta precisió en tasques com la detecció d'idioma i la qualitat del subtitulat.

Paraules clau— Clips automàtics, edició de vídeo, IA generativa, subtítols dinàmics, xarxes socials.

Abstract— This project presents an AI-based tool designed to automatically generate short-form clips from long-form video content, such as podcasts or conferences. The system handles video download, audio transcription, relevant segment detection using LLMs, face-based adaptive reframing, and custom subtitle insertion. The output consists of high-quality, social-media-optimized clips in vertical format, tailored to the user's content preferences. Built with a modular architecture and supported by parallel processing and GPU acceleration in the cloud, the system achieves scalable and efficient performance. Experimental results show significant improvements in processing time and high accuracy in language detection and subtitle quality.

Index Terms— Automated clips, dynamic subtitles, generative AI, social media, video editing.

1 Introducció

En l'actual panorama digital, el consum de continguts audiovisuals breus ha esdevingut una tendència dominant[1], impulsada per l'èxit de plataformes com TikTok, Instagram Reels i YouTube Shorts. Aquestes xarxes socials han transformat la manera com es crea i es distribueix el contingut, fomentant un format ràpid, directe i visualment atractiu, orientat a captar l'atenció de l'usuari en qüestió de segons. Per exemple, un estudi sobre "short-format smartphone video consumption"[2] revela com els usuaris de la Generació Z són especialment receptius a aquest contingut, amb un patró de consum continu (binge-watching) que reforça el comportament repetitiu d'aquests formats.
Pel que fa al costat més comercial, el mercat dels vídeos curts està experimentant un creixement accelerat: el valor global d'aquest sector va ser de 1.520 milions de dòlars el 2022 i es preveu que arribi a 3.240 milions el 2030, amb un creixement anual compost del 10,2%[3].

La popularitat del contingut audiovisual curt a xarxes socials ha propiciat l'aparició de diverses eines comercials que automatitzen, totalment o parcialment, la generació de clips, com ara Opus Clip[4], Wisecut[5], Descript[6], Pictory[7] o Veed.io[8]. Aquestes plataformes utilitzen intel·ligència artificial per tallar, reenquadrar, subtitular i adaptar vídeos llargs a formats breus per a xarxes socials.

En aquest context, el present projecte proposa una solució pròpia basada en IA per automatitzar tot el procés, des de la descàrrega i transcripció dels vídeos fins a la selecció de fragments i l'edició dels clips. amb l'objectiu d'optimitzar la difusió de contingut digital.

El sistema desenvolupat en aquest treball s'ha construït amb una arquitectura modular[9], pensada per ser escalable i adaptable[10]. L'ús de serveis al núvol, la integració amb interfícies web i l'orientació cap a una experiència d'usuari clara i funcional permeten oferir una eina pràctica i completa per a creadors de contingut, equips de comunicació o qualsevol persona interessada en reutilitzar vídeos llargs en format curt.

El sistema es pot provar a través de la següent URL: http://97.107.131.94/ (**usuari**: el seu mail personal, **Contrsasenya:** TestProject2025)

L'Apèndix 1 mostra un conjunt de captures de pantalla i descriu pas a pas com fer-lo servir, de manera que qualsevol persona pugui provar l'eina fàcilment.

Finalment, aquest document recull tot el procés de desenvolupament del sistema, des de l'anàlisi del problema fins a la descripció tècnica dels mòduls i els experiments de validació.

2 Objectius

L'objectiu principal del projecte és desenvolupar un sistema automatitzat que permeti transformar vídeos llargs en clips curts i informatius mitjançant l'ús de models d'intel·ligència artificial i altres eines, amb l'objectiu de facilitar la difusió eficient de continguts audiovisuals a les plataformes digitals i xarxes socials.[11]

Per tal de dur a terme aquesta eina, s'han complert els següents objectius específics:

- Implementar un sistema de descàrrega automàtica de vídeos des de YouTube.
- Desenvolupar un sistema de transcripció automàtica de vídeos utilitzant models de reconeixement de veu.
- Detectar les parts més rellevants del vídeo mitjançant models d'intel·ligència artificial generativa.
- Generar clips a partir dels fragments seleccionats, aplicant tècniques de processament audiovisual com la detecció de rostres, el retall automàtic i el reenquadrament d'imatge.
- Generar i sincronitzar subtítols automàticament a partir de la transcripció del vídeo incloent opcions de personalització visual com tipografia, colors, posició i animació, per tal d'adaptar-los a l'estil dels clips generats.
- Dissenyar i implementar una infraestructura backend escalable per gestionar i accelerar el processament dels vídeos.
- Assegurar l'accessibilitat i la usabilitat del sistema, integrant-lo amb una interfície web funcional.
- Optimitzar el rendiment del sistema mitjançant la paral·lelització de processos i l'ús eficient dels recursos computacionals, aplicant bones pràctiques de desenvolupament.

3 Metodologia

El pas principal és el desenvolupament de l'eina. S'ha dividit el sistema en parts independents, que s'han desenvolupat i millorat per separat. Gràcies a aquest enfocament, ha estat possible anar provant cada mòdul de manera progressiva i aplicar-hi millores sense afectar la resta del sistema.

En la *Figura 3.1* es representa de forma el flux de treball del programa, per un millor enteniment de com funcionen totes les parts juntes.

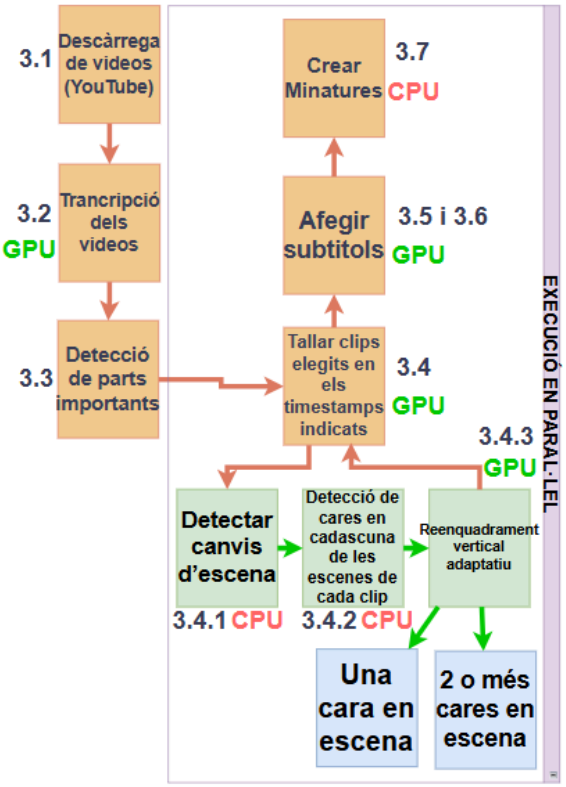


Figura 3.1 - Flux de treball.

A continuació, es descriuen els passos principals de la metodologia.

3.1 Descàrrega de videos

La seva funció és descarregar vídeos d'internet fent servir l'URL del vídeo, en aquest cas, s'ha fet servir YouTube[12] com la font dels vídeos.

L'eina utilitzada per la descàrrega dels vídeos ha sigut yt-dlp[13] que es de codi obert i ens permet descarregar els vídeos a partir de l'URL.

Ja que YouTube fa servir l'estàndard DASH (Dynamic Adaptive Streaming over HTTP) el qual distribueix el vídeo i l'àudio per separat en moltes resolucions, yt-dlp descarrega ambdós components de manera independent per garantir la màxima qualitat. Posteriorment, es combinen en un únic fitxer sincronitzat (per exemple, output.mp4) usant FFmpeg, que també és una eina de codi obert.

3.2 Trancripció dels videos

La seva funció és transcriure l'àudio dels vídeos (input) a text incloent-hi els timestamps d'inici i final de cada paraula (output), independentment de l'idioma.

L'eina usada ha sigut Whisper d'OpenAI[14] que ofereix diversos models, des de "tiny" fins a "large", amb diferents nivells de precisió i requisits computacionals. Per elegir el model de whisper perfecte pel nostre cas hem realitzat un estudi dels diferents models a la máquina g4dn.2xlarge on es pot observar la relació entre el temps d'execució i la precisió (WER) per a un àudio de 30 minuts. A la *Figura 3.2.1* es mostren els resultats de l'estudi

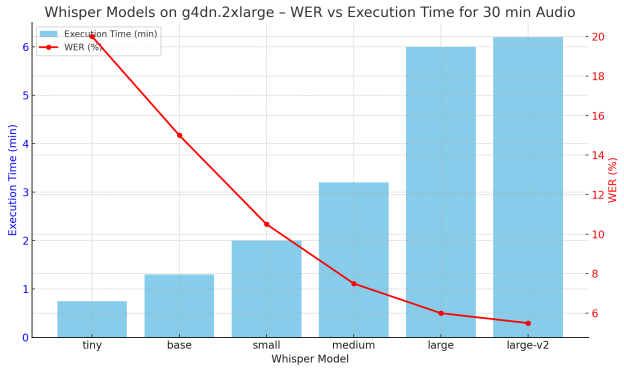


Figura 3.2.1 - Anàlisi per elegir model.

Com que el model medium de whisper ofereix la millor relació entre temps d'execució i WER ha sigut el model escollit.

Un cop obtinguda la transcripció del vídeo amb els timestamps de cada paraula, és necessari agrupar-les en frases completes per facilitar-ne l'anàlisi per part dels models de llenguatge natural (en el nostre cas fem servir l'API de ChatGPT[15]). Si es passessin paraula per paraula, amb tants timestamps, el model podria no comprendre correctament el context de cada idea.

Per aquest motiu, s'ha implementat un procés de segmentació mitjançant un script en Python que agrupa les paraules en frases, utilitzant la puntuació (per exemple, el punt ".") com a delimitador. Cada frase segmentada conserva el timestamp d'inici de la primera paraula i el timestamp final de l'última, cosa que permet identificar amb precisió el moment exacte en què es diu cada frase. En la *Figura 3.2.2* es mostra un exemple del funcionament.

Si Whisper retorna la transcripció següent:

[00:00.00 - 00:00.20] Lo
[00:00.20 - 00:00.40] que
[00:00.40 - 00:00.80] cambió
[00:00.80 - 00:01.00] todo.

Amb el procés de segmentació, s'obté:

[00:00.00 - 00:01.00] "Lo que cambió todo."

Figura 3.2.2 - Anàlisi per elegir model.

3.3 Detecció de parts importants del video mitjançant IA generativa (LLM)

La seva funció és seleccionar a partir del text transcrit complet del vídeo (input), només els fragments més rellevants, indicant-ne els timestamps d'inici i final (output). Aquests fragments definiran els clips finals, i és el LLM qui decideix quants i quins en genera segons el context.

Per detectar els fragments més rellevants, l'usuari selecciona el criteri de selecció a la interfície web (per exemple, "interessant", "viral", "motivacional", "controvertit", "relatable", "divertit" o personalitzat). Segons aquest criteri, el sistema adapta el prompt mitjançant les tecniques de prompt engineering que es troben en l'apartat 5 i el fa servir amb l'API de ChatGPT, que analitza les frases i selecciona automàticament aquelles que millor s'hi ajusten.

Això permet personalitzar el tipus de clips generats segons l'objectiu o el públic al qual es vol dirigir el contingut.

En essència, el LLM actua com un resumidor que navega per l'ambigüitat del que és un highlight[16], permetent flexibilitat segons el tipus de vídeo (no és el mateix un highlight en un partit de futbol que en una xerrada TED[17], i el LLM pot adaptar-s'hi mitjançant el prompt).

3.4 Tallar clips elegits en els timestamps indicats

Un cop el nostre sistema ha identificat les parts rellevants, s'extreuen físicament aquests intervals del vídeo original. Per a cada interval seleccionat, es fa un tall del vídeo amb FFmpeg[18] per generar un clip independent.

En aquesta etapa, cada clip base conserva el format original (per exemple, 1920x1080) i encara no té els subtítols incrustats ni cap transformació aplicada. Aquests clips es faran servir a les etapes posteriors per al reenquadrament, afegit de subtítols i adaptació de format.

3.4.1 Detectar canvis d'escena

Una de les millores clau del sistema és la capacitat de detectar automàticament els canvis d'escena dins de cada clip generat. Aquesta funcionalitat permet identificar transicions visuals significatives i ajustar els enquadraments de manera més precisa i adaptativa.

Els clips es poden classificar segons el seu nivell de dinamisme, és a dir, segons la quantitat i freqüència de canvis d'escena que contenen. Un alt dinamisme implica més transicions visuals i pot requerir més ajustos d'enquadrament per mantenir els rostres centrats. En canvi, clips amb poc dinamisme mantenen una composició més estable i no necessiten tants ajustos. Aquesta classificació és clau per determinar el bon funcionament del sistema posteriorment en l'experiment 3.

Per exemple, si es produeix un canvi d'escena i no s'hi fa cap adaptació, és probable que el rostre que estava ben centrat inicialment quedi desencaixat després del tall.

Per a la detecció de canvis d'escena, es desenvolupa un algorisme propi que funciona a partir de la comparació d'histogrames de color dels fotogrames. A l'*Algorisme 1* es pot veure el pseudocodi de la detecció de canvis d'escena.

Input: *Vídeo d'entrada (format MP4,).*
Output: *Llista de timestamps corresponents als canvis d'escena detectats.*

1. *S'obre el vídeo.*

2. *Es crea una llista per desar els moments en què canvien les*

escenes.

***3.** Es guarda el temps de l'últim canvi d'escena (que comença en 0).*

***4.** Per a cada frame (imatge) del video:*

- *a. Es calcula un histograma de colors (que representa com estan distribuïts els colors).*
- *b. Es compara aquest histograma amb el del frame anterior.*
- *c. Si la diferència entre els histogrames és superior a un cert llindar (threshold histograma) I si han passat com a mínim 0,5 segons des de l'últim canvi (threshold temps):*
  - *i. Es desa el moment i el frame com un canvi d'escena.*
  - *ii. S'actualitza el temps de l'últim canvi.*

***5.** Es retorna la llista amb els timestamps dels canvis d'escena.*

Algorisme 1 - Detecció de canvis d'escena, per la implementació s'ha usat OpenCV[19].

A la *Figura 3.4.1.1* es pot veure un exemple dels histogrames de color que ens podem trobar.

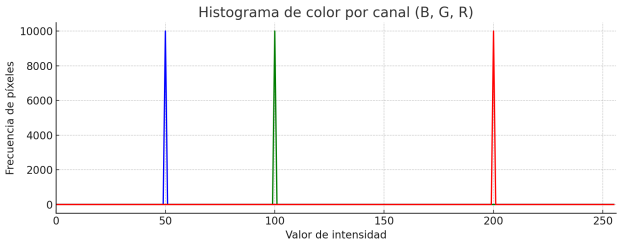


Figura 3.4.1.1 - mostra dels histogrames de color calculats.

Si els pics entre dos fotogrames canvien de manera brusca, és un indici que probablement hi ha hagut un canvi d'escena[20].

3.4.2 Detecció de cares en cadascuna de les escenes de cada clip

L'objectiu és detectar totes les cares que es troben en un mateix frame i retornar les coordenades.

Per dur a terme aquesta tasca s'ha usat una eina basada en el repositori "Group Emotion Recognition" de Samanyou Garg [21, 22]. Tot i que l'objectiu original del repositori és la detecció d'emocions en grups mitjançant visió per computador, en aquest projecte només s'utilitza el mòdul del repositori que permet detectar múltiples rostres en una sola imatge, i no la part de reconeixement d'emocions.

El funcionament d'aquest mòdul dins del sistema és el següent:

Per a cada canvi d'escena dins d'un clip, el sistema analitza el primer frame per detectar si hi ha alguna cara visible. Si en troba més d'una, dona prioritat a la que està més centrada. Aquesta informació s'utilitza després per assegurar que els rostres principals apareguin ben enquadrats en el vídeo final. Si no es detecta cap cara, es considera que no cal fer cap ajustament automàtic d'enquadrament. (l'algorisme amb més detall es pot consultar a l'apèndix *Algorisme 2*)

En funció d'aquesta detecció, el sistema pot decidir com mostrar l'escena:

- Amb un únic rostre enfocat
- Amb múltiples rostres en una pantalla dividida.

En la *Figura 3.4.2.1* es pot veure el resultat de detectar les cares en els frames indicats

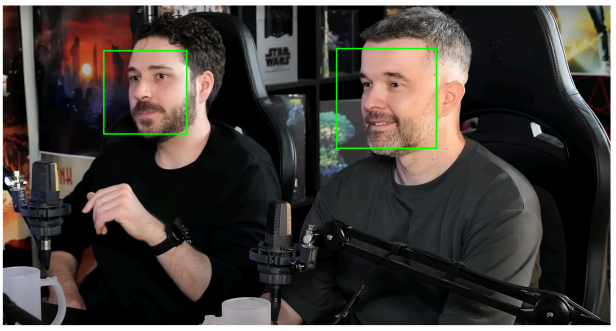


Figura 3.4.2.1 - Detecció de dos cares en la escena.

Després de processar tots els frames de referència, s'obté una llista de posicions dels rostres al llarg del clip, amb els seus temps associats. Aquesta informació permet determinar si, en una determinada secció del clip, hi havia una sola persona o diverses, fet que resulta clau per decidir si cal aplicar un enquadrament únic o bé una pantalla dividida, o combinar els dos al llarg del clip.

3.4.3 Reenquadrament vertical adaptatiu

L'objectiu és retallar el vídeo al format mòbil (9:16), mantenint sempre les cares dins de l'escena de manera automàtica.

Per realitzar aquestes transformacions utilitzo l'eina Ffmpeg. Podem separar el reenquadrament en dos casos:

- **Cas en què només hi ha un rostre en pantalla:** S'aplica un retall en format 9:16 centrat en el rostre. La posició horitzontal del retall es determina pel centre del rostre, i en vertical no es retalla per evitar tallar el cap o els peus.

- **Cas en què hi ha dos rostres en pantalla:** S'aplica una pantalla dividida vertical per mostrar dues persones simultàniament, una a sobre i l'altra a sota. Es generen dos retalls, cadascun centrat en un rostre. Cada retall ocupa la meitat de l'alçada del vídeo (p. ex. 540 px si l'original fa 1080 px). Finalment, s'apilen amb FFmpeg per formar un únic clip vertical.

En la *Figura 3.4.3.1* es mostren els resultats en cada cas.

a)

b)

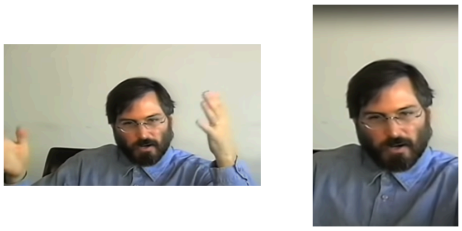


Figura 3.4.3.1 a) - Crop amb dues cares en escena.
b) - Crop amb una cara en escena.

Cal aclarir que cada escena (entre canvis d'escena detectats) es tracta de forma independent, se li aplica el seu propi reenquadrament (retall) i després es tornen a unir tots els fragments retallats per formar el clip complet.

És a dir, que si en una escena del clip només hi ha una persona, se li aplicarà el retall amb un sol rostre en pantalla, però si després, en un canvi d'escena, apareixen dues persones, a aquesta escena se li aplicarà el retall amb dos rostres. Posteriorment, aquests "mini fragments" del clip original s'ajuntaran de nou per formar el vídeo final.

3.5 Creació de subtítols

Els subtítols són un component essencial per a l'accessibilitat i l'atractiu dels clips a les xarxes socials (molts usuaris consumeixen vídeos en silenci).

Utilitzant la transcripció de cada clip amb els timestamps corresponents obtingut en el punt 3.3 el nostre sistema genera subtítols per a cada clip amb el format i estil adequats.

Per afegir els subtítols als vídeos, primer cal generar un fitxer amb tota la informació necessària en format Advanced SubStation Alpha (.ass). Aquest format permet definir no només el text i els timestamps, sinó també aspectes com la posició (coordenades x, y), color, font o mida dels subtítols. Un cop creat aquest fitxer .ass per a cada clip, s'utilitza FFmpeg per incrustar-lo al vídeo final, mantenint la sincronització i l'estil desitjats.

Els fitxers .ass segueixen la estructura[23] marcada en la *Figura A4* de l'apèndix

3.6 Algorisme intel·ligent que posiciona els subtítols dinàmicament, evitant rostres i adaptant mida i posició segons el contingut.

S'han incorporat diversos modes experimentals que afegeixen personalització als subtítols, com ara dinamització de les paraules, i colors dinàmics, amb l'objectiu de fer el text més atractiu i visualment destacat. L'usuari pot triar el tipus de subtítols que vol aplicar des de la web seleccionant entre diferents estils disponibles per adaptar-se a les preferències visuals del clip.

Durant la col·locació dels subtítols, es té en compte les posicions de les cares a l'hora d'escriure les coordenades de cada paraula en el fitxer .ass, ja que en cas d'haver-hi una superposició canviem la coordenada de la paraula perquè no interfereixi en el clip.

Un cop ja sabem com col·locar els subtítols en pantalla, hem de decidir quines paraules són les importants en el clip i que, per tant, aplicarem un estil diferent de la resta de paraules com també podem veure en la *Figura 1.2.1* de l'apèndix amb la paraula "SUCCESSFUL".

Per determinar aquestes paraules importants s'ha desenvolupat un algorisme basat en TF (Term Frequency)[24] que identifica les paraules més rellevants en un fragment d'àudio.
l'Algorisme de decidir i col·locar paraules importants funciona de la següent manera:

Primer es compta quantes vegades surt cada paraula en el vídeo original donant més pes a aquelles que més es repeteixen, posteriorment es descarten les paraules connectores ("el", "de", "i", etc.). El sistema també dona més pes a paraules que són noms propis o que són llargues, ja que normalment són més rellevants.

Les paraules que tinguin una millor valoració final seran les que es mostraran com a paraules importants.

Finalment, abans de mostrar aquestes paraules com a subtítols, es comprova que quedin ben situades a la pantalla i dins dels límits del format vertical, perquè sempre siguin llegibles i visuals.

Actualment, el color de les paraules destacades s'assigna de manera totalment aleatòria, però es preveu millorar aquest comportament en el futur mitjançant algun tipus de sistema de detecció més intel·ligent.

A l'apèndix es pot trobar *l'Algorisme 3* on es mostra l'algorisme de detecció de paraules importants

3.7 Sistema de miniatures (imatge de portada) dels vídeos.

L'objectiu aquí és obtenir miniatures automàticament per fer servir a l'hora de penjar els vídeos a les xarxes socials.

S'ha implementat una primera versió del sistema de generació automàtica de miniatures per als clips, amb l'objectiu de facilitar la publicació directa del contingut a les xarxes socials o plataformes de vídeo. Aquest sistema

permet a l'usuari personalitzar la imatge resultant, incloent-hi efectes visuals i marca d'aigua de forma automatitzada.
A l'apèndix *Algorisme 4* es pot trobar l'algorisme de creació d'una miniatura.

3.8 Processament paral·lel de clips

En el sistema desenvolupat, el processament complet d'un vídeo pot implicar múltiples fases computacionalment costoses, com ara la transcripció automàtica de l'àudio, la detecció de canvis d'escena, la localització de rostres, la generació de subtítols o l'edició dels clips finals. Executades de forma seqüencial, aquestes operacions poden provocar temps d'execució molt elevats, especialment per vídeos llargs.

Per tal de reduir el temps total de processament i millorar la capacitat de resposta del sistema, s'ha optat per una estratègia de paral·lelització. Aquesta consisteix en l'execució simultània de diversos processos mitjançant tècniques de concurrència, aprofitant els recursos de la màquina (8 vCPU i 1 GPU) disponible a la instància AWS g4dn.2xlarge[25].

Els processos que s'executen en paral·lel són els següents:

- Detecció de canvis d'escena.
- Detecció de cares.
- Retalls dels clips.
- Reenquadrament 9:16 centrat en el rostre (o en pantalla dividida si n'hi ha dos).
- Inserció de subtítols als vídeos.

Gràcies a aquesta execució en paral·lel, s'ha aconseguit reduir molt el temps d'execució com es mostra a la *Figura 3.4.1.*

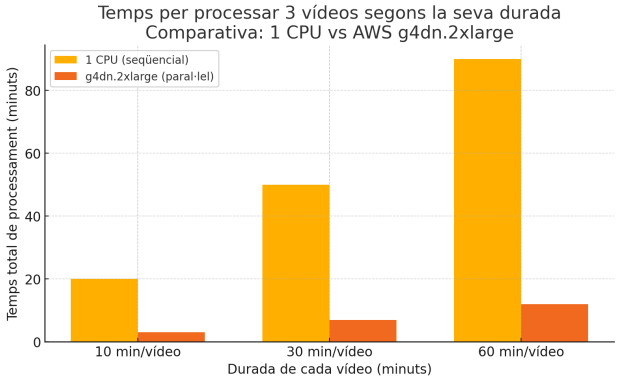


Figura 3.2.1 - Diferències en el temps d'execució.

4 Arquitectura AWS per desplegar l'aplicació al web

Per oferir AI Clip Generator com a servei web, s'ha desenvolupat una arquitectura al núvol utilitzant Amazon Web Services (AWS)[26] que equilibra rendiment, escalabilitat i cost. Aquesta arquitectura és la que ens permet fer servir l'acceleració per GPU i ús de l'execució en paral·lel, sense aquesta arquitectura no seria possible aconseguir un sistema tan ràpid i eficient com l'actual

A la *Figura 3.3.1* es presenta l'esquema del funcionament del servei.

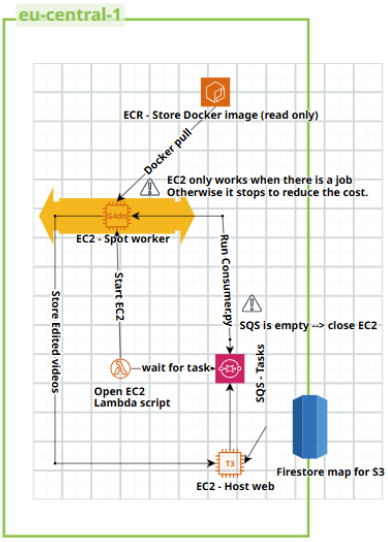


Figura 3.3.1 - Arquitectura AWS.

A l'apèndix es pot trobar cada component de l'arquitectura AWS explicat a detall

5 Prompt engineering

Durant el desenvolupament del projecte, s'han utilitzat models de llenguatge natural (LLMs) per dur a terme la detecció de parts rellevants dels vídeos. Per tal d'obtenir respostes precises i coherents ha estat clau aplicar tècniques de prompt engineering.

Per aconseguir el millor prompt possible que ens permeti elegir les millors parts d'un vídeo, s'ha seguit la guia de prompt engineering d'OpenAI[27].

Les tècniques utilitzades són les següents:

- **Donar identitat a la IA:** Es va definir la identitat del model com una app per extreure les millors parts d'una transcripció.
- **Donar instruccions clares:** Es van formular indicacions directes per evitar ambigüitats.
- **Proveir exemples:** En alguns casos es van utilitzar exemples de com hauria de ser la resposta ideal.
- **Afegir context extra i prohibicions:** Es va afegir context addicional sobre el tipus o autor del vídeo. Així mateix, es van incloure prohibicions per evitar que retorni coses que no tenen res a veure.

El prompt que s'ha fet servir es pot trobar a l'apèndix.

6 Experiments i resultats

Per tal de validar el funcionament del sistema desenvolupat, s'han definit i aplicat diverses mètriques de tipus objectiu, centrades en aspectes com l'eficiència, la precisió del subtitulat o la qualitat visual dels clips. Per motius de temps i viabilitat, les mètriques subjectives com la percepció d'usuari o l'impacte estètic queden com a línia de treball futura.

Cal advertir que els datasets emprats per als experiments no són estàndard, sinó seleccionats manualment per l'autor segons la naturalesa de cada prova. És a dir, no s'utilitzen els mateixos vídeos en tots els experiments. Per exemple, per mesurar la precisió en la detecció d'idioma, es fa servir un conjunt de vídeos en diferents llengües. En canvi, per avaluar la qualitat d'enquadrament o el dinamisme visual, es recorre a clips on predomini el rostre o amb freqüents canvis d'escena.

Aquesta selecció manual introdueix un grau de subjectivitat en la mostra de dades, que s'ha intentat minimitzar escollint vídeos amb característiques representatives. Tot i això, en un entorn de producció o per a estudis més rigorosos, seria recomanable utilitzar un conjunt de dades públic i ben definit.

6.1 Experiment 1: detecció d'idioma

Objectiu: Aquest experiment té com a objectiu avaluar la precisió del sistema a l'hora de detectar l'idioma principal del vídeo d'entrada. Aquesta informació és essencial per assegurar una transcripció correcta.

Dades: Conjunt de 6 vídeos de diferents idiomes.

Mètriques: Per mesurar l'eficiència, s'ha utilitzat la següent mètrica objectiva:

$$Precisió = \frac{Nº\ clips\ amb\ idioma\ ben\ detectat}{Total\ de\ clips} \bullet 100$$

Es a dir si d'un vídeo obtinc 6 clips dels quals 4 s'ha detectat l'idioma correctament significa que en aquell idioma he obtingut (4/6) * 100 = 66,6% d'accuracy.

- **Idioma real del clip:** assignat manualment.
- **Idioma detectat:** retornat pel model automàticament en processar el clip.

Resultats:

Vídeo ID (idioma)	**Nº Clips total**	**Nº Clips correctes**	**Accuracy (%)**
Vídeo 1 (català)	6	5	83,3%
Vídeo 2 (Euskera)	9	7	77,7%
Vídeo 3 (Gallec)	6	5	83,3%
Vídeo 4 (Francès)	7	7	100%
Vídeo 5 (Anglès)	12	12	100%
Vídeo 6 (Castellà)	9	9	100%
Total			90,71%

Els resultats mostren que el sistema és especialment fiable en llengües amb una gran representació al conjunt d'entrenament del model (com el castellà, l'anglès o el francès), on s'ha aconseguit una precisió del 100%.

En canvi, en llengües menys freqüents com el català, l'euskera o el gallec, s'observa una lleugera disminució de la precisió (entre el 77% i el 83%). Aquest fet pot atribuir-se a la similitud acústica amb el castellà o a la menor presència d'aquestes llengües en les dades d'entrenament del model Whisper.

Tot i així, amb una precisió global mitjana del 90,71%, es pot considerar que el sistema presenta un comportament robust i fiable tot i tenir marge de millora.

6.2 Experiment 2: precisió del subtitulat automàtic (Word Error Rate – WER)

Objectiu: Aquest experiment té com a objectiu avaluar la precisió del sistema de transcripció automàtica.

Dades: Conjunt de 3 vídeos de diferents idiomes (els mateixos que abans en català, castellà i anglès). De cadascun d'aquests vídeos agafem simplement un dels clips que duri aproximadament 1 minut.

Mètriques: S'ha utilitzat el Word Error Rate (WER), la mètrica estàndard per avaluar sistemes de reconeixement automàtic de veu. Es calcula mitjançant:

$$WER = \frac{S + D + I}{N} \bullet 100$$

On:

- **S (Substitutions):** paraules incorrectament substituïdes.
- **D (Deletions):** paraules que no apareixen al resultat, però sí a la referència.
- **I (Insertions):** paraules afegides erròniament.
- **N:** total de paraules a la transcripció de referència.

Una puntuació de WER més baixa indica una transcripció més precisa.

Per fer les comparacions es calcula el WER amb un clip curt (~50-60 segons) en cada idioma.

Resultats:

Idioma	Duració Clip s	WER %
Català	59	(12 + 10 + 5) / 200 = 10,5%
Anglès	58	(1 + 3 + 0) / 160 = 2,5%
Castellà	56	(0 + 3 + 0) / 167 = 1,7%

Els resultats mostren que la qualitat de les transcripcions és molt alta, especialment en idiomes com el castellà i l'anglès, on el WER és inferior al 3%. En el cas del català, la precisió es manté acceptable amb un WER del 10,5%, cosa que permet generar subtítols funcionals, encara que amb possibles petites errades.

6.3 Experiment 3: puntuació de composició d'imatge (cara ben enquadrada)

Objectiu: Aquest experiment té com a objectiu avaluar la capacitat del sistema per mantenir les cares ben centrades en els clips, especialment en vídeos amb diferents graus de dinamisme visual, entès com la freqüència de canvis d'escena dins del clip (vegeu Secció 3.4.1).

Dades: Conjunt de 3 videos amb diferents graus de dinamisme.

Mètriques:

$$Precisió = \frac{Segons\ amb\ cara\ enquadrada}{Segons\ totals\ del\ clip} \cdot 100$$

- **Segons amb cara enquadrada:** segons en què una o més cares estan centrades, visibles i no retallades dins del marc.
- **Segons totals del clip:** durada total del fragment de vídeo analitzat.

Aquesta mètrica s'ha mesurat mitjançant una avaluació visual manual, revisant cada clip i marcant en quins segments temporals les cares estan correctament centrades.

Resultats:

Clip ID	Durada total (s)	Segons amb cara ben enquadrada(s)	Percentatge de bon enquadrament
Clip 1 (poc dinàmic)	62	62	100%
Clip 2 (dinàmic)	19	17	89,47%
Clip 3 (molt dinàmic)	15	8	53,3%

Els resultats mostren una correlació clara entre el dinamisme del vídeo i la precisió de l'enquadrament facial:

- En clips amb poca variació de pla, el sistema aconsegueix mantenir el rostre enquadrat el 100% del temps, cosa que demostra un bon comportament del sistema.
- En clips amb moviment moderat, la taxa baixa lleugerament però continua sent molt acceptable (89,47%).
- En clips molt dinàmics, la taxa cau fins al 53,3%, evidenciant els límits actuals del sistema davant de situacions visuals complexes.

Aquestes dades indiquen que el sistema d'enquadrament és molt útil en contextos estàtics o semiestàtics (com podcasts o vídeos informatius), però que podria requerir millores o suport manual addicional en escenaris amb gran variabilitat visual.

6.4 Experiment 4: llegibilitat dels subtítols

Objectiu: Aquest experiment té com a objectiu avaluar si els subtítols generats automàticament són llegibles i estèticament adequats en el context dels clips editats, és a dir que no se sobreposen a les cares, o si ho fan, s'aplica la lògica de transparentar els subtitols.

Dades: Conjunt de 3 vídeos amb diferents estils de subtitulat.

Mètriques:

$$Precisió = \frac{Paraules\ ben\ col{\cdot}locades}{Total\ de\ paraules} \cdot 100$$

Es considera que una paraula està ben col·locada si:

- Apareix completament dins del marc del vídeo.
- No tapa elements importants com la cara del parlant.
- Té una mida i contrast suficient per a ser llegida fàcilment.

Resultats:

Clip ID	Paraules totals	Paraules ben col·locades	Precisió (%)
Clip 1 (One by one)	137	111	81,02%
Clip 2 (Creative)	114	114	100%
Clip 3 (Normal)	115	113	98,26%
Mitjana Global			93,09%

Els resultats mostren una alta precisió en la col·locació de subtítols, amb una mitjana global del 93,09% de les paraules ben posicionades. En dos dels tres casos, la precisió supera el 98%, fet que indica que el sistema funciona molt bé en escenaris convencionals i en estils de presentació creatius gràcies al sistema de transparència.

L'única excepció significativa es dona en el clip amb estil "one by one", on les paraules apareixen individualment animades, fet que pot provocar desajustos puntuals. En aquest cas, s'han detectat paraules que se solapen amb la cara principal del vídeo.

6.5 Experiment 5: temps de processament per minut de vídeo

Objectiu: Aquest experiment mesura el temps que triga el sistema a processar vídeos complets en dos entorns diferents (local + seqüencial i núvol amb GPU + paral·lelització), per tal d'avaluar l'impacte del maquinari en el rendiment global.

Configuració de la prova:
Les proves s'han realitzat en dos entorns diferents:

- Entorn local: CPU 4 cores, RAM 8GB, Sense GPU
- Entorn al núvol (AWS), CPU 8 cores, RAM 16GB, GPU NVIDIA T4

Dades: Conjunt de 3 vídeos de diferent longitud per veure el temps de processament en cadascun d'aquests.

Mètriques: Per mesurar l'eficiència, s'ha utilitzat la següent mètrica objectiva:

$$Temps\ per\ processar\ un\ minut\ de\ vídeo\ = \frac{Temps\ total\ de\ processament}{Durada\ del\ vídeo} \cdot 100$$

- **Temps total de processament:** temps en segons que el sistema triga a completar tot el procés.
- **Durada del vídeo:** temps total del vídeo original, expressat en minuts.

Amb aquesta mètrica determinem quants segons tarda el programa en processar un minut del vídeo original. A menys **s/min** millor.

Resultats:

	Local		Núvol	
Vídeo ID (Duració en mins)	**Total en seg.**	**Seg. per processar un minut**	**Total en seg.**	**Seg. per processar un minut**
V1 (1,39)	251	180,57	64	46,04
V2 (26,56)	3452	129,96	395	14,87
V3 (59,01)	5783	98	419	7,1

L'ús de GPU al núvol redueix significativament el temps de processament, fins a 13 vegades respecte a l'entorn local. Això es deu a la paral·lelització i l'acceleració per maquinari. La diferència s'accentua en vídeos llargs, confirmant la viabilitat del sistema en entorns amb alta càrrega.

7 Conclusions

El projecte AI Clip Generator ha demostrat ser una solució eficient per automatitzar l'edició de vídeos llargs en clips curts, optimitzats per a xarxes socials. La combinació de models de llenguatge (LLM), visió per computador i eines com FFmpeg i Whisper ha permès construir un sistema complet, escalable i fàcilment reutilitzable.

Beneficis i resultats destacats:

- El sistema redueix considerablement el temps de processament quan s'executa en entorns amb GPU, aconseguint fins a 14× més rapidesa respecte a execució local.
- La detecció d'idioma automàtica assoleix una precisió mitjana del 90,71%, amb resultats excel·lents en idiomes com el castellà, anglès i francès.
- El sistema de transcripció aconsegueix un WER (Word Error Rate) entre 1,7% i 2,5% en idiomes majoritaris.
- La detecció facial i reenquadrament ofereix una cobertura superior al 80% en clips estàtics, tot i que baixa fins al 53% en escenes molt dinàmiques.
- Els subtítols generats mostren una precisió del 93,09% pel que fa a la col·locació de paraules dins el clip, assegurant-ne la llegibilitat.

Limitacions:

- El sistema no és òptim per a vídeos molt dinàmics o amb múltiples canvis de pla ràpids.

- L'avaluació de qualitat es basa en un conjunt petit de vídeos seleccionats manualment, fet que pot introduir biaix.
- No s'han pogut implementar les mètriques subjectives proposades (com la rellevància percebuda o retenció d'audiència) per falta de temps.

Aplicacions i treballs futurs: Aquest tipus de tecnologia té aplicació més enllà de les xarxes socials. Pot utilitzar-se per:

- Resumir classes gravades, facilitant l'estudi o la creació de píndoles educatives.
- Millorar l'aprenentatge d'idiomes, seleccionant fragments rellevants segons paraules clau.
- Algorisme per determinar els colors de les paraules importants de forma intel·ligent.

En el futur, es podria ampliar el sistema amb:

- Personalització de criteris de selecció mitjançant feedback actiu de l'usuari.
- Millora de la detecció d'escenes dinàmiques amb models de tracking de rostre.
- Avaluacions subjectives amb usuaris reals per validar qualitativament la utilitat dels clips.

Aquest projecte, a més, ha servit com a exercici pràctic de desenvolupament d'un sistema complet, des del disseny fins al desplegament, consolidant coneixements en IA generativa, arquitectures modulars i optimització de rendiment.

8 Bibliografía

[1] Manic, M. (2024, juliol). Short-Form Video Content and Consumer Engagement in Digital Landscapes. Bulletin of the Transilvania University of Brașov. Series V: Economic Sciences, 17(66), 45–52. https://doi.org/10.31926/but.es.2024.17.66.1.4

[2] Lundqvist, J. (2023). Short-Format Video Consumption: Evaluation of Key Quality and UX Aspects for Generation Z (UPTEC STS 23033) [Tesi de màster, Uppsala University]. DiVA. Recuperat el 10 de juny de 2025, de https://www.diva-portal.org/smash/get/diva2:1773341/FULLTEXT01.pdf

[3] Grand View Research. (2023, març). Short Video Platforms Market Size & Share Report, 2023–2030 [Informe de mercat]. Grand View Research. Recuperat el 10 de juny de 2025, de https://www.grandviewresearch.com/industry-analysis/short-video-platforms-market-report

[4] OpusClip. (s. d.). OpusClip [Lloc web]. Recuperat el 10 de juny de 2025, de https://www.opus.pro/

[5] Wisecut.ai. (s. d.). Wisecut: AI video editor [Lloc web]. Recuperat el 10 de juny de 2025, de https://www.wisecut.ai/?lang=es/

[6] Descript. (s. d.). Descript: AI video editor [Lloc web]. Recuperat el 10 de juny de 2025, de https://www.descript.com/

[7] Pictory.ai. (s. d.). Pictory.ai: AI Video Editor [Lloc web]. Recuperat el 10 de juny de 2025, de https://www.pictory.ai/

[8] VEED.io. (s. d.). Workspaces [Lloc web]. Recuperat el 10 de juny de 2025, de https://www.veed.io/workspaces/

[9] Mbugua, S. (s. d.). On Software Modular Architecture: Concepts, Metrics and Trends [Preprint]. ResearchGate. Recuperat el 21 de juny de 2025, de http://researchgate.net/publication/359922034_On_Software_Modular_Architecture_Concepts_Metrics_and_Trends

[10] Ajiga, D., Okeleke, P. A., Folorunsho, S. O., & Ezeigweneme, C. (2024). Methodologies for developing scalable software frameworks that support growing business needs [Preprint]. ResearchGate. Recuperat el 21 de juny de 2025, de https://www.researchgate.net/publication/383409995_Methodologies_for_developing_scalable_software_frameworks_that_support_growing_business_needs

[11] Chen, C. (2025, gener). The Effectiveness and Impact of Short Videos as Advertising [Preprint]. Proceedings of the 3rd International Conference on Financial Technology and Business Analysis. Recuperat el 21 de juny de 2025, de https://www.researchgate.net/publication/387779705_The_Effectiveness_and_Impact_of_Short_Videos_as_Advertising

[12] YouTube. (2025, 9 de juny). *YouTube* [Lloc web]. Recuperat de https://www.youtube.com/

[13] yt-dlp contributors. (2025, 9 de juny). yt-dlp (versió 2025.06.09) [Repositori]. GitHub. https://github.com/yt-dlp/yt-dlp

[14] OpenAI. (s. d.). Whisper [Lloc web]. Recuperat el 9 de juny de 2025, de https://openai.com/index/whisper/

[15] OpenAI. (s. d.). API Platform [Documentació web]. Recuperat el 21 de juny de 2025, de https://openai.com/api/

[16] Park, Y., Diwan, A., Harwath, D., & Choi, E. (2025, 26 de maig). Rhapsody: A Dataset for Highlight Detection in Podcasts [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.19429

[17] TED Conferences, LLC. (s. d.). TED [Lloc web]. Recuperat el 13 de juny de 2025, de https://www.ted.com/

[18] FFmpeg. (s. d.). FFmpeg [Lloc web]. Recuperat el 9 de juny de 2025, de https://ffmpeg.org/

[19] OpenCV. (s. d.). OpenCV: Open Source Computer Vision Library [Lloc web]. Recuperat el 14 de juny de 2025, de https://opencv.org/

[20] Radwan, N. I., Salem, N. M., & El Adawy, M. I. (2012). Histogram correlation for video scene change detection. En D. C. Wyld, J. Zizka, & D. Nagamalai (Eds.), Advances in Computer Science, Engineering & Applications (Vol. 166, pp. 765–773). Springer. https://doi.org/10.1007/978-3-642-30157-5_76

[21] Garg, S. (s. d.). Group-Emotion-Recognition [Repositori de codi]. GitHub. Recuperat el 14 de juny de 2025, de https://github.com/samanyougarg/Group-Emotion-Recognition

[22] Garg, S. (2019, 3 de maig). Group Emotion Recognition Using Machine Learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1905.01118

[23] SubStation Alpha. (s. d.). ASS File Format Specification [Document tècnic]. TCAX. Recuperat el 14 de juny de 2025, de http://www.tcax.org/docs/ass-specs.htm

[24] Sammut, C., & Webb, G. I. (Eds.). (2011). TF–IDF. In Encyclopedia of Machine Learning (pp. 986–987). Springer. https://doi.org/10.1007/978-0-387-30164-8_832

[25] Amazon Web Services, Inc. (s. d.). Amazon EC2 G4 Instances [Lloc web]. Recuperat el 15 de juny de 2025, de https://aws.amazon.com/ec2/instance-types/g4/

[26] Amazon Web Services. (s. d.). Amazon Web Services [Lloc web]. Recuperat el 9 de juny de 2025, de https://aws.amazon.com/es/

[27] OpenAI. (s. d.). Text generation guide: Chat mode [Documentació web]. Recuperat el 18 de juny de 2025, de https://platform.openai.com/docs/guides/text?api-mode=chat

[28] Amazon Web Services, Inc. (s. d.). Amazon Simple Queue Service (SQS) [Lloc web]. Recuperat el 16 de juny de 2025, de https://aws.amazon.com/es/sqs/

[29] Amazon Web Services, Inc. (s. d.). Amazon Lambda [Lloc web]. Recuperat el 16 de juny de 2025, de https://aws.amazon.com/es/lambda/

[30] Amazon Web Services, Inc. (s. d.). Amazon Elastic Container Registry (ECR) [Lloc web]. Recuperat el 16 de juny de 2025, de https://aws.amazon.com/es/ecr/

9 Apèndix

A1. Interfície web

Accés a la Interfície web:

- URL: http://97.107.131.94/
- Password: TestProject2025

Aquesta és l'URL d'accés a la interfície web del sistema. Es pot accedir-hi amb les credencials proporcionades.

Cada usuari pot processar un màxim d'un vídeo i d'una hora màxim

A1.1. Elecció de criteri / tema del clip

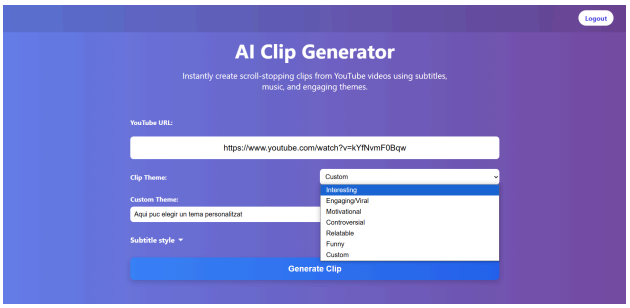
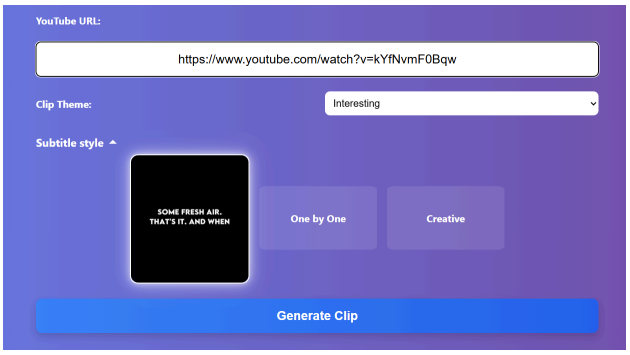


Figura A1.1.1 - UI que permet elegir el tipus de clip.

El primer pas és posar una URL de YouTube en el camp indicat i elegir el criteri que es vol utilitzar per seleccionar els fragments més rellevants del vídeo, l'usuari pot introduir manualment els temes o conceptes d'interès perquè el sistema busqui les parts del vídeo que hi estiguin relacionades.

A1.2. Elecció del estil de subtítols



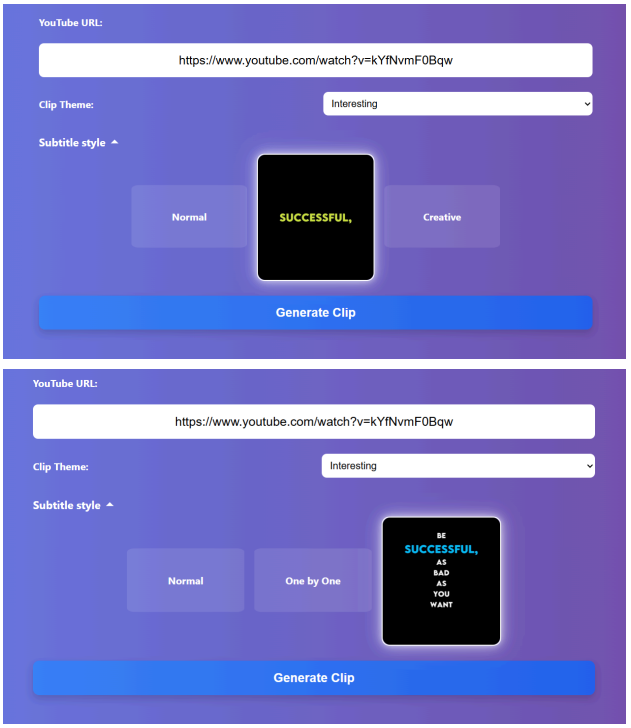


Figura A1.2.1 - Estils a elegir.

El segon pas és elegir el tipus de subtítols que es vol aplicar als clips generats. L'usuari pot triar entre diversos estils visuals predefinits que es poden veure de manera visual a la web.

Per últim donar-li a "Generate Clip".

A2. Algorismes

Aquí s'expliquen els algorismes que no tenien cabuda dins de l'informe.

A2.1. Algorisme detallat de la detecció de cares en un frame

Input: Fotogrames corresponents als inicis dels canvis d'escena, Xarxa neuronal preentrenada per detecció facial.
Output: Coordenades del rostre principal per a cada escena.

1. *Es carrega la xarxa neuronal.*

2. *Per a cada primer frame dels canvis d'escena:*

a. *Es converteix la imatge en un "blob", una forma estandarditzada que prepara la imatge per a la xarxa neuronal.*
b. *Es passa el blob per la xarxa per detectar els rostres.*
c. *Es guarden només els rostres detectats amb una confiança superior al 50%.*
d. *Si es detecten més d'un rostre:*
   i. *Es calcula quin és més a prop del centre de la imatge.*
   ii. *Aquest rostre es considera el principal per enfocar.*

iii. *Els altres es guarden com a rostres secundaris.*
iv. *Si no es detecta cap rostre s'assigna una posició especial* `[-9999, -9999]` *per indicar que no s'ha d'aplicar reenquadrament automàtic.*

Algorisme 2 - Detecció de cares.

A2.2. Algorisme detallat de la detecció de paraules importants mitjançant term freqüency

Input: *Transcripció completa del clip (text amb timestamps).*
Output: *Llista de paraules clau destacades amb la seva posició adaptada per a subtítols visuals.*

1. *S'analitza tot el text del vídeo i es compta quantes vegades apareix cada paraula (freqüència).*
   a. *Com més es repeteix una paraula, més important es considera.*

2. *Es neteja el text del clip eliminant paraules comunes com "el", "de", "i" (stopwords), que no aporten significat.*

3. *S'escullen les X paraules que més es repeteixen al clip.*

4. *A més de comptar repeticions, s'utilitzen altres regles (heurístiques) per decidir si una paraula és important:*
   a. *Si és un nom propi.*
   b. *Si és una paraula llarga (més de X lletres).*
   c. *Aquestes paraules reben més pes.*

5. *Abans de mostrar les paraules importants com a subtítols al vídeo:*
   a. *Es verifica que no surtin dels marges del format vertical (9:16), ja que algunes seran més grans.*
   b. *S'ajusta la seva posició si cal perquè es vegin bé dins de l'enquadrament.*

Algorisme 3 - Detecció de paraules importants.

A2.3. Algorisme detallat de la creació de les miniatures dels clips

Input: *Clip de vídeo amb coordenades de detecció facial, logotip opcional proporcionat per l'usuari.*
Output: *Imatge miniatura en format 9:16 amb la cara principal centrada i opcionalment amb marca d'aigua.*

1. *Es pren el primer fotograma (frame) de cada clip.*

2. *Amb les mateixes coordenades que s'han usat per detectar les cares: Es retalla una imatge (miniatura) amb la cara principal centrada.*

3. *S'apliquen efectes visuals amb OpenCV:*
   a. *S'enfosqueixen lleugerament les vores de la imatge.*
   b. *Es pot desenfocar una part del fons.*
   c. *S'hi destaca el rostre.*

4. *També es pot afegir una marca d'aigua o logotip proporcionat per l'usuari:*
   a. *Es pot triar la posició, mida i opacitat perquè s'adapti a la miniatura.*

Algorisme 4 - Creació de les miniatures dels clips.

A3. Arquitectura AWS

A continuació es descriu l'arquitectura implementada i com interactua cada component.

- **Frontend i servidor web (Flask)**: Hi ha una aplicació web desenvolupada amb Flask que actua com a interfície per als usuaris. Aquest frontend pot residir en un servidor lleuger (per exemple, una instància EC2 t3.small) o fins i tot en un hosting tradicional, ja que la seva funció principal és servir la pàgina, autenticar els usuaris i recollir les sol·licituds de generació de clips. L'usuari introdueix l'URL de YouTube, selecciona opcions (format de subtítols, estil, criteris de highlight, etc.) i envia la petició.

- **Cua de missatges SQS**[28]**:** En lloc de processar la petició directament al servidor web, el sistema l'encua a Amazon SQS (Simple Queue Service). S'ha creat una cua estàndard on cada missatge representa una tasca de processament de vídeo. El missatge conté tots els paràmetres necessaris per ser processat, com la URL del vídeo, opcions seleccionades, etc.

- **Funció AWS Lambda**[29]**:** Quan s'encua un nou missatge a SQS, s'activa automàticament una funció AWS Lambda per gestionar el processament. Aquesta funció s'encarrega de verificar l'estat de la instància EC2 dedicada al processament intensiu. Si la instància està aturada, Lambda la inicia automàticament mitjançant l'API d'AWS, assegurant així que el sistema només consumeixi recursos quan realment hi ha tasques pendents. Aquest enfocament permet mantenir una arquitectura eficient, escalable i de baix cost, ja que el cost de mantenir la instància g4dn.2xlarge activa les 24 hores del dia és extremadament elevat.

- Dins del servidor s'executa un script desenvolupat per escoltar constantment la cua SQS. Quan arriba un missatge (tasca de vídeo), el treu de la cua i inicia un contenidor Docker on s'executa el programa. Finalment, quan el programa acaba d'executar-se, es tanca el contenidor Docker i es notifica a l'usuari per correu electrònic que els clips ja han estat processats.

- Es fa servir el mateix servidor web com a sistema centralitzat d'emmagatzematge dels vídeos. Cada fitxer generat s'organitza en carpetes

separades per usuari i per projecte, cosa que en facilita la gestió i recuperació posterior. Paral·lelament, la informació sobre l'estat del processament i la ubicació exacta dels fitxers es desa a Firebase Firestore, on cada document d'usuari conté referències creuades als recursos emmagatzemats. Aquesta estructura permet que el frontend consulti ràpidament Firestore per mostrar a l'usuari el progrés i els resultats.

- Si passa X temps sense rebre nous missatges, s'inicia l'apagada automàtica de la instància: això es fa mitjançant una crida a la API d'AWS per marcar la instància per a la seva aturada.

- **ECR[30] Docker pull:** Tot i que això no afecta l'execució del programa i es fa de forma manual, si en un futur es volgués escalar el projecte, seria totalment viable, ja que totes les dependències, fitxers i configuracions necessàries estan desades en un repositori Docker al núvol (ECR). Això permet clonar els servidors utilitzats simplement amb un Docker pull.

A4. Estructura fitxers .ass

```
INFORMACIÓ GENERAL DELS SUBTITOLS
(COLOR, LLETRA, TAMANY, ETC. PER
DEFECTE)

[Script Info]
Title: Auto-Generated Subtitles
ScriptType: v4.00+
PlayResX: 1920
PlayResY: 1080
ScaledBorderAndShadow: yes

[V4+ Styles]
Format: Name, Fontname, Fontsize, PrimaryColour,ETC...
Style: Default,LEMON MILK Bold,28,&HFFFFFF,ETC..
```

```
ASSIGNACIÓ DE COLORS, LLETRA, TAMANY, ETC.
A PARAULES CONCRECTES
[Events]
Format: Layer, Start, End, Style, Name, MarginL, MarginR, MarginV, Effect, Text
Dialogue: 0,00:00:00.00,00:00:02.84,Default,ETC...{\fs50}Usually
Dialogue: 0,00:00:00.24,00:00:02.84,Default,ETC...{\fs50}don't
Dialogue: 0,00:00:00.50,00:00:02.84,Default,ETC...{\fs50}like
Dialogue: 0,00:00:00.70,00:00:02.84,Default,ETC...{\fs50}mine
Dialogue: 0,00:00:00.96,00:00:02.84,Default,ETC...{\fs50}finely
```

Figura A4 - Estructura fitxer .ass.

A5. Prompt usat

```
This is an app that transforms long youtube videos into short tik tok form videos
(NOTICE THAT THE VIDEO MAY BE ON ANOTHER LANGUAGE OTHER THAN ENGLISH)

This particular video is from CHANNEL_PLACEHOLDER
 and is titled TITLE_PLACEHOLDER — use that to guide your selections.

I will attach the sentences of the script of a video,
 I need you to tell me which parts are the MOST PROMPT_PLACEHOLDER
and will make the best tiktok.

IMPORTANT, YOU NEED TO AT LEAST JOIN 3 CONSECUTIVE SENTENCES
TO MAKE A PART AND NO MORE THAN 7.
ONLY RETURN THE NUMBER OF THE SENTENCES, NOT THE ENTIRE TEXT.

**If any part contains more than 7 sentences, the output is INVALID**

so one part = MINIMUM 3 SENTENCES AND MAXIMUM 7
FOR EXAMPLE:

3. You have, there's no purpose in your life.
4. You know, people will need to have purpose to get up.
5. They need purpose to perform.
6. You need to get to a point in your life where there's nothing on the docket.
7. There is no 5K, there's no, there's no,
 I'm going to get into school to be this or that.
8. It's still perform to the highest level.
9. Because what people don't get is one day that thing's going to come up.
10 And if you're not constantly performing without purpose,
 you're not going to be ready when the time comes.

If the most important sentences are 5,6,7,8,9,10,
 return: 5,6,7,8,9,10 do this with as many parts as you consider.
DON'T ADD ANY EXTRA TEXT PLEASE NOT EVEN A LIST, JUST THE NUMBERS LIKE THIS:

5,6,7,8,9,10
100,101,102,103,104,105
157,158,159,160

FAILURE TO FOLLOW THESE EXACT RULES WILL RESULT IN AN INVALID OUTPUT

Script:
```

Figura A5 - Prompt usat.

A6. Dataset per realitzar Experiment 1: Detecció d'idioma

El dataset que s'ha elegit té una gran varietat d'idiomes on tenim idiomes comuns com l'anglès, francès o castellà, i també idiomes més minoritaris com el gallec, euskera o català per veure quines són les limitacions que pot arribar a tenir.

Vídeo	Idioma
QUÈ ESTUDIEN L…	Català
Euskera, la lengua …	Euskera
Un home de 72 ano…	Gallec
😨 Comment parler…	Francès
"The Single Greates…	Anglès
El futuro ya está aq…	Castellà

A7. Dataset per realitzar Experiment 2: Precisió del subtitulat automàtic (Word Error Rate – WER)

Vídeo	Idioma
QUÈ ESTUDIEN L…	Català
“The Single Greates…	Anglès
El futuro ya está aq…	Castellà

A8. Dataset per realitzar Experiment 3: Puntuació de composició d’imatge (cara ben enquadrada)

Clip	Vídeo original	Dinamisme
https://youtube.com/shorts/Lq8pZQ80M5k	“The Sing…	poc dinàmic
https://youtube.com/shorts/pqO5VrL7hLY	Probé la …	dinàmic
https://youtube.com/shorts/ggpLci5MZP8	Probé la …	Molt dinàmic

A9. Dataset per realitzar Experiment 4: Llegibilitat dels subtítols

Clip	Vídeo original	Tipus de subtitol
https://youtube.com/shorts/Lq8pZQ80M5k	“The Sing…	One by one
https://youtube.com/shorts/y7k96AM1HHM	A simple …	normal
https://youtube.com/shorts/fDkCdd2NIOs	IlloJuan e…	creative

A10. Dataset per realitzar Experiment 5: Temps de processament per minut de vídeo

Vídeo	Descripció	Durada (min)
Steve Jobs…	Vídeo curt de persona parlant	1:39
5 Años de…	Vídeo mitjà de persona parlant	26:56
The Art O…	Video llarg de podcast entre dos persones	59:01

- E-mail de contacte: Carles.FornesM@autonoma.cat
- Menció realitzada: Computació
- Treball tutoritzat per: Guillermo Eduardo Torres (CVC)
- Curs 2024/25